Best Practices for OAI PMH DataProvider Implementations and Shareable Metadata

DLF/NSDL Working Group on OAI PMH Best Practices

October 2007

Edited by Sarah L. Shreeves, Jenn Riley, and Kat Hagedorn

DIGITAL LIBRARY FEDERATION

Best Practices for OAI PMH Data Provider Implementations and Shareable Metadata

DLF/NSDL WORKING GROUP ON OAI PMH BEST PRACTICES

Edited by:
Sarah L. Shreeves, Jenn Riley, and Kat Hagedorn

Digital Library Federation
Washington, D.C.
2007

Table of Contents

1. Introduction

1.1. BACKGROUND

The Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH) has been widely adopted since its inception in 2001; as of May 2007 there are more than 1,400 active data providers from a wide variety of domains and institutions.[1] The protocol has demonstrated its usefulness as a tool to move and aggregate metadata from diverse institutions. The National Science Digital Library (NSDL), OAIster, American South.org, and the Institute of Museum and Library Services (IMLS) Digital Collections and Content Gateway are examples of metadata aggregations made possible through the OAI PMH. Building on this work, the Aquifer Initiative of the Digital Library Federation (DLF) is developing the American Social History Online aggregation of metadata harvested from DLF members in the Metadata Object Description Schema (MODS). Google™ uses the OAI PMH as a sitemap.[2] For more examples of how widely the OAI PMH has been adopted, see Brogan's 2006 report, *Context and Contributions: Building the Distributed Digital Library*, which does an admirable job laying out the current landscape of digital library services and systems that are based on the OAI PMH.

Metadata harvesting involves two parties: the data provider and the service provider. The data provider is any institution, organization, or individual who exposes metadata (usually describing one or more resources) via the OAI PMH. The service provider uses the OAI PMH to harvest the data provider's metadata. The service provider generally aggregates metadata from many different data providers and creates a database of this metadata. The intention in creating such an aggregation of metadata is to provide users with a way of searching one database, rather than many, to discover resources distributed across the Internet. In addition, service providers may build services beyond the standard search and retrieval service—e.g., aggregations designed to support curriculum development or building personal virtual collections.

The OAI PMH is quite flexible as there are relatively few required pieces for implementation: valid responses to OAI PMH verbs, the use of oai_dc (the Dublin Core Metadata Element Set or unqualified Dublin Core), a unique and persistent OAI identifier, and a datestamp. The *Implementation Guidelines* (Open Archives Initiative 2002b) have a limited technical scope, are intended for a general audience of implementers, and do not describe the consequences of not implementing some of the optional features of the protocol. This has meant that many of the features of the OAI PMH, such as sets and use of descriptive containers, which are quite helpful for service providers have been underutilized. In addition, the need for best practices for the metadata provided through the OAI PMH can best be seen in the work that service providers have had to do to normalize and manipulate their aggregations to ensure a certain threshold of usefulness for end-users. High quality "shareable" metadata will be crucial to the next step in useful metadata aggregations.

Thus, as the protocol has become more widely adopted, several broad areas of concern that would benefit from the establishment of best practices have surfaced—mainly through the documentation of service providers.

In summer 2004, the DLF sponsored a meeting at the California Digital Library. DLF and NSDL-affiliated data and service providers as well as other interested individuals discussed issues about OAI PMH implementations and concerns stemming from the harvesting of metadata from diverse collections. This ad hoc working group brainstormed a large list of areas that needed guidelines or best practices and agreed to establish a wiki, hosted by the NSDL, to write a set of best practices. This work was chiefly facilitated by Kat Hagedorn of the University of Michigan, Sarah L. Shreeves of the University of Illinois at Urbana-Champaign, and Jenn Riley of Indiana University. It was part of an IMLS-funded effort to establish a training program and set of resources on the implementation of the OAI PMH for data providers and DLF member institutions. The best practices work was coordinated through weekly conference calls and twice-yearly meetings held at the DLF Forums.[3]

This document presents the best practices as they existed on the wiki as of April 2007.[4]

[1] See the University of Illinois data provider registry for a current number: **http://gita.grainger.uiuc.edu/registry/**.

[2] See **http://www.google.com/webmasters/sitemaps/docs/en/other.html**.

[3] Much of this documentation is available through a wiki hosted by the University of Michigan. **http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/Archive**

[4] As of August 2007, the wiki is available at **http://webservices.itcs.umich.edu/mediawiki/oaibp/**.

1.2. SCOPE AND AUDIENCE

These best practices are designed to provide data providers, service providers, and system designers with information on how to best implement the OAI PMH on the data provider side. Data providers will also find information on how to create or map metadata that is shareable outside of the local environment and can be used by service providers to support the end-user in finding the variety of rich resources available. System designers in particular may find this document useful in deciding which optional pieces of the OAI PMH to implement. This document is not an introduction to the OAI PMH; it assumes a basic knowledge of the protocol and how it works.

The following sections have been included in this document:

General Areas of Competency

These are general areas of competence that are necessary before the OAI PMH can be implemented and used successfully. Data providers—even if the OAI PMH is already built into their current digital library system—should be conversant with the issues presented here. These areas represent the minimum of proficiency that is necessary to be a "good" OAI PMH data provider.

Best Practices for OAI PMH Data Provider Implementations

This section deals with best practices for implementing the OAI PMH for data providers. Also included in this section are guidelines for some of the optional pieces of the OAI PMH, including sets, branding, rights, and use of the `<about>` container.

Best Practices for Shareable Metadata

This section presents best practices for shareable metadata within the specific context of the OAI PMH, in terms of technical issues (such as XML encoding) and metadata format, semantics, and content.

The working group responsible for this document is aware that there are other areas in which best practices and other community-building tools could be developed. These include best practices for service provider implementations, communications between data and service providers, and a place to share tools and strategies to extend the OAI PMH. These areas, however, were deemed out of scope for this effort. Also out of scope for this document is any discussion of the Open Archives Initiative current development work on Object Reuse and Exchange (ORE).

1.3. GUIDING PRINCIPLES

The following principles guided the development of this document:

- Do not duplicate what has already been developed and is openly available. For this reason, many of these best practices and guidelines link to information outside of this document. The working group has added information and context for these, as needed.
- Give context for best practices. Let the reader understand what the ramifications are for following (or not) the best practice. Give examples whenever possible.
- Use clear and concise language and avoid jargon whenever possible.
- Be prescriptive whenever possible.
- Link within the document as much as possible.
- Give real examples whenever possible. In some cases it was not possible or easy to find real examples, so fabricated examples are given.

1.4 NOTE ON LANGUAGE

Because the best practices were written by a large group of individuals (see section 1.4. for a list of contributors), readers will find that this document changes slightly in tone and voice between different sections. These differences have been mitigated as much as possible.

In addition, because of the current effort within the Open Archives Initiative on Object Reuse and Exchange (ORE) we try to refer explicitly to the OAI PMH as much as possible. However, when speaking of data and service providers and some other specific pieces of the OAI PMH, we refer to these as 'OAI data providers' or 'OAI service providers'. We hope that, given the context of this document, this lack of specificity is forgivable.

1.4 ACKNOWLEDGMENTS

We would especially like to acknowledge the encouragement and support of David Seaman who, as Executive Director of the Digital Library Federation, initiated and funded the activities of this working group.

This document is the result of a large group of individuals with a range of experience with the OAI PMH, including both service and data providers and metadata librarians. We would like to thank and acknowledge all of those who have given their time and expertise to help develop these best practices.

The following individuals contributed to the writing of this document:

- Caroline Arms (Library of Congress)
- Tim Cole (University of Illinois at Urbana-Champaign)
- Naomi Dushay (Colorado State University; this work developed while at Cornell University)
- Muriel Foulonneau (Centre National de la Recherché Scientifique; this work developed while at University of Illinois at Urbana-Champaign)
- Tom Habing (University of Illinois at Urbana-Champaign)
- Kat Hagedorn (University of Michigan)
- Arwen Hutt (University of California – San Diego and formerly at University of Tennessee at Knoxville)
- Diane Hillmann (Cornell University)
- Ann Lally (University of Washington)
- Bill Landis (Yale University; this work developed while at California Digital Library)
- Clay Redding (Library of Congress; this work developed while at Princeton University)
- Jenn Riley (Indiana University)
- Sarah Shreeves (University of Illinois at Urbana-Champaign)
- Jewel Ward (University of North Carolina; this work developed while at University of Southern California)
- Simeon Warner (Cornell University)
- Jeff Young (Online Computer Library Center [OCLC])

The following people formed the original planning group for this effort:

- Naomi Dushay (Colorado State University; this work developed while at Cornell University)
- Kat Hagedorn (University of Michigan)
- Martin Halbert (Emory University)
- Diane Hillmann (Cornell University)
- David Seaman (Dartmouth University; this work developed while executive director of the Digital Library Federation)
- Sarah Shreeves (University of Illinois at Urbana-Champaign)
- Roy Tennant (OCLC; this work developed while at California Digital Library)

The following individuals attended the inaugural meeting in July 2004 at the California Digital Library:

- Caroline Arms (Library of Congress)
- Naomi Dushay (Colorado State University; this work developed while at Cornell University)
- Muriel Foulonneau (Centre National de la Recherché Scientifique; this work developed while at University of Illinois at Urbana-Champaign)
- Kat Hagedorn (University of Michigan)
- Martin Halbert (Emory University)
- Ann Lally (University of Washington)
- Bill Moen (University of North Texas)
- Clay Redding (Library of Congress; this work developed while at Princeton University)

- Jenn Riley (Indiana University)
- Sarah Shreeves (University of Illinois at Urbana-Champaign)
- Robert Tansley (Google; this work developed while at Hewlett-Packard)
- Roy Tennant (OCLC; this work developed while at California Digital Library)
- Simeon Warner (Cornell University)
- Jeff Young (OCLC)

2. General Areas of Competency

While the OAI PMH has been termed "low barrier" (Lagoze and Van de Sompel 2001), there are some general areas of competency that are essential to productively participating as an OAI PMH data provider. Many institutions that now have OAI PMH data providers have little control over the actual implementation of the data provider software because it is packaged within a commercial, proprietary digital library system. However, these institutions and any others participating as data providers have a responsibility to be knowledgeable about three main areas of competency:

- Ability to create quality, shareable metadata

  Organizations wishing to implement OAI data provider services should have a good understanding of metadata standards and should make use of them. They should also consider carefully the shareability of their metadata; often metadata that are useful within a local context are not when pulled out of this environment. See Section 4, "Best Practices for Shareable Metadata" in this document for guidance on creating shareable or interoperable metadata. The NSDL *Metadata Primer* (2005) and the *Descriptive Metadata Guidelines for RLG Cultural Materials* (RLG 2005) are also helpful resources.

- Working understanding of XML, XML namespaces, and XML schemas

  An understanding of XML, XML namespaces, and XML schemas is fundamental for a successful data provider and shareable metadata. This document covers some basic information; the NSDL XML FAQ from the NSDL *Metadata Primer* (2005) and the Wikipedia entry on XML[5] are also helpful resources for beginners.

- Familiarity with the implementation guidelines for the OAI PMH

  This seems obvious, but often—particularly with the use of turnkey systems—would-be data providers do not really understand how the OAI PMH works. There are a number of excellent guides to the protocol, but the protocol specifications themselves (Open Archives Initiative 2002c) and the *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting* (Open Archives Initiative 2002b) are generally accessible. For a general overview of the OAI PMH, see Lagoze and Van de Sompel (2001), the Frequently Asked Questions page for the Open Archives Initiative (Open Archives Initiative 2002a), *OAI for Beginners: The Open Archives Forum Online Tutorial* (Open Archives Forum 2003), and Shreeves (2005).

[5] **http://en.wikipedia.org/wiki/Xml**

3. Best Practices for OAI PMH Data Provider Implementations

3.1. INTRODUCTION

As stated above, the OAI PMH has relatively few requirements for the implementation of a data provider: the valid responses to OAI PMH verbs, the use of oai_dc (unqualified Dublin Core) as a metadata format, unique and persistent OAI identifiers, and a datestamp for each item. The *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting* do not describe the benefits and consequences of some of the optional features of the protocol. This has meant that many useful features of OAI PMH—such as multiple metadata formats, sets, descriptive containers, and others—that are helpful for service providers have been underutilized or have remained unimplemented by data providers. This can be particularly true of data providers built into digital content management systems.

These best practices for OAI data provider implementations are based on the experiences of well established service providers and data providers. The goal is to provide a useful set of guidelines for organizations implementing a stand-alone OAI data provider as well as vendors and software developers who are responsible for OAI data providers included within larger systems.

This chapter is divided into three main components:

- The introduction, which provides background information for this section
- A section on managing the repository lifecycle, which offers general guidelines for various stages in the lifecycle of an OAI repository
- Several sections on the best practices for specific elements of OAI PMH data provider implementations

Throughout, the working group has tried to offer specific, concrete best practices and guidelines whenever possible. However, certain sections present either a range of choices or a spectrum of best practices from ideal to acceptable. These often represent an acknowledgment of the diversity and range of capabilities within the OAI PMH community and, in some cases, areas where a best practice has not clearly evolved. In each section, a basic overview of the protocol definition of the concept is presented, but certain practical knowledge of the OAI PMH is assumed. It is highly recommended that readers have read through the specifications in the *OAI Protocol for Metadata Harvesting* (Open Archives Initiative 2002c) as well as the *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting* (Open Archives Initiative 2002b). The technical specifications of the NSDL *Metadata Primer* (2005) are also recommended. Although written specifically for use with the NSDL, these also are useful for any OAI repository.

These best practices do not offer an assessment of any particular flavor of OAI PMH data provider. Many data providers are open source and freely available,[6] and many digital content management systems now include a built-in OAI data provider. The best practices may be helpful in making determinations about which open-source data provider to select or what functionality to look for in commercial digital content management systems.

3.2. MANAGING THE REPOSITORY LIFECYCLE

Managing the lifecycle of an OAI PMH repository is an important part of a data provider's responsibilities.[7] The most visible part of the lifecycle maintenance is to register the data provider with the Open Archives Initiative, but data providers also need to pay attention to other potential changes in the repository (for example, a change to all OAI identifiers in a repository) and the impact of these changes on service providers.

[6] The OAI Tools page at **http://www.openarchives.org/tools/tools.html** and/or a search on SourceForge at **http://sourceforge.net** will reveal many flavors of data providers.

[7] The terms "repository" and "data provider" are often used interchangeably, although technically the repository refers to the actual server that can process the OAI verbs and the data provider is the entity responsible for the repository.

3.2.3. REPOSITORY END OF LIFE

A repository's end of life might be caused by two different situations:

- All of the resources described by the metadata are no longer available or no longer exist.
- The data provider cannot or is not willing to maintain the OAI repository.

If resources described are no longer available or no longer exist, a data provider should alert service providers that the repository will be taken down and that service providers should purge metadata included in their aggregations. This can be done through direct contact with service providers and announcements on OAI listservs (such as oai-general and oai-implementers). If the OAI repository supports deleted records, all records should be marked deleted. The repository URL should not lead to a 404 error, but should indicate that the repository has been taken down.

If the organization maintaining the OAI repository is no longer able or willing to do this, the data provider should again alert service providers that the repository will be taken down and that service providers should purge metadata included in their aggregations. However, data providers might also investigate alternative options, particularly if the metadata items exposed are relatively static. If the resources described will still be available and will not change, an alternate institution might volunteer to maintain the technical infrastructure of the repository. This could be done through a Static Repository, for example. If an alternate institution is willing to accept the responsibility of the technical maintenance of the repository, then the data provider should ensure that it communicates that information as widely as possible.

3.3. BEST PRACTICES FOR THE IDENTIFY RESPONSE

See the protocol specifications on the Identify response:
http://www.openarchives.org/OAI/openarchivesprotocol.html#Identify

The Identify response contains administrative information about a repository that is useful for service providers for communication and maintenance purposes. It is a best practice for OAI PMH data providers to keep Identify responses accurate and up to date.

3.3.1. BEST PRACTICES FOR MANDATORY ELEMENTS OF THE IDENTIFY RESPONSE

`<repositoryName>`

A human-readable name for the repository.

The repositoryName is often used by service providers to quickly identify possible repositories to harvest and may be used to identify the origin of records within whatever system a service provider builds. Thus the repositoryName should give the name of the repository, identify the organization responsible for the repository if applicable, and whenever possible give a clue as to the content of the repository.

Examples of good repository names are:

The Journal of Community Informatics

Caltech Computer Science Technical Reports

`<baseURL>`

The base URL of the repository.

The base URL should be kept current. See Section 3.2, "Managing the Repository Lifecycle" for more information on handling a change to the base URL.

`<protocolVersion>`

The version of the OAI PMH supported by the repository.

The current version is 2.0 and has been since mid-2002. It is expected that active data providers will be on version 2.0 of the OAI PMH.

3.2.1. REPOSITORY CONFORMANCE AND REGISTRATION

It is a best practice to register an OAI PMH repository with the official OAI registry of data providers. The registry itself is available at **http://www.openarchives.org/Register/BrowseSites**, and the registration site is available at **http://www.openarchives.org/data/registerasprovider.html**.

The primary benefits of registering a data provider in the official OAI registry are—

- Conformance testing to ensure that the data provider meets the OAI specifications;
- Once passed, periodic retesting of your repository for conformance;
- Publicity of the availability of the repository for harvest; and
- Availability for inclusion in other OAI repository registries, including the OAI Registry at the University of Illinois at Urbana-Champaign (UIUC),[8] which is the most comprehensive list of OAI data providers currently available. The UIUC registry picks up new data providers on the official OAI site monthly.

Conformance testing is particularly important. If an OAI repository does not pass the conformance testing, it is likely that service providers will have difficulty harvesting your metadata. Note that the conformance testing offered through the official OAI registry is limited to required pieces of the protocol and does not check the best practices published here.

It is possible to go through the conformance testing and not register your OAI repository. This is a good option for data providers that wish to test their implementations before registering them, or for data providers that are not interested in widely advertising their OAI repository.

3.2.2. REPOSITORY MAINTENANCE

An OAI repository requires a certain level of maintenance to guarantee that the information provided is up to date. Information included in the Identify response (see Section 3.3, "Best Practices for the Identify Response" in this document for more information) or any of the descriptive containers at the repository or set level should be kept up to date. It is particularly important that the e-mail of the current contact person be included in the Identify response.

Data providers should make an effort to know who is regularly harvesting their OAI repository. This can be done through checking server logs. Having this information can ease communication issues, particularly when there are substantial changes to or downtime for an OAI repository.

If an OAI repository is going to be unavailable for a certain period of time, it is a best practice that this be communicated to the service providers that regularly harvest that repository. A human-readable page should also be published with information about how long the repository is likely to be down and with contact information for the repository administrator. The repository URL should not lead to a 404 error. The information that the repository will be or is offline should be mentioned in the repository Identify response, appropriate error messages, and/or a human-readable page description.

If the OAI repository's location (base URL) changes, a redirection mechanism, appropriate harvesting error message, and/or human-readable page can help ensure that service providers are informed. Service providers that regularly harvest the OAI repository should be informed directly. The data provider should also reregister the repository at the OAI site and ensure that all other registries are aware of the updated repository base URL.

If there are other major changes to the OAI repository (e.g., reorganization of sets, a change in the OAI identifiers used) that will affect how records are harvested (particularly in incremental harvests), it is a best practice that these changes be communicated with the service providers that regularly harvest the repository.

It can also be useful for data providers to communicate other information about their repository, such as how often the repository changes in terms of number of records and/or sets (see Section 3.10, "Best Practices for Sets" in this document), and information about rights over the metadata (see Section 3.11.1, "Best Practices for Expressing Rights over Metadata" in this document). This information should be communicated with service providers that regularly harvest the repository. This information can also be included in the `<description>` containers for both the repository and the set. Generally, any information that might have an impact on service provider routines should be communicated, either through direct contact or through repository or set descriptions.

[8] **http://gita.grainger.uiuc.edu/registry/searchform.asp**

`<earliestDatestamp>`

The lower limit of all datestamps recording changes, modifications, or deletions in the repository.

See Section 3.5, "Best Practices for Datestamps" in this document for more information.

`<deletedRecord>`

The value will be no, transient, or persistent.

See Section 3.6, "Best Practices for Deleted Records" for more information.

`<granularity>`

The finest granularity of the datestamp.

See Section 3.5, "Best Practices for Datestamps" in this document for more information.

`<adminEmail>`

The e-mail address of an administrator of the repository.

The Identify response may contain multiple instances of adminEmail. This value allows service providers to communicate with data providers when they encounter difficulty in harvesting or to report other problems. The value should be a plain e-mail address such as someone@somewhere.org and not a mailto: URI. The adminEmail should include the e-mail of the person responsible for the implementation and maintenance of the repository. In addition, the e-mail address of the person responsible for the records exposed by the OAI repository is useful. In any case, the administrator should be able to respond to problems and questions or direct service providers to the appropriate person. Please note that many digital library management systems that include an OAI repository have a default value in the adminEmail element. This should be updated to the appropriate e-mail address.

3.3.2. BEST PRACTICES FOR OPTIONAL ELEMENTS OF THE IDENTIFY RESPONSE

`<compression>`

See the protocol specifications on record compression in Section 3.1.3:
http://www.openarchives.org/OAI/openarchivesprotocol.html#ResponseCompression.

See Section 7 of the Implementation Guidelines for Repository Implementers:
http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#ResponseCompression.

The compression encoding supported by the repository. The recommended values are those defined for the Content-Encoding header in Section 14.11 of RFC 2616 describing HTTP 1.1.[9]

It is a best practice that the OAI repository includes the encodings it supports in this element.

`<description>`

The description container allows data providers to describe various aspects of the OAI repository. As stated in the OAI PMH specifications, each description container must be accompanied by the URL of an XML schema describing the structure of the description container. There are several such schemas available and described within Section 3.1 of the *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting:* **http://www.openarchives.org/OAI/2.0/guidelines.htm**. These are described more fully below, but it should be noted that the description container is extensible, and other schemas may be used as long as an XML schema is in use.

[9] **http://www.faqs.org/rfcs/rfc2616.html**

3.3.2.1. BEST PRACTICES FOR USE OF THE DESCRIPTION CONTAINER

Repository Description:

It is a best practice to include a description of the OAI repository within the Identify response, particularly if the repository does not support sets or does not include set descriptions (see Section 3.10 "Best Practices for Sets" for more information). The repository description should include such information as the type of resources described by the metadata in the repository, how often the repository is updated, documentation about metadata practices, the number of items in the repository, etc. This type of administrative information is often important for service providers in determining whether and how often to harvest an OAI repository and what post-harvest processing to perform on the metadata. It is possible to provide such a description using the unqualified Dublin Core schema. Community-specific guidelines were developed for ePrint repositories.[10] Other communities such as the Open Language Archives Community (OLAC) have developed their own schema to describe repositories.[11]

This is an example of an ePrint description:

```
<description>
      <eprints
      xsi:schemaLocation="http://www.openarchives.org/OAI/1.1/eprints
      http://www.openarchives.org/OAI/1.1/eprints.xsd">
            <content>
                  <URL>http://cds.cern.ch/</URL>
            </content>
            <metadataPolicy>
                  <text>Free and unlimited use by anybody with obligation to
                  refer to original record</text>
            </metadataPolicy>
            <dataPolicy>
                  <text>Full content, i.e., preprints may not be harvested by
                  robots</text>
            </dataPolicy>
            <submissionPolicy>
                  <text>Submission restricted. Submitted documents are subject
                  to approval by OAI repository admins.</text>
            </submissionPolicy>
      </eprints>
</description>
```

Friends:

An OAI repository can refer to related repositories by using the friends schema in the `<friends>` container.[12] Use of this container allows service providers to find other repositories that might be of interest. The use of the `<friends>` container is erratic; as of May 2007 only about 11 percent of active repositories included this container in their Identify response. As such, these best practices are neutral about the value of including it.

[10] **http://www.openarchives.org/OAI/1.1/eprints.xsd**

[11] **http://www.language-archives.org/OLAC/1.0/olac-archive.xsd**

[12] **http://www.openarchives.org/OAI/2.0/friends.xsd**

This is an example of a friends container from a National Aeronautic and Space Administration (NASA) repository:

```
<description>
      <friends
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/friends/
      http://www.openarchives.org/OAI/2.0/friends.xsd">
            <baseURL>http://techreports.larc.nasa.gov/ltrs/oai2.0/</baseURL>
            <baseURL>http://naca.central.cranfield.ac.uk/cgi-bin/nph-oai.cgi
            </baseURL>
            <baseURL>http://ston.jsc.nasa.gov/collections/TRS/oai/</baseURL>
            <baseURL>http://trs.nis.nasa.gov/perl/oai/</baseURL>
            <baseURL>http://naca.larc.nasa.gov/oai2.0/</baseURL>
            <baseURL>http://eprints.riacs.edu/perl/oai/</baseURL>
            <baseURL>http://celestial.eprints.org/cgi-bin/oaia2/arXiv.org
            </baseURL>
            <baseURL>http://www.biomedcentral.com/oai/2.0/</baseURL>
            <baseURL>http://www-dev.osti.gov/oai2.0/</baseURL>
            <baseURL>http://genesis2.jpl.nasa.gov/perl/oai</baseURL>
            <baseURL>http://www.giss.nasa.gov/cgi-bin/gpol2/</baseURL>
      </friends>
</description>
```

OAI Identifier:

See Section 3.4.1, "Best Practices for the OAI Identifier Description Container" in this document.

Branding:

See the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *XML Schema to Hold Branding Information for Collections:* **http://www.openarchives.org/OAI/2.0/guidelines-branding.htm**.

A service provider should always include information in whatever service it builds about the data provider that provided the metadata. This may occur at the aggregation or record level. Branding the record or group of records should clearly indicate to the user where a record came from and potentially add to the trust and confidence in that record (though this has not been well tested). From the point of view of the data provider, branding provides an opportunity for marketing and establishment of reputation. While this can be achieved by simply noting the name of the institution, it is also possible to provide an icon in the branding container at either the repository level (in the Identify response) or the set level. In both cases, the branding schema should be used.[13]

Service providers may choose to ignore the branding information; data providers may want to contact likely service providers to see whether or not such branding is useful. Use of the `<branding>` container is very low; as of May 2007, less than 1 percent of data providers give any branding information, and few service providers display the branding when provided. As such, these best practices are neutral about the value of including it.

If used, it is recommended that the icon be 88 pixels wide by 31 pixels, as is standard practice within the RSS community. The icon should be used within a data provider's local environment; this will facilitate the brand recognition for users. A data provider may provide both repository- and set-level branding, but should be aware that a service provider may choose one icon over the other; it is good practice to provide branding at either the set or the repository level, but not both.

[15] **http://www.openarchives.org/OAI/2.0/branding.xsd**

The following is a branding example from arXiv.org that is found in their Identify response:

```
<description>
      <branding
      xmlns="http://www.openarchives.org/OAI/2.0/branding/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/branding/
      http://www.openarchives.org/OAI/2.0/branding.xsd">
            <collectionIcon>
                  <url>http://arxiv.org/OAI/arXivLogo.png</url>
                  <link>http://arxiv.org/</link>
                  <title>arXiv e-print archive</title>
                  <width>88</width>
                  <height>31</height>
            </collectionIcon>
      </branding>
</description>
```

This would typically be expressed in HTML as

```
<a href="http://arxiv.org/">
      <img src="http://arxiv.org/OAI/arXivLogo.png" alt="arXiv e-print archive"
      width="88" height="33"/>
</a>
```

Toolkit:

A data provider may include a description of the toolkit used in the OAI repository using the toolkit schema.[14] This information may be useful to service providers in determining the technical capabilities of the OAI repository. It is a best practice to include a toolkit description if at all possible.

This is an example of the toolkit description that comes with the OAICat data provider software.

```
<description>
      <toolkit
      xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit
      http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd"
      xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit">
            <title>OCLC's OAICat Repository Framework</title>
            <author>
            <name>Jeffrey A. Young</name>
            <email>jyoung@oclc.org</email>
            <institution>OCLC</institution>
            </author>
            <version>1.5.42</version>
```

[14] **http://oai.dlib.vt.edu/OAI/metadata/toolkit.xsd**

```
            <toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif
            </toolkitIcon>

            <URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>

      </toolkit>

</description>
```

Rights:
The data provider may include a rights manifest statement at the OAI repository level. See Section 3.11.1, "Best Practices for Expressing Rights over Metadata" in this document for more information.

3.4. BEST PRACTICES FOR OAI IDENTIFIERS

See the protocol specifications on OAI identifiers:
http://www.openarchives.org/OAI/openarchivesprotocol.html#UniqueIdentifier.

See also the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Specification and XML Schema for the OAI Identifier Format:* **http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm**.

The OAI-PMH 2.0 specification requires that each item in the repository have a unique identifier `<oai-identifier>`. The OAI identifier is specific to the item from which records are disseminated. For example, an item may have metadata available for harvest in both unqualified Dublin Core and MARC21. Both of these records will have the same unique OAI identifier, but will differ in their metadataPrefix (oai_dc and marc21 respectively) and may differ in their datestamp. Note that an OAI identifier does NOT refer to the identifier for a resource (for example, what might be contained within the `<dc:identifier>` element in a metadata record).

The OAI identifier for a specific repository item should not change over time for the same object. If the OAI identifier for a specific item must be changed, it should never be reused for a different item. If an item is deleted, its OAI identifier should also be marked deleted. See Section 3.6, "Best Practices for Deleted Records" for a full discussion.

The OAI identifier for any item should not exceed 128 characters to be efficiently handled by all kinds of databases and file systems. Although not specified in the protocol, the length of the OAI identifier might affect processing of the record by the service provider.

Unless already using an established URI schema, OAI repositories should conform to the *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting:* Specification and XML Schema for the OAI Identifier Format. Use of this specification will mean that the repository identifiers are globally unique within the oai namespace. The advantage for repositories of adopting this naming convention is that record identifiers are resolvable via future resolution services or services such as OCLC's Extensible Repository Resource Locators (ERRoLs) for OAI Identifiers.[15]

3.4.1. BEST PRACTICES FOR THE OAI IDENTIFIER DESCRIPTION CONTAINER

If OAI data repositories implement the OAI Identifier format discussed above, they should expose their compliance with the `<oai-identifier>` format by including a `<description>` container in their Identify response. For example,

```
<description>

      <oai-identifier
      xmlns="http://www.openarchives.org/OAI/2.0/oai-identifier"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai-identifier
      http://www.openarchives.org/OAI/2.0/oai-identifier.xsd">

            <scheme>oai</scheme>

            <repositoryIdentifier>arXiv.org</repositoryIdentifier>
```

[15] **http://www.oclc.org/research/projects/oairesolver/default.htm**

```
            <delimiter>:</delimiter>
            <sampleIdentifier>oai:arXiv.org:quant-ph/9901001</sampleIdentifier>
      </oai-identifier>
</description>
```

The <repositoryIdentifier> must be an Internet domain name (not a literal numeric IP address) that is registered to the organization that controls the OAI repository. Best practice is to use the domain name where the OAI service itself resides. For example, if the base URL of the OAI data repository is **http://oai.some.edu/path/one/oai.asp**, then the repository identifier should be "oai.some.edu". If multiple OAI providers come from the same domain, it is acceptable to create new domains specifically for use as identifiers. For example, the base URL for two providers might be **http://oai.some.edu/path/one/oai.asp** and **http://oai.some.edu/path/two/oai.asp**, so their repository identifiers could be "one.oai.some.edu" and "two.oai.some.edu".

It is possible for the base URL of a repository to change over time. However, the repository identifier should never change once it is established, so later in the life of an OAI repository, the domain of the base URL may differ from the <repositoryIdentifier>.

OAI identifiers, and thus repository identifiers, are case sensitive (even though Internet domain names are not). Therefore, the best practice is to always use all lower case for repository identifiers.

For the <sampleIdentifier> it is a best practice to use an actual real identifier as the value.

3.5. BEST PRACTICES FOR DATESTAMPS

See the protocol specifications on datestamps and selective harvesting:
http://www.openarchives.org/OAI/openarchivesprotocol.html#SelectiveHarvestingandDatestamps

See the protocol specifications on UTC datetime: **http://www.openarchives.org/OAI/openarchivesprotocol.html#Dates**

The purpose of datestamps in the OAI PMH is to support incremental or selective harvesting. Each metadata record in a repository must have a datestamp. If more than one record is available from an item (perhaps an unqualified Dublin Core record and a MARC21 record), then the datestamps may change independently. The datestamp appears in the <header> of an OAI record.

Datestamps allow a service provider to keep an up-to-date copy of metadata from a repository by periodically harvesting only those records that have changed since a particular date and time. Such incremental harvesting addresses a scalability issue by providing an alternative to completely reharvesting metadata from a repository. Datestamps are not to be confused with the dates that may be included within metadata records, for example within the <dc:date> element of a Dublin Core record.

It is a best practice for data providers to include accurate and updated datestamps in their OAI repository. Only changes to the underlying item that have no effect on the OAI record should go unrecorded. If a service provider is performing incremental harvests, updated, added, and deleted records will be harvested only if their datestamps accurately reflect the time the record was added, updated, or deleted in the OAI repository.

It is a best practice for the datestamp to reflect the date and time at which a record change was actually made available from the OAI repository. If, for example, a particular institution edits the metadata items on 2005-08-10, and then makes available the OAI records on the next day (2005-08-11), the datestamp should correspond to 2005-08-11 rather than 2005-08-10. This allows a service provider to accurately harvest changed records. Datestamps must never be backdated because that might result in the change being missed by an incremental harvest.

Repositories may record datestamps with either date, or date and time, precision. This is referred to as the datestamp "granularity." The granularity must be used consistently for all records within a repository and must be declared in the <granularity> element of the Identify response (see Section 3.3 of this document). To avoid problems with comparison of datestamps from around the world, they must always be specified in Coordinated Universal Time (UTC). It is a best practice that repositories use seconds granularity where practical. This allows a service provider to incrementally harvest to the finest specificity.

3.6. BEST PRACTICES FOR DELETED RECORDS

See the protocol specification on deleted records:
http://www.openarchives.org/OAI/openarchivesprotocol.html#DeletedRecords.

The OAI PMH states that repositories must declare one of three levels of support for deleted records in the `<deletedRecord>` element of the Identify response:

- `<deletedRecord>no</deletedRecord>`
- `<deletedRecord>persistent</deletedRecord>`
- `<deletedRecord>transient</deletedRecord>`

If there is support for deleted records (whether persistent or transient) and an OAI record is deleted, the datestamp must be the date and time that the record was deleted, and the OAI header must contain an attribute `status="deleted"` (i.e., `<header status="deleted">`) and must not include metadata or about containers. The OAI header might look like this:

```
<header status="deleted">
      <identifier>oai:arXiv.org:hep-th/9801010</identifier>
      <datestamp>1999-02-23</datestamp>
      <setSpec>physic:hep</setSpec>
      <setSpec>math</setSpec>
</header>
```

Note that the deleted status is a property of an OAI record. If, for example, a repository provides records in oai_dc, oai_marc, and marc21 and decides to delete the records in oai_marc, it can mark these records deleted without affecting the records in oai_dc and marc21. But if the item (with the unique id) is deleted, the records disseminated in all three metadata formats should be marked deleted.

It is a best practice is to support persistent information about deleted records. This allows service providers to track which records have been deleted and purge the appropriate records from their service. Maintenance of persistent information about deleted records has an added benefit of helping to ensure that a repository does not reuse OAI identifiers because they continue to be used for the deleted record.

To appreciate this best practice, consider the impact of not supporting persistent information about deleted records. The OAI PMH is designed to support incremental harvesting by service providers by datestamps.[16] This means that service providers are able to harvest only those records that have been added, modified, or deleted since the last harvest of the repository. To determine whether a record has been added, modified, or deleted since the last harvest, the protocol relies on a change in the datestamp (see Section 3.5, "Best Practices for Datestamps" in this document). Because of this reliance on datestamps, if a repository does not support information (whether persistent or transient) about deleted records and does, in fact, delete records, a service provider conducting an incremental harvest has no way to know that these records have been deleted. This has two specific implications:

- The records that have been deleted from the data provider's repository will still appear within the service provider's end product.

  Example: Repository A has decided to withdraw access to 100 digitized photographs because of a copyright dispute. It deletes the metadata items from its database, and the records are no longer disseminated via its OAI repository. Repository A does not support information about deleted records. When Service Provider 1 incrementally harvests Repository A for records that have changed since the last harvest, the ListRecords request (with a from and until argument) returns only new records added since the last harvest. Service Provider 1 adds the new records to its service but does not make any other changes. An end-user searching Service Provider 1's database finds an interesting metadata record, clicks the URL pointing to the content described by the metadata, but is directed to an error page because that content has been deleted. This reflects badly both on Service Provider 1 and Repository A.

[16] See the protocol specification on selective harvesting for more information:
http://www.openarchives.org/OAI/openarchivesprotocol.html#SelectiveHarvesting

- The service provider will need to conduct regular full harvests to ensure that its data match the repository.

  While conducting periodic full harvests is generally good practice for service providers, conducting regular full harvests can present a scalability problem. If a repository does not support information about deleted records, regular full harvests are, unfortunately, essential. Full harvests are the only way to ensure that the records the service provider has match the records the OAI repository is providing.

If maintenance of persistent information about deleted records is not possible, OAI repositories should consistently maintain transient information about deleted records. Transient information should be maintained for a minimum of six months to allow service providers that harvest sites irregularly to harvest the information about deleted records. If a repository supports persistent or transient information for deleted records, a service provider will be able to purge the appropriate records from its database or service and maintain greater consistency with the original repository.

As of May 2007, 36 percent (494) of 2.0-compliant sites support *persistent* information about deleted records, 10 percent (138) of 2.0-compliant sites support *transient* information about deleted records, and 54 percent (727) of 2.0-compliant sites *do not support* deleted records.

3.7. BEST PRACTICES FOR RESUMPTION TOKENS

See the protocol specifications on flow control and use of the resumptionToken:
http://www.openarchives.org/OAI/openarchivesprotocol.html#FlowControl.

See Section 5 of the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Guidelines for Repository Implementers on Incomplete-List Responses and the Use of resumptionToken Elements:*
http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#resumptionToken.

Data providers may also be interested in the section on Flow Control and Load Balancing in the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Guidelines for Repository Implementers:*
http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#FlowControlAndLoadBalancing.

Resumption tokens are an option within the OAI protocol that allows data providers to institute a measure of flow control. Resumption tokens allow data providers to chunk responses to list requests, specifically the `ListRecords`, `ListIdentifiers`, and ListSets requests. An empty `<resumptionToken>` attribute must be included to indicate that the list request is complete.

The entities contained in an incomplete list response (i.e., the OAI records, headers, or set) must be intact and complete. A repository may not issue half of an OAI record in one incomplete list response and the second half in the next.

A further requirement is that a repository must be able to accept the same resumption token more than once and return the same response (that is, the contents of the chunk are the same). The only time when this may not be the case is when records in the complete list request have been added, modified, or deleted and are thus out of the datestamp range of the initial request.

It is a best practice for resumption tokens to be implemented in any sizeable (more than 2 MB) OAI repository. Implementation of resumption tokens is beneficial for both the data provider (limits the load on its server) and for the service provider (can process responses in reasonable chunks). The incomplete list size should be reasonable—ideal response size is probably between 0.5 and 2 MB—though this may depend on the capabilities of both the data provider and the service provider. If responses are too big, they may be difficult to retrieve reliably via HTTP. If responses are too small, a great deal of extra network traffic is required to harvest a repository's records. The number of records returned per resumption token will often depend on the size of the metadata, not necessarily the number of records.

A data provider does not need to format a resumption token so that it is understandable to a service provider.

A data provider should decide whether or not to allow resumption tokens to expire (the best practices are neutral on this issue). However, if resumption tokens do expire, it is a best practice to include the `expirationDate` attribute in the resumption token. Expiration dates should be set to allow harvesters adequate time to process the last incomplete list received. The *Implementation Guidelines* recommend that expiration dates be valid for at least tens of minutes.

It is a best practice to include both the `completeListSize` and `cursor` attributes. The `completeListSize` attribute is often the only place where there is an indication of the total number of records that will be included in the complete response, and thus is a useful indicator of the size of the OAI repository. This information might also be recorded in the Identify response as well as set descriptions.

Resumption tokens in the response should not be URL encoded. This is different from an OAI request, in which resumption tokens MUST be URL encoded. It is a best practice not to use characters in resumption tokens that require URL encoding.

It is a best practice to issue a `badResumptionToken` to stop the sequence of requests if a repository cannot continue to complete a request. A harvester is expected to start again from the initial list request. This might occur for a variety of reasons, including if significant changes, additions, or deletions have been made to a repository that would affect the idempotency of the resumption token. It is a best practice to include human-readable text that explains the `badResumptionToken` response (for example: The resumptionToken has expired).

This is an example of a resumption token from the **http://hal.ccsd.cnrs.fr/oai/oai.php** repository:

```
<resumptionToken expirationDate="2005-07-26T16:57:24Z"
completeListSize="31979" cursor="4">lr42e519f4d1e58</resumptionToken>
```

3.8. BEST PRACTICES FOR PROPER MIME-TYPING

Server software, browsers, and the HTTP standard utilize and rely on the designation of "content types" (or MIME types) to make successful use of shared data on the Internet. For the delivery of OAI PMH–compliant XML, the OAI PMH requires the MIME-type setting to be text/xml. Usage of any other MIME type, such as text/plain, text/html, etc., is not acceptable.

Most out-of-the-box OAI data provider toolkits should handle this typing automatically. Similarly, recent versions of server applications like Apache HTTP server, Apache Tomcat, JBoss, and Microsoft IIS server all automatically recognize common XML file extensions (such as .xml, .xsl, etc.) as text/xml files. For toolkits and applications that do not—or for data providers that create their own OAI tools locally—the MIME-type setting should be set to text/xml to ensure compliance with the protocol. Setting the MIME type to text/xml ensures that XML and OAI-compliant tools, such as harvesters, can adequately retrieve and process the data of interest.

An easy way to see if your server is delivering XML as text/xml is to use a fifth-generation or newer browser to retrieve a given instance (i.e., performing a GetRecord request). If the data come back in the browser looking like a plain text file, odds are the MIME type is set to text/plain. However, if the XML has been formatted in different colors, or the data has collapsible/navigable points, it is likely being served as text/xml. Better yet, more technical solutions include the following:

- Looking at the "Page Info" or "Properties" utilities provided by most browsers.
- Checking your Web server logs.
- Viewing the HTTP header responses delivered to your Web browser. The most common browsers have plug-ins or add-ons that allow you to view these data.

3.9. BEST PRACTICES FOR HTTP SERVER RESPONSES

See the protocol specifications on OAI PMH error responses:
http://www.openarchives.org/OAI/openarchivesprotocol.html#ErrorConditions.

See the protocol specifications on HTTP Status codes:
http://www.openarchives.org/OAI/openarchivesprotocol.html#StatusCodes.

Servers implementing the OAI PMH must conform to HTTP status code definitions and report relevant HTTP transport layer status via those status codes.

It is a best practice to supply useful HTTP messages that convey information about the availability of your repository. For HTTP messages used when redirecting to new locations (HTTP 301), load balancing or throttling high traffic rates (HTTP 302), etc., be sure to use the "location" HTTP field to convey how the service provider should proceed.

Most Web server applications provide the ability to deliver custom HTTP error messages based on the requests for resources they receive. It is a best practice to make use of these capabilities. Data providers should convey the fullest level of information possible to allow others to determine what the nature of the problem is. Also, it is recommended that data providers share when the repository will be available again for usage.

3.10. BEST PRACTICES FOR SETS

See the protocol specifications for sets: **http://www.openarchives.org/OAI/openarchivesprotocol.html#Set**.

See the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Guidelines for Repository Implementers on Sets:* **http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#Sets**.

The protocol defines a set as "an optional construct for grouping items for the purpose of selective harvesting." Set structure may be flat, hierarchical, or overlapping. Items may belong to more than one set. Each set must have a `<setSpec>` that is used within OAI PMH requests and a `<setName>` that is a human-readable string. Each set may also have a `<setDescription>`, which is an optional and repeatable container that can hold XML-encoded data about the set. If an OAI repository has implemented sets, a service provider may still choose to harvest all records (ListRecords) or identifiers (ListIdentifiers) without reference to the sets. A data provider may not require that its repository be harvested only by set.

The best practice for use of sets is, in many ways, an extremely fluid and evolving area. This section attempts to distinguish between what is an accepted best practice, and what is an area that is either still under discussion or may be dependent on the domain of which a data provider is part.

Service providers can use sets to selectively harvest sets of records that are appropriate for their service. For example, the National Science Digital Library might harvest only those sets that include science-, engineering-, and mathematics-oriented records. The IMLS Digital Content Gateway might harvest only those sets and repositories that include metadata describing material in collections that have been funded by the Institute of Museum and Library Services. Sets can help a service provider avoid harvesting an entire OAI repository, which can reduce the workload on the data provider's server.

Sets and set descriptions have the potential to aid service providers further. The use of set descriptions in particular can provide both technical and descriptive information about the set of metadata records. This information might help a service provider not only determine whether to harvest a set, but also how often to incrementally harvest or reharvest the records and what metadata formats the records are available in. In addition, the information potentially could be used to provide contextual information about the resources described by the metadata to the end-user.

Service providers can run into a number of obstacles when using sets to selectively harvest:

- Interpreting how a repository has organized sets and determining which sets to harvest.

  At issue here is that the `<setName>` is not human understandable and/or there is no `<setDescription>` provided. Set name (because it is required) is often used to interpret the content of the set and whether its items should be harvested. However, like the use of the `<description>` container in the Identify response (see Section 3.3.2.1. of this document), much more detailed information could be recorded in the set description, including how often the set is updated, how many records it contains, etc.

  Also at issue may be the large number of sets to sort through. As of May 2007 the largest number of sets in a single OAI repository was 8,699. The average number of sets was 97, and the median number of sets was 8.

- Knowing when there are records that belong to no sets.

  The OAI PMH allows items to belong to no sets. However, there is no standard way for the data provider to communicate this to service providers.

- Knowing when there are empty sets.

  Data providers can expose sets with no records. This happens particularly in data providers that are built into digital library systems, such as DSpace. When a collection is created in DSpace, a corresponding OAI PMH set is created. Again, there is no standard way for a data provider to communicate this to service providers.

- Understanding relationships between sets.

  While there is a mechanism to express relationships between hierarchical sets, there is no mechanism to express relationships between overlapping sets. The only way to know which items belong to multiple nonhierarchical sets is to harvest the identifiers or records that contain the header information.

- Knowing when a set structure has been substantially changed.

  Data providers may not communicate when there is a major change in the set structure, as suggested in Section 3.2, "Managing the Repository Lifecycle" in this document.

Many of these obstacles point to a need for data providers to include documentation about sets and a need for data and service providers to communicate with one another around set issues.

3.10.1. WHEN SETS SHOULD BE USED

The use of sets within an OAI repository is optional; it is simply a way to organize metadata to support selective harvesting by service providers. Because service providers have many different criteria for what metadata they want to harvest and how they store and manipulate metadata once harvested, it is unlikely that all sets will be useful for all service providers. For this reason, if a data provider wants to be included in a specific harvesting service, it should communicate with the service provider to understand (1) whether sets would be beneficial for the service provider, and (2) if so, what sort of organization would be most useful.

That said, if the OAI repository contains a large number of items describing resources from a variety of different collections, or if a data provider wishes to distinguish between one group of items and another group, it is a best practice to use the set concept in the OAI PMH to do this.

3.10.2. HOW ITEMS SHOULD BE ORGANIZED INTO SETS

The protocol does not specify how sets should be defined, organized, or arranged within a repository, and current practice varies widely. Sets are organized according to any number of factors: subject, type of material, a traditional library or museum collection, publication status, originating institution, internal departments, and access restrictions. Set organization may be dictated by the program or software used by the data provider as well as by internal workflow and organization.

If a data provider wants to be included in a specific harvesting service, it should communicate with that service provider to understand what sort of organization of sets is most useful to the service provider. In addition, the data provider may want to understand how other data providers in its domain are organizing their OAI repositories. The Deutsche Initiative für Netzwerkinformation (DINI), for example, has made some recommendations for set organization for the higher education community in Germany that suggest using subject classification, publication type, and/or document type as possible organizing structures (DINI 2003).

There is no single best practice for the organization of sets. The most realistic recommendation is that data providers organize sets in a way that best meets the needs of their primary service provider *and* can be easily done within their own internal workflow.

It can be useful to organize the metadata items into sets according to the collections of resources they represent. The concept of collections might vary considerably. However, the internal conception of collections that data providers have is usually helpful to service providers and allows them to selectively harvest. Many times, the concept of collection actually corresponds to a unity of topics, material type, and person in charge of the original collection. Most of the time each collection has its own homepage in the data provider's local system. Collection information can help provide crucial contextual information to supplement the metadata within the set.

3.10.2.1. RELATIONSHIPS BETWEEN SETS

Set structure can be flat, strictly hierarchical, or overlapping. Items may belong to more than one set; this is typically indicated in the `<header>` container of an OAI record. For example,

```
<header>
      <identifier>oai:lcoa1.loc.gov:loc.pnp/cph.3a01895</identifier>
      <datestamp>2005-01-07T21:13:30Z</datestamp>
```

```
      <setSpec>lcmiscvis</setSpec>
      <setSpec>app</setSpec>
</header>
```

Independent of what is in the <header> container of an OAI record, hierarchical sets can also be indicated through the use of colons in the <setSpec>. For example, <setSpec>Functions:Algebra</setSpec> indicates that there is a <setSpec>Functions</setSpec> set that includes <setSpec>Functions:Algebra</setSpec> which can be harvested separately from <setSpec>Functions<setSpec>. However, not all data providers follow this convention. If sets are hierarchical, it is a best practice to use the : convention indicated in the OAI PMH. If there are hierarchical sets and the : convention is not used, it is a best practice to document the relationship between sets in the set description.

Sets may also be overlapping, but not strictly hierarchical. For example, the University of Michigan repository has organized its sets by collection (all records within a set describe resources belonging to a specific collection), and has provided specific sets for the use of specific service providers (e.g., it also provides a set corresponding to all scientific material within its collections).[17] There is no formalized way to indicate overlapping sets within the protocol as there is with hierarchical sets. It is a best practice to document overlapping sets in the set description (discussed further below).

3.10.2.2. EMPTY SETS

Whenever possible, data providers should not expose empty sets. In some cases exposure of empty sets is a function of the data provider software in use, but whenever possible these should be suppressed in the actual OAI repository.

3.10.2.3. SET MEMBERSHIP

Within the OAI PMH, items belong to sets. The implications of this is that a set may contain items that are disseminated in only one metadata format and other items that are disseminated in two or more metadata formats, depending on the data provider implementation. There is no formal way within the OAI protocol to indicate what metadata formats the items in a set are available in. It is a best practice that items within sets should all be able to be disseminated in the same metadata format. Items that are disseminated in different metadata formats should be organized into distinct sets. For example, if a data provider generally exposes all items in a set in unqualified Dublin Core, MARC, and MODS, but has a new batch of items to add that are only available in unqualified Dublin Core and MODS, these items should be placed into a new set rather than added to the other one. If this is not possible, the data provider should add explicit documentation to the set description noting the inconsistency in metadata format availability.

If a repository has implemented sets, it is a best practice for all items to belong to at least one set; that is, there should be no items which do not belong to any set. A general set can be created for items that cannot be categorized, for example, "All uncategorized technical reports." If a data provider includes items that do not belong to a set, it is useful to note this information in the repository description within the Identify response. Service providers may harvest only by set and may never realize that additional records are available unless such information is noted.

3.10.3. HOW SETS SHOULD BE DESCRIBED

There are two main ways for a repository to provide some descriptive information about its sets.

3.10.3.1. SET NAME

It is a best practice that the <setName> should not only be human readable, but human understandable. Set names such as MWB01 or c72 may have some meaning to a data provider, but they will not to the repository's primary audience, the service provider. Service providers often rely on the set name to interpret the set, particularly if the <setSpec> is not human readable/understandable or if there is no set description.

[17] See **http://quod.lib.umich.ed/cgi/b/broker20/broker20/?verb=ListSets**.

[18] **http://gita.grainger.uiuc.edu/registry/searchform.asp**

3.10.3.2. SET DESCRIPTION

If a repository has sets implemented, it is a best practice to include a <setDescription>. This allows service providers to understand what items are in a set and how it is organized. It also allows registries such as the UIUC registry to enable better keyword searching from responses to OAI requests.[18] However, what precisely should be included within the <setDescription> container is still an open issue.

As of October 1, 2005, 227 (22 percent) of OAI data providers that used sets included a <setDescription> container. All but one used unqualified Dublin Core for the set description.

An analysis conducted in October 2005 of 109 repositories using the set description revealed the following rough breakdown of how it is used:

- A link to the top page of the journal issue containing the records in the set (dc:description element only): 49 (45 percent)
- Description of the resources described by the metadata contained in the set (dc:description element only): 45 (41 percent)
- <setDescription> included but is either empty or contains the letter c: 9 (8 percent)
- Title or publisher of the collection of resources described by the metadata contained in the set and a URL to the collection homepage: 3 (3 percent)
- Description of the resources described by the metadata contained in the set (more than dc:description): 2 (5 percent)
- Record count only: 1 (1 percent)

Note that the majority of the repositories that use the <setDescription> container describe the collection of resources represented by the metadata contained in the set, but do not strictly describe the set itself as a collection of metadata records. Of these 109 data providers, only 5 (5 percent) included descriptive, technical information about the set itself (e.g., metadata formats available).

The distinction between the description of the resources represented by the metadata in the set and the description of the set itself as a collection of metadata records and as a technical mechanism in and of itself is an important one. Service providers often need both types of information; knowing the type of resources that are represented by the metadata in the set is important for selection purposes, while knowing what metadata formats are available within a specific set and how often it is updated is very useful to production harvesting activities, particularly if that information is encoded in a machine-readable format.

The working group recommends making a distinction between set description (as in the collection of metadata records themselves and the set's relationship to other sets) and collection description (as in the collection of resources represented by the metadata records in the set). However, placement of set description information and how it is associated with collection description of the resources is not by any means resolved. A discussion paper developed at UIUC (in consultation with the working group but not strictly under its aegis) proposed a method for handling set descriptions, but it has not been developed further (Foulonneau and Shreeves 2005).

The following is useful information to include in a <setDescription> container:

- The formats in which the data can be harvested

  For a repository that offers sets using different metadata formats, the data provider should indicate which metadata format(s) the records in the set adhere to. It is important for service providers interested in harvesting more than unqualified Dublin Core to know which formats are available for each set. As stated above, there is no formal way in the protocol to do this.

- Accrual periodicity

  Identifying the frequency with which items are added to a collection can help a service provider schedule harvests.

- Access rights if valid for all records contained in the set

  This, however, should be repeated within the individual OAI record. Rights should be expressed using the <rightsManifest> package. Note that this is specifically for rights over the metadata, not the content described by the metadata. See Section 3.11.1, "Best Practices for Expressing Rights over Metadata" in this document.

- The relation to other sets

  If sets overlap, the data provider could indicate how the sets interact by using <dcterm:isPartOf> or another appropriate metadata element in the set description.

- An approximate number of records contained in the set, if the set is relatively static.

  This can also be encoded in the resumption token (see Section 3.7, "Best Practices for Resumption Tokens").

- Any branding information for the set using the <branding> container

  See the subsection on branding in Section 3.3.2.1, "Best Practices for Use of the Description Container" in this document.

In addition, if the metadata within the set represents a collection (however defined), a collection description can be very helpful for service providers for selection purposes. For example, a service provider may want to harvest only metadata records focused on American history; it may be possible to use collection descriptions to determine which sets to harvest. However, it is important to not use the set description to convey essential information for the individual items within a set. While some institutions (notably UIUC) have experimented with set descriptions to convey essential context to the actual aggregation, because set descriptions cannot be reliably harvested within the OAI PMH (i.e., set descriptions do not have OAI identifiers or datestamps), they are most useful—at least for now—for informational purposes (Foulonneau et al. 2005).

In all cases, standard encoding schemes and/or controlled vocabularies should be used if possible to facilitate machine processing of the information.

The following are two examples of set descriptions.

A set description may be a simple <dc:description> field such as this one (in bold) in the Library of Congress American Memory data provider. This is embedded in a much more extensive description of the resources represented by the metadata in the set:

```
<setDescription>
      <oai_dc:dc
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
      http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title xml:lang="en">Records for California As I Saw It
            (books)</dc:title>
            <dc:creator>Library of Congress</dc:creator>
            <dc:description>Transcribed text with illustrations of 186 works
            documenting the formative era of California's history through
            eyewitness accounts. The collection covers the dramatic decades
            between the Gold Rush and the turn of the twentieth century. It
            captures the pioneer experience; encounters between Anglo-Americans
            and the diverse peoples who had preceded them; the transformation
            of the land by mining, ranching, agriculture, and urban
            development; the often-turbulent growth of communities and cities;
            and California's emergence as both a state and a place of uniquely
            American dreams.</dc:description>
            <dc:description>Set characteristics for calbkbib: Source records
            are MARC (from LC catalog); MODS or oai_dc records are dynamically
            generated using generic transformation when harvested. dct:
            accrualPolicy: Closed. Contains about 200 records. Records in set
            calbkbib are also in set lcbooks.</dc:description>
            <dc:type xml:lang="en">text</dc:type>
            <dc:type xml:lang="en">books</dc:type>
            <dc:type xml:lang="en">printed materials</dc:type>
            <dc:type xml:lang="en">collection</dc:type>
            <dc:coverage xml:lang="en">1849-1900</dc:coverage>
```

```
            <dc:coverage xml:lang="en">California</dc:coverage>
            <dc:subject xml:lang="en">Ethnic groups--California.</dc:subject>
            <dc:subject xml:lang="en">Law and politics--California.
            </dc:subject>
            <dc:subject xml:lang="en">California--History.</dc:subject>
            <dc:subject xml:lang="en">California--Biography.</dc:subject>
            <dc:subject xml:lang="en">California--Gold discoveries.
            </dc:subject>
            <dc:contributor
            xml:lang="en">Library of Congress, General Collections
            </dc:contributor>
            <dc:rights
            xml:lang="en">http://memory.loc.gov/ammem/cbhtml/cbres.html
            </dc:rights>;
            <dc:relation
            xml:lang="en">http://memory.loc.gov/ammem/cbhtml/cbhome.html
            </dc:relation>
      </oai_dc:dc>
</setDescription>
```

The following example is a sample record from the Michigan State University repository. It attempts to distinguish between the description of the set and the description of the collection itself. Note that no particular collection description schema has been named here since this is still evolving. See the Dublin Core Collections Application Profile[19] for the closest to canonical schema.

```
<setDescription>
      <cld:set>
            <dc:identifier>lib.msu.edu:fap</dc:identifier>
            <dc:title>Feeding America: The Historic American Cookbook Project
            </dc:title>
            <dc:format>oai_dc</dc:format>
            <dcterm:accrualPeriodicity>Infrequently</dcterm:accrualPeriodicity>
            <dcterm:extent>77 records</dcterm:extent>
            <dcterm:accessRights>No restrictions</dcterm:accessRights>
            <dc:rights>No restrictions</dc:rights>
            <dcterm:abstract>Records for cookbooks digitized as part of the
            Feeding America: Historic American Cookbooks Project
            </dcterm:abstract>
            <dc:description>Simple Dublin Core records created from
            documentation for each cookbook; oai_dc subject uses LCSH; oai_dc:
            type uses DCMIType vocabulary; oai_dc:coverage (1) uses TGN
            </dc:description>
            <dc:language xsi:type="dct:ISO639-2">eng</dc:language>
```

[19] **http://dublincore.org/groups/collections/collection-application-profile/**

```
            <cld:isLocatedAt>Michigan State University Libraries Digital &
            Multimedia Center</cld:isLocatedAt?>

            <cld:isAccessedVia xsi:type="dct:URI">
            http://oai.lib.msu.edu/OAIHandler </cld:isAccessedVia>

            <dcterm:references>

                  <cld:collection>

                        <dc:identifier>
                        http://digital.lib.msu.edu/projects/cookbooks</dc:
                        identifier>;

                        <dc:title>Feeding America: The Historic American
                        Cookbook Project</dc:title>

                        <dcterm:abstract>Online collection of some of the most
                        important and influential American cookbooks from the
                        late 18th to early 20th century.</dcterm:abstract>

                        <dcterm:extent>77 items</dcterm:extent>

                        <dc:language xsi:type="ISO639-2">eng</dc:language>

                        <dc:type xsi:type="cldtype">Collection of Texts
                        </dc:type>

                        <dcterm:accessRights>No restrictions
                        </dcterm:accessrights>

                        <dcterm:accrualPolicy xsi:type="DCCDAccrualPeriodicity?"
                        >Passive</dcterm:accrualPolicy>

                        <dc:subject xsi:type="LCSH">Cookery, American
                        </dc:subject>

                        <dc:subject xsi:type="LCSH">Cookery -- United States --
                        19th century</dc:subject>

                        <dc:subject xsi:type="LCSH">Cookery -- United States --
                        20th century</dc:subject>

                        <dc:creator>Michigan State University Libraries. Digital
                        & Multimedia Center.</dc:creator>

                        <cld:isLocatedAt>Michigan State University, 100 Library,
                        East Lansing, MI, 48224, USA</cld:isLocatedAt>

                        <cld:isAvailableVia
                        xsi:type="URI">http://digital.lib.msu.edu/projects/
                        cookbooks</cld:isAvailableVia>

                  <cld:collection>

            </dcterm:references>

      </cld:set>

</setDescription>
```

3.11. BEST PRACTICES FOR ABOUT CONTAINERS

See the protocol specification on about containers: **http://www.openarchives.org/OAI/openarchivesprotocol.html#Record**.

See Section 3.4 of the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: **http://www.openarchives.org/OAI/2.0/guidelines.htm**.

An optional piece of the OAI PMH that allows a repository to include self-referential metadata about the records themselves is the <about> container. It may have only one child element. If a data provider needs to provide multiple pieces of information about the item, the <about> container should be repeated within the record. For example, if there is included both a rights expression and a provenance expression, each of these appears in a separate <about> container. Note that the <about> container refers to the metadata record itself, not the resource.

There are two common uses of the <about> container: rights and provenance (discussed below). Other information may also be included; however, the contents must conform to an XML schema.

3.11.1. BEST PRACTICES FOR EXPRESSING RIGHTS OVER METADATA

See the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Conveying Rights Expressions about Metadata in the OAI PMH Framework:* **http://www.openarchives.org/OAI/2.0/guidelines-rights.htm**.

The Rights Expression Specification provides a general framework and an XML schema for stating rights about metadata records. It is a best practice to follow these guidelines in describing rights for metadata at the record, set, and repository level. Note that these guidelines are not for stating rights about the resources described in the metadata. Conversely, within the descriptive metadata section of the record, rights elements should describe only the resource itself and not its metadata.

If rights statements are included, it is a best practice to include them at the record level. Data providers should not assume that providing rights information at the repository level in a <description> container or at the set level in <setDescription> is sufficient.

One important caveat is that the OAI PMH does not provide a mechanism for enforcing restrictions or conditions that you place on the use of your descriptive metadata. The data provider should think carefully before attaching restrictions or conditions on the use by service providers of the descriptive metadata associated with items. Doing so imposes burdens on service providers that wish to harvest this metadata and incorporate it in a specific service, and therefore can restrict the exposure of resources gained by participating in OAI PMH.

Some data providers use the OAI PMH to surface metadata for harvesting in a more closed environment. In these cases, a rights statement should be provided relating to use of the descriptive metadata in the records.

The following example contains rights information about a metadata record:

```
<about>
      <rights
      xmlns="http://www.openarchives.org/OAI/2.0/rights/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/rights/ http://
      www.openarchives.org/OAI/2.0/rights.xsd">
            <rightsDefinition>
                  <oai_dc:dc
                  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
                  oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
                  xmlns="http://purl.org/dc/elements/1.1/"
                  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
                        <publisher>The University of Chicago Library</publisher>
```

```
            <rights>No rights to the use of these metadata are granted
            except by prior agreement.</rights>
            </oai_dc:dc>
        </rightsDefinition>
    </rights>
</about>
```

3.11.2. BEST PRACTICES FOR USE OF THE PROVENANCE CONTAINER

See the Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *XML Schema to Hold Provenance Information in the "about" Part of a Record:* **http://www.openarchives.org/OAI/2.0/guidelines-provenance.htm**.

The `<about>` container may contain provenance information. Data providers generally do not add `<provenance>` containers; they are usually added by service providers when they re-expose harvested metadata via OAI PMH. It is a best practice for service providers to add a provenance element to all harvested records that will be re-exposed through the OAI PMH to ensure traceability of metadata records.

There is a schema specifically for `<provenance>`.[20] The node should contain the harvesting history of the record, through embedded `<originDescription>` elements, and should indicate whether the record has been altered.

The following example contains provenance information about a metadata record:

```
<about>
    <provenance
    xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/provenance
    http://www.openarchives.org/OAI/2.0/provenance.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://www.openarchives.org/OAI/2.0/provenance">
        <originDescription altered="true" harvestDate="2005-1-15">
            <baseURL>http://oai.lib.uchicago.edu/</baseURL>
            <identifier>oai:lib.uchicago.edu:AEP-AKS25</identifier>
            <datestamp>2004-10-28T22:21:25Z</datestamp>
        <metadataNamespace>http://cicharvest.grainger.uiuc.edu/schemas/
        QDC/</metadataNamespace>
        </originDescription>
    </provenance>
</about>
```

3.12. BEST PRACTICES FOR OAI PMH STATIC REPOSITORIES

See Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: *Specification for an OAI Static Repository and an OAI Static Repository Gateway:* **http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm**.

The OAI Static Repository protocol provides a low-barrier solution for data providers to make their metadata collections available to the world. The data provider writes an XML file in a specific metadata format, which is made OAI PMH–harvestable through an OAI Static Repository Gateway, operated by a third party. The data provider contacts an OAI Static Repository Gateway owner and provides direction to the HTTP-accessible XML file.

[20] **http://www.openarchives.org/OAI/2.0/provenance.xsd**

Static repositories can support multiple metadata formats. As with all OAI repositories, unqualified Dublin Core (oai_dc) is required, but other formats are encouraged. Make the required additions to the `<ListMetadataFormats>` section of the static XML file, and add another `<ListRecords metadataPrefix="my_format">` section containing all of the records in the alternate metadata format.

In general, the size of a single static repository XML file should not exceed 2.5 megabytes. This is the same recommended maximum size for a single resumption token's worth of data when using the standard OAI PMH. This is the maximum size XML file that can be conveniently handled by some XML parsers. However, certain gateways may have different limits, either higher or lower, so check with the gateway to be certain of its limit.

OAI static repositories do not allow sets. Current best practice is to provide multiple static XML files, and therefore multiple static repositories, for distinct collections a data provider may have.

For more information about OAI PMH static repositories, see Hochstenbach, Jerez, and Van de Sompel (2003) and the DLF IMLS workshop material, "OAI Data Providers" (Habing 2006).

4. Best Practices for Shareable Metadata

4.1. INTRODUCTION

4.1.1. SCOPE

This chapter covers four main areas:

- The introduction, which provides useful background explaining the idea of "shareable" metadata
- Best practices for shareable metadata content, which include general recommendations on authoring shareable metadata, documenting decisions made about the metadata, and recommendations on authoring certain classes of metadata elements
- Best practices for technical aspects of metadata, which include recommendations for use of namespaces, XML schemas, and character encoding
- Final preparations, or how to check for shareability before implementing an OAI data provider

Throughout, the working group has tried to offer specific, concrete best practices and guidelines whenever possible. However, certain sections present either a range of choices or a spectrum of best practices from ideal to acceptable. These often represent an acknowledgment of the diversity and range of capabilities within the OAI PMH participants and, in some cases, areas where a best practice has not clearly evolved.

Readers will notice that the best practices do not focus on the use of unqualified Dublin Core within this document, but instead present best practices for a range of metadata formats. This is intentional, as the working group wishes to encourage data providers to offer records not only in unqualified Dublin Core (as required by the OAI PMH), but also in other metadata formats as appropriate.

The best practices assume a certain practical knowledge of metadata, although there is an attempt to provide basic definitions and examples. These best practices are also specifically focused on use of metadata expressed in XML within the OAI PMH context, although the principles and guidelines developed here should also be generally appropriate for other interoperability protocols and systems. The last caveat is that these best practices are not meant to guide the implementation of local-used metadata, only shared metadata.

For more general information about metadata as well as some specific case studies, please see the following publications:

- Caplan, Priscilla. 2003. *Metadata Fundamentals for All Librarians*. Chicago: ALA Editions.

  A good introduction to a range of metadata issues and formats. Aimed at librarians, but quite accessible to other communities as well.

- Gill, Tony, Anne J. Gilliland, and Mary S. Woodley. [n.d.] *Introduction to Metadata: Pathways to Digital Information*. Online edition, version 2.1, ed. Murtha Baca. Los Angeles, CA: Getty Research Institute. **http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html**.

  Another good basic introduction to metadata.

- Hillmann, Diane I., and Elaine Westbrooks, eds. 2004. *Metadata in Practice*. Chicago: ALA Editions.

  Presents a series of case studies from libraries, museums, and other communities working with metadata. Contains some specific case studies focused on OAI and metadata interoperability.

- National Science Digital Library. 2005. *Metadata Primer*. **http://metamanagement.comm.nsdlib.org/outline.html**.

  The NSDL primer for institutions interested in contributing metadata to the NSDL via the OAI PMH, but also more broadly applicable.

- RLG. 2005. *Descriptive Metadata Guidelines for RLG Cultural Materials*. **http://www.rlg.org/en/pdfs/RLG_desc_metadata.pdf**.

  Guidelines for institutions interested in contributing to RLG's Cultural Materials database, but very useful for any institution with cultural heritage materials.

4.1.2. WHY BEST PRACTICES FOR SHAREABLE METADATA ARE NECESSARY

Participants in the OAI PMH are many and diverse. Each data provider has its own needs and methods for describing its resources; therefore, metadata from one data provider may look very different from metadata from any other data provider, even when in the same metadata format. This diversity, however, makes it difficult for OAI PMH service providers to aggregate metadata from multiple data providers together in a meaningful way. However, the goal of these best practices is not to ask data providers to make all metadata more consistent to ease the burden for service providers, but rather, to offer guidance on how to author metadata that can be used successfully outside of its local environment. Often the shared metadata is not optimized for sharing; that is, it loses meaning and context when pulled out of its local environment. The more interoperable or shareable the metadata, the more robust and useful are the services that can be built on top of it.

The best practices included here represent the consensus of participants from a range of communities. As such, they are, for the most part, not specific to a particular metadata format or to a particular community, but instead offer general guidelines and best practices. The working group fully expects and encourages the further adaptation of these best practices for use by specific communities and domains.

4.1.3. QUALITY METADATA AND SHAREABLE METADATA

Thomas R. Bruce and Diane I. Hillmann (2004) discuss seven characteristics of quality metadata:

Completeness. Two aspects of this characteristic are described: choosing an element set allowing the resources in question to be described as completely as economically feasible, and applying that element set as completely as possible.

Accuracy. This characteristic is defined as the metadata being correct, factual, and conforming to syntax of the element set in use.

Provenance. Here provenance refers to providing information about the expertise of the person(s) creating the original metadata and its transformation history.

Conformance to expectations. Metadata elements, use of controlled vocabularies, and robustness should match the expectations of a particular community. This aspect of metadata quality is particularly problematic for OAI PMH data providers, as sharing metadata via OAI PMH allows it to be used by a wider variety of communities than previously targeted.

Logical consistency and coherence. This characteristic is defined as element usage matching standard definitions, and consistent application of these elements.

Timeliness. Two concepts make up this characteristic of metadata quality. *Currency* refers to metadata keeping up with changes to the resource it describes. *Lag* refers to a resource's availability preceding the availability of its metadata.

Accessibility. Proper association of metadata with the resource it describes and readability by target users contribute to this characteristic.

Quality metadata may or may not be shareable. That is, metadata may be of high quality within its local context, but for various reasons may be compromised when it is taken out of this context. Shareable metadata should, of course, have the above characteristics of quality metadata. However, there are some additional characteristics that make quality metadata more useful in a shared environment:

Proper context. In a shared environment, metadata records will become separated from any high-level context applying to all records in a group, and from other records presented together in a local environment. It is therefore essential that each record contain the context necessary for understanding the resource the record describes, without relying on outside information.

Content coherence. Metadata records for a shared environment need to contain enough information so the record makes sense standing on its own, yet exclude information that only makes sense in a local environment. This can be described as sharing a "view" of the native metadata (Lagoze 2001).

Use of standard vocabularies. The use of standard vocabularies enables better integration of metadata records from one source with records from other sources.

Consistency. Even high-quality metadata will vary somewhat among metadata creators. All decisions made about application of elements, syntax of metadata values, and usage of controlled vocabularies should be consistent within an identifiable set of metadata records so those using this metadata can apply any necessary transformation steps without having to process inconsistencies within such a set.

Technical conformance. Metadata should conform to the specified XML schemas and should be properly encoded.

4.1.4. BENEFITS OF CREATING SHAREABLE METADATA

Creating shareable metadata requires an investment of time. However, there are many benefits gained from making this investment.

The first and perhaps most significant benefit to creating shareable metadata is that it will be interoperable, or meaningful, when combined with metadata from other sources. By using metadata schemas and rules for creating metadata values similar to those used by others, your resources can meaningfully appear in search results alongside related resources from other metadata providers.

When creating truly shareable metadata, your resources are more likely to be found when pooled together with resources from other providers. Inconsistencies or gaps in descriptions of your metadata may mean that your resources will not be retrieved by searchers. Shareable resources will receive more exposure, and end-users will have the opportunity to make previously unseen connections between your resources and those from other metadata providers.

Finally, creating shareable metadata increases the number of access points for your resources available to end-users. Aspects of a resource not previously explicitly described are often added when metadata creators think in terms of shareable metadata.

4.1.5. STRIKING A BALANCE BETWEEN OAI PMH DATA PROVIDERS AND SERVICE PROVIDERS

Exposing metadata via the OAI PMH opens up content to uses data providers might not have imagined. Data providers cannot possibly tailor their metadata specifically for all the uses a service provider might have. Therefore, service providers will generally expect to do at least a minimal amount of processing of harvested metadata to achieve their goals. Service providers also generally have computing and processing power directed at massaging certain aspects of metadata that are relatively straightforward to normalize, such as dates. However, data providers most often know more about the resources being described than any given service provider. Data providers, therefore, can facilitate the service providers' work by following the guidelines for shareable metadata and by creating clear documentation describing the meaning and structure of the metadata they expose.

4.2. APPROPRIATE REPRESENTATION OF RESOURCES

- Think about uses of your metadata in an aggregated environment, and tailor your OAI PMH records for those purposes.
- Use metadata formats appropriate to your resources and your intended communities.
- Use metadata elements and construct values for those elements appropriate for a shared environment.
- Include the appropriate context for the resource in an OAI PMH metadata record.

It is important to conceive of shared metadata as a specific view of a resource. Good shareable metadata may not be appropriate for use in a local environment. Data providers may have local "master" metadata records from which they create versions of that metadata for specific purposes, including the OAI PMH. When creating records for use with the OAI PMH, institutions should envision the primary uses of this metadata and support these uses in their records. As of 2006, the vast majority of OAI service providers focused on resource discovery; therefore, data providers frequently tailor records meant for exposure via the OAI PMH for this purpose. Shared records must include the right amount of information. If too little information is included, records are not discoverable in a shared environment. If too much information is included, users of the shared metadata must wade through tangential information to target items that interest them. However, when in doubt, err on the side of including more information rather than less.

4.2.1. APPROPRIATE METADATA FORMATS

Metadata formats (sometimes called metadata schemas or elements sets) are groups of defined elements or fields that allow you to say something—descriptive, technical, or administrative—about a resource. Within the OAI PMH context, most exposed metadata are descriptive, that is, a set of elements that allows you to describe the resource for purposes of resource discovery or other services that rely on knowing what the resource is about.

Many communities have well-established metadata formats that are maintained as standards. It is a best practice to use supported, standard metadata formats. See Section 4.4, "Use of Multiple Metadata Formats" and Section 4.5, "Potential Metadata Formats for Use with OAI PMH" for more information.

Metadata formats should be appropriate to both the resources described and to the communities expected to use the resource. As an extreme example, the Metadata Object Description Schema (MODS) format is not a good format to describe the hierarchy of an archival finding aid; the Encoded Archival Description (EAD) is a much better format for this task. On the flip side, one would not use EAD to describe a single photograph; MODS might be a better choice for this type of resource.

4.2.2. APPROPRIATE METADATA CONTENT

The information you enter into the fields of a metadata format should describe the resource at an appropriate level, given the content described and the context in which it is likely to be used. You should consider element choices and controlled vocabularies that best represent the diversity and range of resources in your collection, not focus on one or two difficult-to-describe resources.

Many communities maintain controlled vocabularies (such as the Library of Congress Subject Headings) and encoding schemes (such as the W3C encoding scheme for date and time) as standards. It is a best practice to use appropriate standard controlled vocabularies and encoding schemes, like the use of standard metadata formats. It is a best practice to identify the controlled vocabulary in use if it is possible to do so. Note that in the required OAI PMH metadata format, oai_dc, it is not possible to do this. See Section 4.9, "Providing Supplemental Documentation to OAI PMH Service Providers" for more information.

Administrative metadata, which is primarily used internally to manage a digital collection and which often includes technical and preservation metadata, generally should not be exposed in OAI PMH records. Consider carefully what information is most useful to a service provider or end-user. An example of inappropriate exposure of administrative metadata might be multiple date fields describing updates to content or a reference to the type of scanner used to digitize a photograph. If administrative information is included in your metadata, you might consider exposing a "view" of your metadata (as discussed above) that excludes those fields.

4.2.3. APPROPRIATE CONTEXT

Problems with insufficient metadata frequently occur because metadata creators see a particular element or piece of information as unnecessary, given the larger context of the internal site or collection. However, when these individual metadata records are disassociated from their original context (as happens when the records are aggregated), the record becomes confusing or unusable. An example might be where all items in a collection are associated with a particular person, but the metadata records for the individual items do not include the person's name. This is what Robin Wendler (2004) has termed the "on a horse" problem.

For example, if a collection of digitized material is entirely about Mark Twain, the metadata describing this material may not include the subject heading "Mark Twain" or "Samuel Clemens," because it is obvious from the context of the collection that all resources are about Mark Twain. But when this metadata is made available via OAI PMH, it no longer has its original environment to provide that context. Wendler's descriptive term "on a horse" comes from a collection about Theodore Roosevelt. Photographs of Roosevelt on a horse simply had the singularly nonuseful (out of its local context, at least!) descriptive entry, "On a horse."

For example, determine what the following record describes:

```
<oai_dc:dc
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">

      <dc:title>Bowie County Texas (County Number 19, Supplementary Sheet D)
```

```
      </dc:title>
      <dc:creator>Texas Transportation Planning and Programming Division.
      </dc:creator>
      <dc:subject>Texarkana</dc:subject>
      <dc:subject>Kennedy Lake</dc:subject>
      <dc:subject>Coca Cola Lake</dc:subject>
      <dc:subject>Hobo Jungle Park</dc:subject>
      <dc:publisher> Libraries, University </dc:publisher>
      <dc:identifier>http://www.university.edu/raw/tcbowid1.html</dc:identifier>
</oai_dc:dc>
```
[21]

Most humans could interpret from the clues here ("Transportation Planning and Programming," "County Number 19, Supplementary Sheet D") that this is a map (and it is). But nowhere in the metadata record is this recorded. Any automatic processing that is done at the item level would not sort this record properly.

4.3 GRANULARITY OF DESCRIPTION

- Expose metadata records at the smallest level of granularity appropriate for the resources being described.
- Do not expose individual metadata records for digital objects that are only subordinate parts to a single item, unless they are unique objects unto themselves.

Data providers should expose metadata records at the smallest level of granularity appropriate for the resources being described. For many resources, this will be at the single-item level. Item-level description is most appropriate for resources whose individual characteristics are of primary importance to end-users, where the differences between a resource and other similar resources are significant. E-prints, published materials, letters, photographs, and paintings are some categories of resources for which item-level description may be useful.

For other types of resources, item-level description is not as useful. These resources include those in which a creator or manager has organized or grouped component parts into a coherent whole that serves as the resource of interest to end-users. Learning objects, Web sites, and some types of archival collections are categories of resources for which records describing groups of items might be most useful. For these resources, describe the characteristics of the group that would be of primary interest to the end-user.

Following is a portion of a record in oai_dc describing an archival collection (from AIM25: Archives in London and the M25 Area):[22]

```
<oai_dc:dc>
      <title>Christian Concern for Southern Africa</title>
      <creator>Christian Concern for Southern Africa</creator>
      <description>Records, 1966-1993, of Christian Concern for Southern Africa
      (CCSA), comprising papers on the constitution of the CCSA; its Executive
      Committee and Annual General Meeting papers; finance papers and examples
      of many of CCSA's publications and reports. Also included are files of
      correspondence between CCSA and churches and religious organizations,
      affiliated support groups and British companies in South Africa. Papers
      also include those of the Oil Working Group, which contain material on
      the Royal Dutch/Shell Group; the mass lobby of Parliament (17 June 1986)
      for 'Sanctions against Apartheid' organized by CCSA; and the Ethical
```

[21] We are not providing the link to this item in order not to single out the creator; however, this is a real record that has been slightly altered to maintain anonymity.

[22] **http://www.aim25.ac.uk/cgi-bin/oai/OAI2.0?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:aim25.ac.uk:64**

```
      Investment Research Service, founded as an independent offshoot of
      CCSA.</description>
      <publisher>School of Oriental and African Studies</publisher>
</oai_dc:dc>
```

It is rarely useful in an aggregated environment to expose records for multiple digital items that together make up a complete resource, for example, individual page images in a digitized book. It is a best practice that, unless each component part has significantly different descriptive information, metadata should instead be exposed for the intellectual resource (e.g., the book); access to the individual parts (e.g., the pages), if they have been digitized, should be supplied through navigation in the data provider's local environment. The same is true for collections of individual resources that vary only in a very small and, for the end-user, insignificant way—for example, a collection of coins that vary only in their accession number. These should be described as a collection, and access to the individual items should be made available within the data provider's local environment.

See also Section 3.10, "Best Practices for Sets" in this document for more information on creating logical groupings of records in an OAI repository.

4.4. USE OF MULTIPLE METADATA FORMATS

- Use of metadata formats in addition to unqualified Dublin Core (oai_dc) is both allowed and encouraged.
- Choose metadata formats to supplement unqualified Dublin Core that are expressed as an XML schema and are common in communities to which your resources are of interest.
- Metadata formats used must be listed in response to a `ListMetadataFormats` request.
- Indicate metadata formats used for records within a given set in the set description.

There is an unfortunate perception that the OAI PMH allows exposure of only unqualified Dublin Core records and that the OAI PMH can expose only a single metadata record for each item.

The reality is that the OAI PMH is designed to support records in multiple metadata formats for each item. An item can be exposed in as many metadata formats as desired as long as those metadata formats have an XML schema available for validation. Thus an item could be exposed as a MODS record, MARCXML record, qualified Dublin Core record, as well as the required unqualified Dublin Core record.

The OAI PMH does require an unqualified Dublin Core (i.e., the Dublin Core Metadata Element Set 1.1) record to be available for every item. For this purpose the Open Archives Initiative makes available an XML schema for unqualified Dublin Core and has reserved the metadata prefix oai_dc for this schema.[23] However, in addition to this required oai_dc record, records in other metadata formats can be provided for any or all of the items a repository includes.

It is a best practice that, in addition to unqualified Dublin Core, repositories expose the richest possible metadata formats available for all items in the repository. Why include additional metadata formats? Unqualified Dublin Core cannot express some of the complexities many OAI repositories wish to communicate about their resources (and service providers wish to know!). In addition, the oai_dc schema does not include a way to convey the controlled vocabularies and encoding schemes in use. Metadata formats that are more semantically complex can support a variety of uses, including ones not anticipated by the OAI repository. By supplying additional metadata formats that have the semantic richness to more clearly express meaning, data providers can help service providers make better use of their metadata.

The choice of additional metadata formats should be based on the robustness of description desired for the resources in question; the commonly used metadata schema in the community in which the resources will be primarily used; and, if applicable, the needs of a service provider by whom a repository specifically wishes to be harvested. Any number of additional metadata schemas may be used to reach desired audiences. However, metadata formats used with the OAI PMH must have an XML schema available for validation.[24] See Section 4.5, "Potential Metadata Formats for Use with OAI PMH" for a selected list of possible metadata formats.

All metadata formats available for harvest must be included in the response to a `ListMetadataFormats` request.

[23] **http://www.openarchives.org/OAI/2.0/oai_dc.xsd**

[24] See **http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces** for more information.

If multiple metadata formats are available and sets are implemented, it is a best practice to include in the set description(s) the metadata formats that are available for the items in a particular set. This is because the protocol does not require that all items be available in all metadata formats (besides unqualified Dublin Core), and there is no way in the protocol to request the `metadataPrefixes` in use for a specific set. For example, one set may include items available in both oai_dc and MODS, while a second set may include items available only in oai-dc. If some, but not all, items in a set are available in an additional metadata format(s), it is recommended that additional sets, corresponding to the additional format(s), be built. See Section 3.10, "Best Practices for Sets" in this document for further information.

As discussed in Section 4.6, "Crosswalking Logic," repositories with metadata formats other than unqualified Dublin Core may benefit from first attempting to crosswalk their native metadata to qualified Dublin Core, then "dumbing down" to unqualified Dublin Core. This strategy allows the repository to make iterative small changes rather than one significant change, and may ensure the best result with the least loss of data or specificity.

See Section 4.11.1, "XML Schemas and Namespaces" for discussion of some technical issues.

4.5. POTENTIAL METADATA FORMATS FOR USE WITH THE OAI PMH

The use of multiple metadata formats (at least one in addition to the required unqualified Dublin Core) is strongly encouraged. The following list of potential metadata formats for use with OAI PMH is necessarily incomplete. To see the range of metadata formats currently in use by OAI data providers, see the *Distinct Metadata Schemas* report from the OAI Registry at the University of Illinois at Urbana-Champaign.[25]

Categories for the Description of Works of Art—Lite (CDWA Lite)

CDWA Lite may be a good option for data providers that

- Wish to describe works of art and material culture, and
- Intend their metadata to be used by specialist audiences in the art domain.

The CDWA Lite XML Schema version 1.1 is available at **http://www.getty.edu/CDWA/CDWALite/CDWALite-xsd-public-v1-1.xsd**.

Encoded Archival Description (EAD)

EAD may be a good option for data providers that

- Wish to share records about archival collections where expressing the relationships between items and the context of the collection as a whole is important, and
- Have as a primary audience for resources described via OAI records a community well versed in archival descriptive practices.

The EAD 2002 schema is available at: **http://www.loc.gov/ead/ead.xsd**.

Electronic Theses and Dissertations Metadata Standard (ETD-MS)

ETD-MS may be a good option for data providers that

- Primarily expose metadata about electronic theses and dissertations, and
- Wish to contribute to aggregations of electronic theses and dissertations such as the Networked Digital Library of Theses and Dissertations.[26]

The ETD-MS XML schema is available at **http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd**.

MARCXML

MARCXML may be a good option for data providers that

- Locally describe resources in MARC according to AACR2, and
- Have as a primary audience for resources described via OAI PMH records the core library community.

The MARCXML XML schema is available at **http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd**.

[25] **http://www.gita.grainger.uiuc.edu/registry/ListSchemas.asp**

[26] **http://www.ndltd.org**

Metadata Object Description Standard (MODS)

MODS may be a good option for data providers that

- Locally engage in descriptive practices heavily influenced by resource description standards in libraries, and
- Have as a primary audience for resources described via OAI records a community well versed in library descriptive practices, yet want robust records in a format accessible to service providers outside the core library community.

The MODS v.3.2 XML schema is available at **http://www.loc.gov/standards/mods/v3/mods-3-2.xsd**. See also the *DLF Aquifer Implementation Guidelines for Shareable MODS Records*, which gives specific guidance on construction of MODS records for exposure via the OAI PMH and was based heavily on these best practices.

Qualified Dublin Core

Qualified Dublin Core may be a good option for data providers that:

- Have a need for more granularity of description than is available in unqualified Dublin Core, but not a fundamentally different approach to resource description;
- Use controlled vocabularies that they wish to specify within their metadata records; and
- Have resources of interest to many different knowledge communities with disparate descriptive metadata practices.

There is no single canonical qualified Dublin Core XML schema. However, an XML schema for qualified Dublin Core can be created through the importation of the necessary namespaces and schemas. See the information available at

http://dublincore.org/schemas/xmls/qdc/dc.xsd

http://dublincore.org/schemas/xmls/qdc/dcterms.xsd

http://dublincore.org/schemas/xmls/qdc/dcmitype.xsd

VRA Core

VRA Core may be a good option for data providers that

- Wish to share works of visual culture and the images that document them, and
- Wish to group sets of images related to these works together.

The VRA Core has two XML schemas:

- The unrestricted version at **http://www.vraweb.org/projects/vracore4/vra-4.0.xsd** specifies the basic structure of the schema.
- The restricted version at **http://www.vraweb.org/projects/vracore4/vra-4.0-restricted.xsd** extends the unrestricted schema by adding controlled type lists and date formats.

4.6. CROSSWALKING LOGIC

- Map metadata from more robust formats to simpler ones.
- Plan for both mapping values between fields and for transforming data values themselves to meet the expectations of the target metadata format.
- Repeat elements when your target metadata format allows it.
- Include titles and appropriate context in your mapped metadata.
- Exclude indications of unknown or inapplicable data, and artifacts of descriptive practices not applicable to the target metadata format.
- Stepped crosswalking may be beneficial.

Ideally, crosswalking from one metadata format to another should be done in a manner that limits loss of data or specificity. Crosswalking may be done to convert records from a local format to a community standard or to convert records from a richer to a simpler metadata format. Most data providers that are not using unqualified Dublin Core as their primary format will need to do some crosswalking to provide the required unqualified Dublin Core (oai_dc) records for the OAI PMH. Always crosswalk from a richer metadata standard to a simpler one; mapping from a simple schema to a richer one will not usually yield any extra usable data. After determining the metadata formats to expose via the OAI PMH, decide whether to use standard crosswalks that are often available or to develop one.

In some cases, a direct crosswalk from a very rich format to a simple one involves loss of information and context. One way to prevent unnecessary loss and improve the accuracy of mapping is to crosswalk the richer format into an intervening format as part of the process. For example, qualified Dublin Core can serve as an intervening format between richer metadata and unqualified Dublin Core (oai_dc). This stepped approach can be used simply as a conceptual tool for improving the quality of mapping, or as a stage in actual data transformation. By making transformations in a series of small steps rather than one large step, repositories may find it easier to match the semantics of elements in the source format to those of Dublin Core.

Performing data transformation in stages has the added benefit of creating records in the intervening metadata format (in our example, qualified Dublin Core) that can then be exposed via the OAI PMH, in addition to the required unqualified Dublin Core and any other metadata formats. As mentioned in Section 4.4, "Use of Multiple Metadata Formats" in this document, data providers are encouraged to provide metadata in more than one format.

To create a custom crosswalk, the primary task is to develop the logical rules for transforming the existing metadata records into those to be made available via the OAI PMH. This involves defining the steps, as specifically as possible, necessary to change the metadata into the version to be exposed. These logical rules may operate at several levels and require distinct decisions about data mapping, data value transformations, data to include, and data to exclude.

4.6.1. DATA MAPPING ISSUES

Map the complete contents of one element to another, as is. This case occurs when a metadata element in a local implementation matches exactly the semantics of a target element, and the existing value(s) in this field is formatted exactly in the same way the target element value should be. The transformation rule in this case simply involves copying the value in the local field to the field in the target schema.

Splitting data in one element into two or more elements. The target metadata format may have separate elements for data that the local format stores in a single element; e.g., publisher name and place, or first and last names. Developing mapping logic in this case would require identifying the rules for deciding what part of an existing element goes to each new element in the target schema.

Splitting multiple values in a single local element into multiple iterations of a single element in the target format. This case occurs when the local implementation allows for multiple values of the same type to be placed together in a single element, and your target metadata format recommends repeating elements rather than "packing" multiple values of the same type into a single element. The transformation rule in this case requires specifying the characters within the element in the local implementation that indicate the end of one value and the beginning of the next (for example, ";" or ",").

4.6.2. DATA VALUE MAPPING ISSUES

Translating anomalous local practices into a more generally useful value. For example, because of limitations of local software, many data providers store date ranges as a comma-delimited set of individual years; e.g., the date range 1890–1895 would be expressed as 1890, 1891, 1892, 1893, 1894, 1895. Such software-specific work-arounds should be translated back to the original date range value for shared records.

Transforming data values to expected syntax and practice. The syntax of data values in local elements may not match the expected syntax of a given metadata format or the descriptive practices of the community a data provider might be trying to reach with shared records. For example, if local practice dictates that supplied titles be indicated with [brackets] and the target aggregator and user community does not understand the meaning of brackets, the use of brackets in shared records can be problematic. The elements affected must be identified and the appropriate mappings made when possible.

4.6.3. DATA TO INCLUDE

Titles. While Dublin Core does not require any element, OAI PMH service providers commonly use the Dublin Core title element as the core of a brief results display. Think carefully before exposing Dublin Core records without titles. If it is determined that a title is not an appropriate element for the described resources, document this decision, along with information about the fields in your records most useful for a brief results display, either in a set description or in other appropriate documentation. See Section 4.9, "Providing Supplemental Documentation to OAI PMH Service Providers" and Section 4.10.1, "Titles in Shared Records" for further discussion.

Repeating elements. Whenever using a metadata format that allows for repeating elements, always repeat the element for each value rather than "packing" multiple values into a single element. It is easier for the service provider to merge them for display than divide them to process the various values within an element.

Preserving context. Data providers should ensure that their record, when standing alone, makes sense outside its local context. For example, in a collection of Russian images, each record should contain reference to Russia. Any context useful for information discovery or for information display that is not included in the local record should be automatically added to each individual record during mapping.

4.6.4. DATA TO EXCLUDE

Values indicating that information is unknown or the element is not applicable. It is possible that a record will have elements with values that are essentially empty or values that indicate that the relevant information is unknown or not applicable such as `<dc:date>--</dc:date>` or `<dc:subject>XXX</dc:subject>`. In several cases, the value, or lack thereof, might be considered as information (the fact that the object has no date), but generally this is of little interest for the purpose of information retrieval. Take for example, `<dc:date>ND</dc:date>` or `<dc:date>undated</dc:date>` or `<dc:creator>unknown</dc:creator>`: this information may be useful in a local environment, but it creates problems in a shared environment. Shared records should not include these values. Similarly, if there are no data for an element in the schema being exposed, the OAI record should refrain from using that element, rather than presenting it as empty, for example, `<dc:date></dc:date>` or `<dc:date/>`.

Junk values. When mapping values from one metadata format to another, do not include values representing artifacts of description from the original records, such as `<dc:creator>et al.</dc:creator>` or `<dc:creator>and</dc:creator>`.

4.7. DESCRIBING VERSIONS AND REPRODUCTIONS

- Adhere to the one-to-one principle when practical.
- When it is necessary to provide access to multiple versions of a resource, carefully select a strategy from options used by other data providers in the OAI PMH community.
- Do not provide unnecessary information that does not serve the primary purpose of the shared record.

Many resources described by metadata records shared via the OAI PMH exist in multiple versions. An image may exist as a film negative, a digitized TIFF master image, and three sizes of JPEG derivative images. A text may exist as a TEI file, an unedited ASCII text file, and a PDF document of digitized page images. Metadata is also used (particularly by museums) to describe physical objects, as well as reproductions of those objects in various formats.

Many metadata standards emphasize the description of only one manifestation of an object in a metadata record. Hillmann (2005) defines this concept as the "one-to-one principle":

> **The One-to-One Principle.** In general Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand in for one another. For instance, a jpeg image of the Mona Lisa has much in common with the original painting, but it is not the same as the painting. As such the digital image should be described as itself, most likely with the creator of the digital image as Creator or Contributor, rather than the painter of the original Mona Lisa. The relationship between the metadata for the original and the reproduction is part of the metadata description, and assists the user in determining whether he or she needs to go to the Louvre for the original, or whether his/her need can be met by a reproduction.

The one-to-one principle was designed to forestall the common problem of metadata that described more than one manifestation of a resource. Because Dublin Core is dependent on syntax for structure, and many (if not most) of the syntaxes used for Dublin Core have no mechanism for relating statements when more than one description is embedded in a single record, interoperability is severely compromised. This is a particular problem when metadata from many sources, with different practices regarding versions, are aggregated to support discovery over a broad range of materials. This can be compounded because of the requirement of oai_dc in the OAI PMH, particularly if oai_dc is the only format provided.

It is important to note that metadata formats other than Dublin Core have similar difficulties maintaining connections between descriptive elements within records describing multiple versions of resources. However, some metadata formats—particularly those coming from the visual resources and museum communities—have made a concerted effort to separate description of the resource (or "work," although this is different from a FRBR work!) and what are considered visual surrogates (or versions). CDWA Lite and VRA Core are two metadata formats that separate these two concepts. Other metadata formats also have a means internal to the format itself to describe an object and a version of that object. In MODS, for example, the top-level record could be used to describe a physical book, but the `<relatedItem>` element could conceivably be used to describe a digitized version of that book.

However, complete adherence to the one-to-one principle, particularly for shared records, is often impractical. The difficulties of recombining versions for user displays—and the current primitive state of linking mechanisms between related records—should be considered when planning for the creation of shared metadata records. The primary consideration in determining what to share about a resource or set of connected resources should be how useful that record will be in an aggregation.

Below are described some common compromises—listed in no particular order—between practice and the one-to-one principle, with some descriptions of advantages and disadvantages of each. Note that these are for shared records, particularly the required oai_dc records. Much more satisfactory approaches may be possible in the local environment, but these tend to be flattened out in the shared environment.

A. THE INTELLECTUAL CONTENT AND ENTRY PAGE APPROACH

In this approach, the metadata description is relatively generalized and focused most on the content of the resource rather than the carrier. There is little technical detail, except, perhaps, a list of digital formats in which the item is available, in repeating elements. The identifier for the record leads the user to an HTML page with links to the versions available, including instructions, caveats, etc.

Advantages:

Fairly clean, easily updated without necessarily needing to update metadata frequently.

Metadata record creation can emphasize provision of information on topic, contributors, etc., without worry about reconciling versions of these records.

Disadvantages:

Aggregators cannot easily get to content through entry pages when content is used to support additional indexing.

When format information is present only on entry pages, downstream services cannot easily support user filtering by format.

Examples:

arXiv.org - **http://arXiv.org/oai2?verb=ListRecords&metadataPrefix=oai_dc**

Most DSpace OAI data providers use this approach.

B. THE "LINKING" APPROACH

This approach uses the `<dc:relation>` element or other linking strategies to connect related records for versions and reproductions. The links may be URIs or other identifying strings, or they may be citations. Linking approaches may vary; some may require reciprocal links, and others may be one-way only.

Advantages:

Unambiguous when using URIs or standard numbers.

Generally does not require explanation to be interpreted by others.

Disadvantages:

Expensive and sometimes difficult to maintain.

Does not scale well to complex relationships.

Example:

The K-MODDL project: **http://kmoddl.library.cornell.edu/oai/oai2.php?verb=ListRecords&metadataPrefix=oai_dc**

Note that in its native interface, the K-MODDL project uses several vocabularies to manage relationships between a wide variety of media and versions all relating to a collection of physical models demonstrating mechanical principles.

C. CONTENT OF THE PHYSICAL ITEM AND CARRIER OF THE DIGITAL ITEM APPROACH

This approach is similar to compromise A in that the intellectual content of the physical item is of primary importance in the metadata record, but the characteristics of the digital carrier(s) are also described.

Advantages:

Fairly clean, easily updated without necessarily needing to update metadata frequently.

Metadata record creation can emphasize provision of information on the intellectual content.

Disadvantages:

Users may want to know more about the physical carrier.

If multiple carriers are described, must distinguish between them.

Example:

USC Digital Archive. Columbia, South Carolina Aerial Photos Pilot Project:

http://digital.tcl.sc.edu/cgi-bin/oai.exe?verb=ListRecords&metadataPrefix=oai_dc&set=AP

4.8. LINKING FROM A RECORD TO A RESOURCE AND OTHER LINKING ISSUES

- URIs in shared metadata records should point to the described resource.
- URIs in shared metadata records should be permanent and persistent.
- URIs in shared metadata records should include indication of the protocol used.
- Provide a single URL or an indication of the key URL pointing to a resource in appropriate context.

Also see Section 4.10.7, “Identifiers in Shared Records.”

For the most part, the metadata exposed via the OAI PMH describe Internet-accessible resources, i.e., the metadata contain URLs that lead a user to the digital resource. This is not always the case: some OAI repositories contain metadata that describe analog resources that are not available digitally. Either way, appropriate links are an important piece of the ultimate shareability of the record. In the case of metadata describing digital resources, the location to which users are sent after clicking on a URL is of critical importance for the end-users, but the URL link also affects the credibility of both the data and service providers. It is the URL that most often links the metadata record in the service provider’s aggregation with the actual resource in the data provider’s repository. Send users to the wrong place—or to an empty page—and they may leave both the data and the service providers’ sites altogether. End-users may also become frustrated if they must wade through many layers of indirection (i.e., numerous mouse clicks) before viewing the resource itself. For those sites using the OAI PMH as a Google sitemap, the URLs are perhaps of even greater importance, because Google relies on the URLs, not the descriptive metadata, to spider the site.

A metadata record may describe a single resource available as a single digital file. As discussed in the previous section, a single intellectual resource can have multiple representations or versions. Resources may also have multiple parts, such as multipage texts (as discussed in Section 4.3, “Granularity of Description”). A metadata record describing an analog resource may have a URL pointing to the institution’s homepage or a page describing the collection. Metadata records may also include URLs to pages describing conditions of use, copyright information for the resource, related resources, etc. Thus, numerous links may be included in a single metadata record.

As a result of this variability, the specific level of representation a URL should point to cannot easily be prescribed, but the following best practices should be followed whenever possible.

In all cases, it is a best practice to use some form of standard, persistent URL (such as a PURL, Handle, or a resolvable DOI) if possible. If this is not possible, it is a best practice to keep URLs up to date within the metadata available via an OAI data provider. Link rot is a significant problem for service providers, and one that is entirely out of their control except through reharvesting updated metadata records from the data provider.

Strictly speaking, a URL[27] must be a valid URI[28]. Implementers should refer to the most current documentation to determine which characters are reserved and unreserved and which may need to be escaped. URIs require that non-ASCII characters and some ASCII characters be escaped. In practice, this may require special treatment of certain characters, such as spaces, in URLs. See also Section 4.11.2, "Character Encoding Issues."

It is a best practice to include the appropriate scheme prefix (i.e., http://, ftp://, etc.).

4.8.1. URLS THAT POINT TO DIGITAL RESOURCE(S)

It is a best practice to provide one, primary URL or an indication of a single primary URL that is a link to the resource with its contextual material (e.g., metadata, navigation to the collection homepage). For example, in the case of a Dublin Core record, a single `<dc:identifier>` element should contain the URL that points to the resource. This element should not be repeated unless with information that is not a URL (for example, an ISBN). In the case of a MODS record, `<location><url>` allows a usage attribute set to "`primary display`" to indicate the primary URL. Thus the `<location><url>` element could be repeated, albeit not with `usage="primary display"`.

The use of one primary URL or the indication of such allows service providers to know what URL to provide as the link to the resource. In cases of multipart resources or resources with multiple manifestations, this best practice means that data providers must make a decision about what is the most relevant or appropriate place for a user to get access to the resource. Secondary URLs that take users to other versions or manifestations of the resource should be included in other elements (such as `<dc:relation>`).

An addendum to this best practice: if it is possible to indicate the URL that points to the object in context, then it is useful to provide direct links to the digital objects described to allow machine processing or spidering of those objects. For example, some service providers use links to images to make thumbnails to include in the search results of an aggregation.

The best practice of the URL pointing to the resource in context is important because many service providers do not display the full metadata record harvested from a data provider. Thus, if the primary link to a resource is to a stand-alone version of the resource (such as a JPG image only), an end-user will have no context except for the metadata on the service provider's site. This does not serve the end-user well, nor does it serve the data provider well as the end-user cannot easily navigate to other parts of the data provider's collection. At a minimum, the URL should point to a page that contains the resource and a navigation bar that allows users to reach the collection homepage. It is highly desirable that this page also include the descriptive metadata for the resource. For those sites using the OAI PMH as a Google sitemap, the context is perhaps of even greater importance, because Google relies on the URLs, not the descriptive metadata, to spider the site. If a URL goes to a JPG image only, that may be problematic.

If it is not possible to provide a URL that links to the resource, the end-user should be able to access the resource with a minimum number of clicks (at most two) with a minimum amount of effort. For example, in the case of the arXiv.org data repository, the URL provided takes the user to a page with basic metadata and a selection of formats to choose from (see **http://arxiv.org/abs/chem-ph/9403001** as an example). The resource is only one click away. The same is true of DSpace repositories.

It is particularly bad practice to include as the primary URL for a resource the collection homepage with the expectation that the end-user will conduct what is probably an additional search to find the relevant resource. End-users' frustration with this particular practice is described by Shreeves and Kirkham (2004):

> [The subjects] reported a significant slowing of their efforts when a pointer, or active link, within a record led them to another institution's Web site in which they had to execute an additional search. The subjects clearly believed that a live URL in a search result should immediately display the digital object of interest. One student described the interaction as being comparable to going to McDonald's "and upon walking up to the counter, the employee hands across directions to Burger King across town."

[27] **http://www.w3.org/Addressing/URL/Overview.html**

[28] See **http://www.ietf.org/rfc/rfc2396.txt**, **http://www.ietf.org/rfc/rfc2732.txt**, and **http://www.ietf.org/rfc/rfc3986.txt**.

4.8.2. METADATA RECORDS DESCRIBING ANALOG RESOURCES

If there is no digital object available, this should be indicated in the metadata, and either contact information for the institution or collection or a URL to a page with this contact information should be provided. However, it is a best practice to clearly distinguish this URL from those that will lead a user to digital content. For example, in a Dublin Core record, a URL to an institution or collection homepage should not be included in the `<dc:identifier>` element but in the `<dc:relation>` element. In a MODS record, the `<location>` element and `<physicalLocation>` subelement should be used, not the `<url>` subelement. This helps to mitigate confusion for users and service providers.

4.8.3. OTHER LINKS IN METADATA RECORDS

Among the other links that may be included in metadata records are

- Conditions of use
- Copyright information
- Access restrictions
- Collection or institution homepage
- E-mail addresses for contacts
- Related resources or collections

In all cases, the links should be current and clearly distinguished from the primary URL for the described resource by including these in elements other than that used by the primary URL.

4.9 PROVIDING SUPPLEMENTAL DOCUMENTATION TO OAI PMH SERVICE PROVIDERS

- Provide documentation on choices made when providing metadata for exposure via OAI PMH.

There are two parties to an OAI PMH mediated transaction: the data provider and the service provider. They exchange metadata in the context of a protocol that allows a good bit of information about the metadata to be communicated within the protocol itself. However, as the OAI PMH world becomes increasingly diverse, the service provider may need to know more than what the protocol requires or encourages the data provider to expose in order to make the best use of harvested metadata. Consequently, the data provider should:

- Make careful use of the opportunities within the OAI PMH (such as `<about>` containers) for documenting practices.
- Make additional information available for any service providers that wish to know more about the metadata and its origins, changes, and capabilities.

Regardless of the metadata format(s) data providers expose via the OAI PMH, it is always a good idea to provide documentation for the decisions and standards in use for the exposed metadata. Such documentation can help service providers better aggregate your metadata with others because the service provider will have a better understanding of how to interpret and normalize the metadata. This documentation is especially important when a data provider is exposing only a low-granularity metadata format like oai_dc, as it can help service providers to make sense of the choices the data provider has made. It can also be important for the data provider itself to keep track of the decisions about shared records that it has made over time!

Bruce and Hillmann (2004) assert that optimal metadata provision should include the following:

- An expression of metadata intentions based on an explicit, documented application profile, endorsed by a specialized community, and registered in conformance with a general metadata standard. They add that an XML schema does not express intention.
- A source of trusted data with a known history of regularly updating metadata, including controlled vocabularies. This includes explicit conformance with current standards and schemas.
- Full provenance information, including nested information, as original metadata is harvested, augmented, and reexposed. This may not record changes at the element level, but it should reference practice documentation that describes augmentation and upgrade routines of particular aggregators.

Note that this last point is particularly relevant to aggregators that are reexposing metadata harvested from elsewhere.

4.9.1. WHAT TO DOCUMENT

Data providers should consider documenting the following:

- Data source and creation decisions and history. For example,

  Is the data crosswalked from another data source? Is that data source available in its native form? If so, where?

  Example documentation: A table that illustrates the mapping of the local metadata format to the metadata format that is exposed for harvest.

  Is the data created by humans or machines?

  Example documentation: Describe the creation methodology or refer to a description.

- Use of controlled vocabularies and content standards. For example,

  What vocabularies are used, for what elements, and under what circumstances? (This is especially important if only oai_dc is used.)

  Example documentation: Indicate that the subjects are assigned using Library of Congress Subject Headings (LCSH), and your genre and form terms are assigned using Getty Art & Architecture Thesaurus (AAT), either within the metadata format itself, if possible, or in outside documentation when not possible (as in the case of oai_dc).

  Is the whole vocabulary available to be used with this data, or is only a subset approved for use?

  Are some terms used from local lists and not specified in the namespaces (possibly not formally documented)? Is any documentation available on local vocabularies, and if so, where?

  Information about the descriptive content standards that are used locally (such as the Anglo-American Cataloging Rules [AACR2], Describing Archives: A Content Standard [DACS], or Cataloging Cultural Objects [CCO]) will help service providers harvesting your metadata to understand the context for your metadata. It may be useful to them in normalizing and transforming your metadata content so that it works most efficiently for end-users of their Web-based portal or service.

- Names practice. For example,

  Order of names (direct order or surname, forename)

  Completeness of names (initials or full names)

  Any additions to names (courtesy or academic titles, affiliations)

  Authority source for names (availability of name variants)

- Dates practice. For example,

  Parsing rules if not encoded

- Identifiers. For example,

  For unencoded or local identifiers, describe the source of the identifier and any parsing or validation rules.

- Quality control measures. For example,

  Are controlled vocabulary values updated when the source vocabulary changes?

  Are validation routines run regularly on encoded data?

- Updating practices and schedules. For example,

  How often are new or changed records added to the database?

  How many new or updated records are added per week or month?

- Set decisions and specifications (see also Section 3.10.3). For example,

  Are sets added or removed on a regular basis?

  Are all records included in sets?

  Is there overlap between sets?

In addition, supplemental documentation is a good way to provide special information and eliminate guesswork for service providers harvesting your metadata via the OAI PMH. As one example, many types of resources for which metadata records are exposed may not have a meaningful formal title, for instance, a set of satellite images or a set of images of fossils. In both of these cases, end-users might find a subject, geographic location, or type of resource more meaningful than a title supplied by a cataloger or the service provider in the absence of a formal title. Nonetheless, many service providers will need to identify a title-equivalent field to be used in automated citation generation or in search results lists. It is very helpful for data providers, who are usually the content experts, to supply information regarding a preferred title-equivalent metadata field for sets in which records do not contain titles.

4.9.2. HOW AND WHERE TO DOCUMENT

Data providers should consider documenting the practices above in the following locations:

- The metadata itself, wherever possible

  Many metadata formats allow the documentation of controlled vocabularies and encoding schemes in use. Data providers should take advantage of this as much as possible.

- The OAI PMH responses

  Repository description contained in the Identify response (see Section 3.3).

  Set specifications documented in set descriptions (see Section 3.10.3).

  Individual record descriptions in `<about>` containers (see Section 3.11).

- External documentation

  Web pages. References to documentation Web pages from within OAI Identify responses or `<about>` containers can lead service providers to additional documentation.

  Application profiles can document individual provider or community decision making in human and machine readable versions. See Heery and Patel (2000) for a description of an application profile. See also the *Dublin Core Application Profile Guidelines* (CEN 2003).

4.10. RECOMMENDATIONS FOR CLASSES OF DATA ELEMENTS

These recommendations for classes of data elements are not specific to any one metadata format, but should be applicable across a range of formats and content standards. These recommendations are also focused specifically on shared records, not necessarily on records used within local systems.

4.10.1. TITLES IN SHARED RECORDS

- Provide a title in every shared record whenever possible. Supply one if necessary, according to established standards.
- Express multiple titles in repeated fields.
- Make the distinction between title and subtitle clear through the metadata format used or through standard punctuation.
- If there is no title, consider carefully the ramifications of not supplying one.
- Format titles consistently within an OAI PMH set or repository.

4.10.1.1. TYPES OF TITLES

Titles are an extremely important access point for resources and are frequently used in brief record displays to assist end-users in deciding whether to investigate a resource further. Typically, aggregators will build indexes based on the titles in metadata records and will almost always display titles in the results of a search query. It is a best practice to provide a title in the shared record if at all possible.

Titles can originate from two main sources: the creator of the resource or the creator of the metadata describing the resource. When available, titles assigned by the creator of the resource are preferred and should be transcribed as accurately as possible.

In cases where there is no original title for a work, a descriptive title created by the metadata author may be supplied. This should be created in accordance with an established standard, and the title should describe, as concisely as possible, the content of the work. Consider carefully, however, whether to use the notation indicated by a content standard for a supplied title. For example, the AACR2 standard uses brackets to indicate a supplied title. However, if records are to be shared outside of the library community, the brackets may cause problems in indexing for the aggregator and problems in interpretation for the user community. The same is true for conventions such as adding "[electronic resource]" at the end of titles that are electronic resources (Tennant n.d.).

If the nature of the items is such that a title is neither provided nor useful (satellite images, for example), the provider should not use default text in the title field indicating that there is no title, as this strategy often backfires in an aggregated environment. If possible, other descriptive information about the resource should be provided in a field other than the title field to allow adequate indexing of the metadata record as well as to support end-user decision making. In the example below, the photographer's name and related keywords provide an adequate substitute for the lack of a title.

4.10.1.2. FORMATS OF TITLES

If using a metadata format that does not allow for specification of subtitle, title and subtitle may be collapsed into a single field, preferably separated by a colon or other punctuation:

```
<dc:title>The story of my life; or, The sunshine and shadow of seventy years/
by Mary A. Livermore... with hitherto unrecorded incidents and recollections
of three years' experience as an army nurse in the great Civil War, and
reminiscences of twenty-five years' experiences on the lecture platform... to
which is added six of her most popular lectures... with portraits and one
hundred and twenty engravings from designs by eminent artists... </dc:title>
```
[29]

Multiple titles should be coded in separate fields. If one title is intended to be the primary one, and no distinction is available in the element name, the most important title should appear first. For example,

```
<title>Social Science Research Building</title>

<title>Series II: Buildings and Grounds</title>

<title>10th Anniversary Conference 1</title>
```
[30]

4.10.2. NAMES IN SHARED RECORDS

- Include all known names expected by your community of practice in shared records.
- Format names consistently within an OAI PMH set or repository.
- Provide as granular an encoding of a name as possible in the metadata schema being used.
- Express multiple names in repeated fields.

4.10.2.1. USE OF NAMES

Names are a critically important component in shared metadata. Names are most likely to appear as creators of and/or contributors to the intellectual content of the resource, but may appear also in subjects, as rights holders, as publishers, or in other contexts. Names in metadata may be used to:

- Select the appropriate resource;
- Provide proper citations for the resource;
- Determine the intellectual property issues that may accompany uses of the resource;
- Provide useful sorting of resources; and
- Collocate resources by or about the same entity.

[29] **http://quod.lib.umich.edu/cgi/b/broker20/broker20/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:lib.umich.edu:4728109.0001.001**

[30] **http://oai.lib.uchicago.edu:8180/?verb=GetRecord&metadataPrefix=oai_dc&identifier=lib.uchicago.edu:apf2-07489**

The name(s) of the creator(s) of the resource should be included in the shared metadata record whenever known. The creator(s) name(s) should be clearly distinguished from names of persons or corporate bodies that are subjects of the resource being described or other named entities in the metadata record.

The guidelines below generally apply to all uses of personal and corporate names in a metadata record.

4.10.2.2. CHOICE OF NAMES

Determination of the importance of a name and how many should be included in a metadata record about a particular resource should be based on the practice or expectation of the domain or community providing the data. For example, if the community expects all authors to be listed as creators, this should determine practice. If a content standard like the AACR2, DACS, or CCO is used to determine practice in this area, it should be documented by the provider. See Section 4.9, "Providing Supplemental Documentation to OAI PMH Service Providers."

4.10.2.3. FORMAT OF NAMES

Names should be entered consistently within an OAI PMH set or repository, based on the content practices of the domain or community providing the information. Consistency is important because it allows the OAI service provider to more easily process the metadata.

In traditional library practice, personal names are rendered "surname, forename" and literary warrant is used to determine appropriate completeness of the name. In that community, external authority files (such as the Library of Congress Name Authority Files) control the format of names, determined according to standard content rules (generally AACR2). In a context where names are text values, "surname, forename" practices allow more appropriate sorting.

In other communities with different content standards (or lacking content standards), direct order names are acceptable but should be applied consistently. Some communities traditionally use only forename initials rather than full forenames. Although this practice complicates the determination of whether "Smith, A." who publishes about chemistry is the same as "Smith, A." whose topic is golf, this practice is acceptable for records created by and intended to be used in these communities.

Corporate names are complicated by issues of hierarchical ordering and frequent changes. Traditional libraries follow fairly complex rules about order of hierarchy and whether all levels must be expressed, but in other communities such considerations are not relevant. In general, best practice mandates only that acronyms for organizations be consistently spelled out, so that communities not familiar with the acronyms can make use of the data.

4.10.2.4. ADDITIONS TO NAMES

In some communities, titles, dates, or academic affiliations of the person names are included in the name element to reduce ambiguity. Before including extra information within a name element, data providers should ask whether the addition to the name can reasonably be considered an extension of the name itself. Using these criteria, a date or title is normally acceptable, but an affiliation is not, since the affiliation is actually the name of a separate entity.

It is important, when considering what information to add to a name in the context of a resource description, to limit the addition of information that serves to describe the person, rather than the resource. In an ideal system, the person or corporate body would be described unambiguously somewhere else, and the relationship between them expressed in the metadata description for the resource, but few current systems allow this. See the *DCMI Abstract Model* for an example of a model that allows the expression of related metadata records (Powell et al. 2005).

4.10.2.5. ENCODINGS OF NAMES

The metadata format, content standard, and authority file in use will often prescribe the granularity of encoding of a name.

In MODS, for example, a name heading for a creator can be broken into several parts:

```
<mods:name type="personal">
      <mods:namePart type="family">Cushman</mods:namePart>
      <mods:namePart type="given">Charles Weever</mods:namePart>
      <mods:namePart type="date">1896-1972</mods:namePart>
      <mods:role>
            <mods:roleTerm authority="marcrelator" type="text">photographer
            </mods:roleTerm>
            <mods:roleTerm authority="marcrelator" type="code">pht
            </mods:roleTerm>
      </mods:role>
</mods:name>
```
[31]

In Dublin Core, the entire name heading must be entered into a single creator, contributor, publisher, or subject element, as appropriate:

```
<dc:creator>Cushman, Charles Weever, 1896-1972</dc:creator>
```
[32]

Note that choosing to expose metadata formats that provide for more granular encodings of names will allow more robust use of this data by OAI service providers.

Multiple names should be encoded in separate fields and never packed into a single field. For example,

```
<dc:creator>Donohue, Timothy G.</dc:creator>
<dc:creator>Salo, Dorothea</dc:creator>
```
[33]

4.10.3. DATES IN SHARED RECORDS

- Date elements in shared records should contain values important for discovery of the resource by end-users.
- When providing multiple dates in a shared record, clearly indicate the relationship of each to the resource, and repeat the relevant date element for each date.
- Include easily parsable values in date elements whenever possible.
- Present dates in a consistent format, according to established machine-readable standards.
- Format dates consistently within an OAI PMH set or repository.

4.10.3.1. USE OF DATES

Dates can be a very important part of a shared record. Service providers often use dates to allow sorting of search results or the limiting of search results. The date often appears in the brief display. Users may use dates to help select the appropriate resources or to take advantage of sorting or limiting capabilities.

[31] **http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=GetRecord&metadataPrefix=mods&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07959**

[32] **http://oai.dlib.indiana.edu/phpoai/oai2.pvp?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07959**

[33] **http://www.ideals.uiuc.edu/dspace-oai/request?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:www.ideals.uiuc.edu:2142/11**

4.10.3.2. CHOICE OF DATES

It is a best practice to provide a date if at all possible and that the date provided refers to an event(s) most relevant to discovery of the resource by end-users. This is particularly true when oai_dc is the only metadata format provided because there is no good way to indicate the differences between multiple dates.

For unpublished born-digital content, the best date to use is often the date of creation of the digital resource. For published resources that were born-digital, the most relevant date is generally the date of publication.

For analog resources that have been digitized, the choice is more complex. However, in general, the date of creation of the original analog resource, not the date of creation of the digital reproduction, is the most relevant to end-users searching for the resource. The inclusion of the date a resource was digitized is often confusing for users.

Following is an example of useful date information:

```
<dc:title>Whistle a tune</dc:title>
<dc:creator>Von Tilzer, Albert [composer]</dc:creator>
<dc:creator>Fleeson, Neville [lyricist]</dc:creator>
<dc:publisher>New York: A.V.T. Music Pub.</dc:publisher>
<dc:date>1922</dc:date>
<dc:identifier>http://digital.library.ucla.edu/apam/librarian?ITEMID=AVWAT
</dc:identifier>
```
[34]

Here, the date reflects when the sheet music was published, rather than when this particular digital version of it was created. This allows a service provider to provide access to this material by the date it was written.

Following is an example of confusing date information:

```
<title>Lieutenant General Jubal Anderson Early C.S.A.: Autobiographical Sketch
and Narrative of the War between the States</title>
<creator>Jubal Anderson Early</creator>
<publisher>Philadelphia; London: J. B. Lippincott Company, 1912</publisher>
<date>2003-04-24T13:15:52Z</date>
```
[35]

Here, the date in the `<date>` element reflects when the digital resource was created or made accessible. Note that the date the analog resource was published is buried in the publisher element. Although this example reflects greater adherence to the Dublin Core one-to-one principle (see Section 4.7) than the previous example, the record as it stands is confusing and difficult for a service provider to process. In particular, a service provider would find providing access by 1912 (the date the analog text was published) difficult.

4.10.3.3. USE OF MULTIPLE DATES

Multiple dates (for example, date of creation and date of last modification) should be used only if they can be properly distinguished from one another. When providing multiple dates, repeat the relevant date elements for each distinct date value. Dates in your shared records should be as unambiguous as possible. Ideally it should be obvious from the record itself what events the dates in it refer to. This, however, is not possible in unqualified Dublin Core; dates carefully distinguished in another metadata format will map to multiple `<dc:date>` elements with no distinction between them. In some cases, this problem can be solved by "dumbing down" or crosswalking only the most important date when creating a unqualified Dublin Core expression of a richer metadata record.

[34] **http://digital.library.ucla.edu/oai/sheetmusicdp?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:librart.ulca.edu:sheetmusic/AVWAT**

[35] The links to the records used as confusing examples are not provided because the working group did not want to single out particular organizations, but readers should know that these are real examples.

Following is an example in MODS, distinguishing between two types of dates. Note also the use of the keyDate attribute. Use of this attribute confirms for the service provider that the <dateCreated> element is the appropriate element to index.

```
<mods:titleInfo>
      <mods:title>Marlene's Katy Palomares Road Alameda county</mods:title>
</mods:titleInfo>
<mods:name type="personal">
      <mods:namePart type="family">Cushman</mods:namePart>
      <mods:namePart type="given">Charles Weever</mods:namePart>
      <mods:namePart type="date">1896-1972</mods:namePart>
</mods:name>
<mods:originInfo>
      <mods:dateCreated encoding="w3cdtf" keyDate="yes">1955-04-08
      </mods:dateCreated>
      <mods:copyrightDate encoding="w3cdtf">2003</mods:copyrightDate>
</mods:originInfo>[36]
```

Following is the same item represented in unqualified Dublin Core, displaying only the most relevant date for retrieval purposes:

```
<dc:title>Marlene's Katy Palomares Road Alameda county</dc:title>
<dc:creator>Cushman, Charles Weever, 1896-1972</dc:creator>
<dc:date>1955-04-08</dc:date>[37]
```

It may be tempting, when using a metadata format that does not allow the encoded qualification of dates, to put information in the actual element value qualifying a date, such as the following:

```
<title>Champaign County, Illinois</title>
<creator>United States. Agricultural Adjustment Administration</creator>
<creator>Woltz Studios, Inc. Aerial Survey</creator>
<date>Created: 1940-06-19</date>
<date>Issued: 1940-01-01</date>
<date>Scanned and Processed: 1998-06-01</date>
<type>image</type>[38]
```

This practice, however, makes it difficult for a service provider to process the metadata as the service provider will need to separate out the label from the value and determine which dates to use for indexing purposes. As stated above, it is a best practice to expose only the most relevant date for discovery of the resource (in this case either 1940-06-19 or 1940-01-01).

[36] **http://oai.dlib.indiana.edu/phpoai/oai2.pvp?verb=GetRecord&metadataPrefix=mods&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07958**

[37] **http://oai.dlib.indiana.edu/phpoai/oai2.pvp?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07958**

[38] The links to the records used as confusing examples are not provided because the working group did not want to single out particular organizations, bot readers should know that these are real examples.

4.10.3.4. FORMATS OF DATES

The first consideration when choosing a date format are the requirements of the metadata format in use. The recommendations below should be followed if your format of choice provides for multiple formats or gives no guidance at all.

Ideal Practice

Ideally, dates should be in a standard, machine-readable format, such as ISO8601[39] or, if the date represented is a single date with a known year, the W3CDTF profile of ISO8601.[40] These machine-readable dates allow service providers to easily process them for chronological sorting of search results and date searching of resources.

Following is an example in unqualified Dublin Core:

```
<dc:title>Marlene's Katy Palomares Road Alameda county</dc:title>
<dc:date>1955-04-08</dc:date>
```[41]

Following is an example in MODS:

```
<mods:titleInfo>
      <mods:title>Marlene's Katy Palomares Road Alameda county</mods:title>
</mods:titleInfo>
<mods:originInfo>
      <mods:dateCreated encoding="w3cdtf" keyDate="yes">1955-04-08</mods:
      dateCreated>
      <mods:copyrightDate encoding="w3cdtf">2003</mods:copyrightDate>
</mods:originInfo>
```[42]

Adequate Practice

When it is not possible or practical to encode dates in ISO8601 or W3CDTF, it is essential to format dates consistently and to give service providers enough information in the format used to interpret the dates properly. Pick a format that is commonly used or understood within your knowledge domain, to better allow service providers within that community to use the records. Once you choose a format, use it consistently within your repository or within a set. Provide as much information as possible about the format chosen and how to interpret it in a set description. The following examples show a date format that, if used consistently, could be understood and parsed by a service provider.

Example in unqualified Dublin Core:

```
<dc:title>Africa - Stores</dc:title>
<dc:description>Grocery store in Tangier bazaar.</dc:description>
<dc:date>[ca.1924]</dc:date>
```[43]

Example in MARCXML:

```
<marc:datafield tag="245" ind1="0" ind2="0">
      <marc:subfield code="a">Africa - Stores</marc:subfield>
      <marc:subfield code="h">[graphic].</marc:subfield>
</marc:datafield>
```

[39] See **http://www.iso.org/iso/en/prods-services/popstds/datesandtime.html**.

[40] See **http://www.w3.org/TR/NOTE-datetime**.

[41] **http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07958**

[42] **http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=GetRecord&metadataPrefix=mods&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07958**

[43] **http://memory.loc.gov/cgi-bin/oai2_0?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:local.loc.gov:loc.pnp/cph.3a03261**

```
<marc:datafield tag="260" ind1=" " ind2=" ">
      <marc:subfield code="c">[ca.1924]</marc:subfield>
</marc:datafield>
<marc:datafield tag="520" ind1="0" ind2=" ">
      <marc:subfield code="a">Grocery store in Tangier bazaar.</marc:subfield>
</marc:datafield>
```
[44]

4.10.4. SUBJECTS IN SHARED RECORDS

- Choose subject values from relevant controlled vocabularies consistently and explicitly.
- Repeat subject information in more specific fields when they are available in the metadata format used.
- Express multiple subjects in repeated fields.

4.10.4.1. USE OF SUBJECTS

In addition to the guidelines presented in this section, *Descriptive Metadata Guidelines for RLG Cultural Materials* (RLG 2005) contains useful advice on subject headings in metadata records.

Subjects are an important part of a shared record. Service providers often index subjects with title and description elements. When working with homogeneous metadata, subjects can be used to browse aggregations, and subject terms often appear in brief display. Users will use subjects to find and select appropriate resources. The subject values describe subject content represented in or by the work and typically answer such questions as *who*, *what*, *where*, and *when*. There are, of course, some cases when subject terms may not be appropriate. However, in many instances, using appropriate subject values can greatly enhance users' ability to locate relevant works.

4.10.4.2. CHOICE AND FORMAT OF SUBJECTS

It is a best practice to use relevant controlled vocabularies for subjects consistently and explicitly. The controlled vocabularies chosen (there may be more than one used) should be relevant to the resource and known to the community to which the resources described would hold the most interest. The metadata format(s) used should allow the unambiguous specification of the vocabulary serving as the source of the subject headings applied to shared records. The use and specification of a controlled subject vocabulary allow an OAI service provider to provide improved search services and to build browse indexes—particularly among all data providers using the same controlled vocabulary.

Depending on the materials being described, vocabularies such as the Library of Congress Subject Headings (LCSH), National Library of Medicine's Medical Subject Headings (MeSH), Sears List of Subject Headings, the Library of Congress Thesaurus for Graphic Materials I: Subject Terms (TGM I), or the Getty Art & Architecture Thesaurus (AAT) might be good choices. The most specific subject term available in the vocabulary should be applied.

Metadata authors may also wish to supply uncontrolled terms. This should be done as a supplement to a controlled vocabulary.

When using a precoordinated subject system such as LCSH, repeat any information that may also belong in other fields in the metadata record (e.g., geographic place) in the appropriate, more specific field. For example,

```
<dc:subject>Gardening—Florida—History</dc:subject>
<dc:coverage>Florida</dc:coverage>
```

Apply as many subject terms as necessary to describe the item in question. Place each subject term in its own element rather than placing multiple values within a single element (when using a metadata scheme that allows repeating elements). Placing multiple terms in a single element makes it more difficult for a service provider to parse and index the terms contained within it individually.

[44] **http://memory.loc.gov/cgi-bin/oai2_0?verb=GetRecord&metadataPrefix=marc21&identifier=oai:lcoa1.loc.gov:loc.pnp/cph.3a03261**

Following is an example of multiple subject values in separate elements:

```
<dc:title>Map and profile of the proposed Paterson and Dover Rail Road and
Paterson and Ramapo Rail Road.</dc:title>

<dc:creator>Allen, J. W.</dc:creator>

<dc:creator>Paterson and Dover Railroad.</dc:creator>

<dc:subject>Paterson and Dover Railroad.</dc:subject>

<dc:subject>Paterson and Ramapo Rail Road.</dc:subject>

<dc:subject>Railroads--New Jersey--Maps.</dc:subject>

<dc:subject>Iron industry and trade--New Jersey--Maps.</dc:subject>
```
[45]

Packing multiple, unrelated subject values into a single element is not good practice. These will need to be parsed by the service provider. However, this is sometimes unavoidable as the system in use automatically "packs" the subject element when exposing metadata via the OAI PMH.

Following is an example of packing subject values in a single element. Note the variety of punctuation used in the subject field. The service provider will need to figure out what demarcates a new subject value.

```
<dc:title>Lewis (Payne) Powell</dc:title>

<dc:subject>Payne, Lewis, 1845-1865--Portraits; Gardner, Alexander,
1821-1882</dc:subject>

<dcterms:spatial>Washington (D.C.)</dcterms:spatial>
```
[46]

When using only the required minimum unqualified Dublin Core (oai_dc) or another format that does not allow the specification of the source of subject vocabulary, the next best option is to apply terms from a single standard controlled vocabulary consistently within an OAI set or an entire repository, and specify the vocabulary within the set description or Identify response. See Section 4.9 for more information.

If a metadata provider is not using a standard controlled vocabulary, the provider should provide subject access via a local vocabulary or list used consistently across an OAI PMH set or repository and specify the vocabulary within the set description or Identify response. It is a best practice to make a local vocabulary list available via documentation that is publicly available.

4.10.5. LANGUAGE OF CONTENT IN SHARED RECORDS

- Supply a language element when relevant to the resource.
- Format the value of the language element according to the rules of the metadata format in use.
- Express multiple languages in repeated fields.
- Supply the language of the metadata record only in a metadata element specifically designed for this purpose.

4.10.5.1. USE OF LANGUAGE

In the current online environment—particularly when making metadata available via the OAI PMH—making assumptions about the languages users want to find is no longer viable; it is a best practice to provide the language of textual, audio, and visual materials whenever appropriate. Users will use the language of a resource to narrow search results and to select appropriate resources.

[45] **http://memory.loc.gov/cgi-bin/oai2_0?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:lcoa1.loc.gov:loc.gmd/g3811p.rr005130**

[46] The links to the records used as confusing or bad examples are not provided because the working group did not want to single out particular organizations, but readers should know that these are real examples.

4.10.5.2. CHOICE OF LANGUAGE VALUES

It is a best practice is to include the language of a resource when it is appropriate to do so. For textual, audio, and video materials, for example, language can be an important access point for end-users and therefore should be recorded. For graphic materials, however, the material itself may have no obvious connection to a specific language, so none should be recorded.

Following is an example of an item for which a language value is not relevant:

```
<dc:title>Houses in Eureka valley San Francisco from Collingwood near
22nd St.</dc:title>

<dc:creator>Cushman, Charles Weever, 1896-1972</dc:creator>

<dc:date>1955-03-14</dc:date>

<dc:type>Cityscape photographs</dc:type>

<dc:type>StillImage</dc:type>

<dc:identifier>http://purl.dlib.indiana.edu/iudl/archives/cushman/P07697
</dc:identifier>
```
[47]

Following is an example of an item for which a language value provides a useful access point:

```
<dc:language>ger</dc:language>

<dc:title xml:lang="ger">Othello, der Mohr von Venedig</dc:title>

<dc:creator>Shakespeare, William</dc:creator>

<dc:source>Druckausg.: Wien : Wallishausser, 1806 Original: ULB Münster
</dc:source>

<dc:identifier>http://miami.uni-muenster.de/resolver/urn:nbn:de:hbz:6-
85659521923</dc:identifier>
```
[48]

When multiple languages apply to a resource, encode each language value in a separate field for language in your target metadata schema.

Following is an example of multiple languages in a shared metadata record:

```
<dc:title>Central Saint Martins College of Art &amp;amp; Design, Art and Design
Archive and The Teaching Examples Collection</dc:title>

<dc:language>en-GB</dc:language>

<dc:language>de</dc:language>

<dc:language>la</dc:language>
```
[49]

[47] **http://oai.dlib.indiana.edu/phpoai/oai2.php?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:oai.dlib.indiana.edu:archives/cushman/P07697**

[48] **http://miami.uni-muenster.de/servlets/OAIDataProvider?verb=GetRecord&metadataPrefix=ozi_dc&identifier=oai:uni-muenster.de:2227**

[49] **http://ahds.ac.uk/srvc-oai/provider?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:ahds.ac.uk:va-CSM-1**

4.10.5.3. FORMAT OF LANGUAGE VALUES

Metadata formats frequently indicate that language values should use either a code from a specified standard or a term from a controlled list. Language values in shared records should appear in a format dictated by the metadata format in use. If it is possible to specify the controlled vocabulary or encoding scheme in use, data providers should do so.

Two frequently used content standards for language values (both terms and codes) are *ISO 639-2— Codes for the Representation of Names of Languages*,[50] for which the Library of Congress serves as the registration authority, and *RFC 3066—Tags for the Identification of Languages*, [51] an Internet current best practices document. RFC 3066 facilitates the use of both the original two-letter ISO 639 (1988) and three-letter ISO 639-2 (1998) codes, which may be especially good for those who have been using the two-letter language codes.

Note that most service providers have the capacity to expand the codes to the full name of the language. It is not necessary to provide both the code and the full name.

Following is an example from a MODS record with an encoding standard in use and specified:

```
<language>
      <languageTerm authority="iso639-2b"
      type="code">eng</languageTerm>
</language>
```

4.10.5.4. LANGUAGE OF METADATA RECORD

It is a best practice to record the language of the metadata record when the metadata format used provides a specific element or attribute for this purpose. Do not use a language element that is meant for the resource to include the language of the metadata.

Following is an example from a MODS record:

```
<recordInfo>
      <recordContentSource authority="marcorg">DLC</recordContentSource>
      <recordCreationDate encoding="w3cdtf">1982-11-08</recordCreationDate>
      <recordChangeDate encoding="w3cdtf">2002-03-20</recordChangeDate>
      <recordIdentifier source="DLC">30012858</recordIdentifier>
      <recordOrigin>Derived from a MARC record using the Library of Congress
      stylesheet</recordOrigin>
      <languageOfCataloging>
            <languageTerm authority="iso639-2b">eng</languageTerm>
      </languageOfCataloging>
</recordInfo>
```

4.10.6. GEOGRAPHIC PLACE NAMES IN SHARED RECORDS

- Choose geographic place values from relevant controlled vocabularies consistently and explicitly.
- Be as specific as possible when a hierarchy or controlled vocabulary indication is not possible.
- Place geographic place names or locations in most appropriate elements.

[50] **http://www.loc.gov/standards/iso639-2/php/code_list.php**

[51] **http://www.ietf.org/rfc/rfc3066.txt**

4.10.6.1. USE OF GEOGRAPHIC PLACE NAMES

In addition to the guidelines presented in this section, *Descriptive Metadata Guidelines for RLG Cultural Materials* (RLG 2005) contains useful advice on use of geographic place names in metadata records.

Geographic place names are extremely useful for both aggregators and users. Geographic place names could be used to provide alternate, map-based interfaces to aggregated metadata or to provide a faceted analysis of search results. Users will use geographic place names to find and select appropriate resources. It is a best practice to provide geographic place names from a controlled vocabulary in designated fields whenever appropriate.

4.10.6.2. SPECIFICITY

In choosing a geographic place name, the most specific place name or location available should be applied. Use of a controlled vocabulary can assist in providing disambiguated geographic locations. When a controlled vocabulary is not in use or does not require disambiguation, it is recommended that sufficient context be added to the location to differentiate it from other places with the same name. It is a best practice, when the controlled vocabulary in use allows, to not use abbreviations for geographic place names.

It is a best practice to apply as many geographic place name terms as necessary to describe the item in question. Unrelated place name values should be in separate elements.

4.10.6.3. CONTROLLED VOCABULARIES

The best approach for supplying geographic place names or location information is to consistently and explicitly use relevant controlled vocabularies. The controlled vocabularies chosen (more than one may be used if appropriate) should be relevant to the resource described and frequently used within the community to which the materials being described would hold the most interest. Depending on the materials being described, vocabularies such as the Thesaurus of Geographic Names (TGN), the U.S. Geologic Survey's Geographic Names Information System (GNIS), or the Alexandria Digital Library (ADL) Gazetteer may be useful as controlled vocabularies for geographic location information.

It is also worth mentioning that the LCSH contain geographic place name headings. When using a controlled vocabulary like LCSH, which mixes topical and geographic location terms in a single vocabulary, you should map terms that are unambiguously geographic in nature (e.g., Calgary (Alta.)) to a geographic place or location field in your metadata format if one exists, while mapping explicitly topical terms that include geographic subdivisions (e.g., Central business districts--Alberta--Calgary) to a subject or topical field in your metadata format. If possible you should also map the geographic subdivisions to the geographic place or location field in your metadata format if one exists. This allows service providers to take better advantage of the geographic names in their design of the search (or other) service.

4.10.6.4. FORMAT AND ENCODING OF GEOGRAPHIC PLACE NAMES AND LOCATIONS

When using a metadata format that allows repeating elements, each distinct geographic place or set of geographic coordinates should be placed in its own element. For example, if you are describing a text that discusses both Santa Fe, New Mexico, and Boise, Idaho, these two distinct locations would be placed in two separate elements, rather than being packed into a single element.

Hierarchies

When encoding hierarchical data, it is recommended to encode the individual components of the hierarchy, if your metadata format supports this.

MODS is a schema that allows the encoding of hierarchies in geographic places:

```
<subject>
      <hierarchicalGeographic>
            <country>United States</country>
            <state>Kansas</state>
            <county>Butler</county>
```

```
            <city>Augusta</city>
      </hierarchicalGeographic>
</subject>
```

When mapping hierarchical geographic elements into a metadata schema that does not support the encoding of hierarchies, concatenate the components of the hierarchy into a single geographic element. Here is the above example mapped into unqualified Dublin Core:

```
<coverage>United States – Kansas – Butler – Augusta</coverage>
```

Coordinates

Several metadata standards allow the encoding of geographic coordinates. The Dublin Core Metadata Initiative (DCMI) Point Encoding Scheme provides an alternate strategy for encoding a geographic location as "a point in space using its geographic coordinates," as does the DCMI Box Encoding Scheme for encoding a geographic location as "a region in space using its geographic limits." MODS provides elements for the encoding of cartographic data, including coordinates, scale, and projection. It is a best practice to provide geographic coordinates in shared records if expected by the target aggregator or community of practice.

Ideal Practice

Ideally, the metadata format in which shared records are exposed should allow the unambiguous specification of the vocabulary used as the source for each geographic place name applied to those records.

MODS is an example of a metadata schema with an authority attribute that supports unambiguous specification of controlled vocabularies used, encoding of geographic hierarchies, and encoding of geographic coordinates. Note that the expression of geographic place in all of these elements may not be necessary.

```
<subject authority="lcsh">
      <geographic>Santa Cruz County (Calif.)</geographic>
</subject>
<subject authority="tgn">
      <hierarchicalGeographic>
            <continent>North and Central America</continent>
            <country> United States</country>
            <state>California</state>
            <county>Santa Cruz</county>
            <city>Santa Cruz</county>
      </hierarchicalGeographic>
</subject>
<subject authority="gnis">
      <cartographics>
            <coordinates>36 58 27 N, 122 01 47 W</coordinates>
      </cartographics>
</subject>
```

Qualified Dublin Core also allows the specification of controlled vocabularies used:

```
<dc:subject xsi:type="dcterm:LCSH">Central business districts--Alberta--
Calgary</dc:subject>
```

```
<dcterm:spatial xsi:type="dcterm:LCSH">Calgary (Alta.)</dcterm:spatial>
```

```
<dcterm:spatial xsi:type="dcterm:LCSH">Alberta</dcterm:spatial>
```

Good Practice

If limited to unqualified Dublin Core (oai_dc) or another metadata format that does not support the explicit specification of the controlled vocabulary source(s) used to apply each geographic place name, the next best option is to apply terms from a single standard controlled vocabulary consistently within an OAI set or an entire repository. Specify the vocabulary within the set description, Identify response, and/or external documentation. See Section 4.9 on documentation of practices.

Unqualified Dublin Core (oai_dc) is an example of a metadata format that does not support specification of controlled vocabularies used:

```
<dc:coverage>Santa Cruz County (Calif.)</dc:coverage>
```

Include a note in documentation that controlled vocabulary in use is LCSH.

```
<dc:coverage>North and Central America--Canada--Quebec--Matapedia, Lac</dc:
coverage>
```

Include a note in documentation that controlled vocabulary in use is TGN.

```
<dc:coverage>36 58 27 N, 122 01 47 W</dc:coverage>
```

Include a note in documentation that encoding scheme in use is GNIS.

Adequate Practice

At minimum, a data provider should use a local vocabulary or list to apply geographic place names in a consistent fashion across all records in an OA PMH set or repository. Geographic names should be placed in the appropriate geographic element where one exists.

4.10.7. IDENTIFIERS IN SHARED RECORDS

- Include recognized standard identifiers when available.
- Include a URI or DOI linking to the resource when available.
- Explicitly encode the nature of an identifier provided.
- Express multiple identifiers in repeated fields.

4.10.7.1. USE OF IDENTIFIERS

See also Section 4.8, "Linking from a Record to a Resource and Other Linking Issues."

Identifiers are a way to unambiguously identify a resource. For analog materials, identifiers might be a standard record number such as an ISBN or ISSN or a classification number. Digital materials may also have an ISBN or ISSN, but they may also have some sort of standard digital identifier such as a URI (Uniform Resource Identifier) or a DOI (Digital Object Identifier). This section discusses both analog and digital identifiers.

As discussed in Section 4.8, identifiers can be critical to the discovery of the resources (digital or analog) described by metadata. An ISBN provides an unambiguous identifier for a specific edition of a monograph and aids the end-user in finding that specific edition. A DOI will do the same thing for an article in an online journal, for example.

4.10.7.2. SELECTION OF IDENTIFIERS

When a recognized standard identifier (ISBN, ISSN, DOI, etc.) exists for an item (or for the work of which it is a manifestation), it is a best practice to include such an identifier in the metadata. This holds for both digital and analog resources. In cases where no globally recognized standard identifier exists, a local identifier should be included in the metadata.

Ideally the URI or DOI for a digital resource will resolve (i.e., be a URL) to the resource. If the digital identifier does not locate the digital resource, it is important to provide, in addition to the identifier, a URL (Uniform Resource Locator) that will direct a user to the resource. In aggregations of metadata describing digital resources, the link between the metadata and the actual resource is crucial; without it the user will not be able to reach the resource without time-consuming work-arounds, which affects the credibility of both service providers and data providers.

In many cases, a URL will be the only identifier included in a metadata record. A URL may not actually be the "unambiguous identifier" for the resource as URLs often change. Furthermore, the URL included in the metadata record may not actually resolve to the resource itself, but rather to a page with a link to the resource, to a page with the resource and metadata, or to a collection homepage. For more information about linking to the resource, see Section 4.8.

4.10.7.3. FORMATTING IDENTIFIERS

If using a standard identifier, it is a best practice to format the identifier according to that standard. This is especially true for identifiers that are meant to be machine processable such as URIs[52] and DOIs.[53]

When using a metadata format that allows explicit coding of the type of identifier, best practice is to code this information, including information for local identifiers.

Following is an example of coding a standard identifier in qualified Dublin Core:

```
<dc:identifier xsi:type="dcterms:URI">http://mathworld.wolfram.com/
</dc:identifier>
```

Following is an example of coding a local identifier in MODS:

```
<mods:identifier displayLabel="IU Archives number" type="local"> P07959
</mods:identifier>
```

In general, when using identifiers that are not actionable and cannot be adequately referenced within the metadata itself, it is useful to provide a prefix. For instance, ISSNs and ISBNs can be prefixed, allowing them to be more easily understood by humans and machines.

Following is an example in unqualified Dublin Core:

```
<dc:creator>Ashford, Nicholas</dc:creator>

<dc:date>1994</dc:date>

<dc:identifier>ISBN 92-807-1442-2</dc:identifier>

<dc:title>Government Strategies and Policies for Cleaner Production</dc:title>
```
[54]

4.10.7.4. MULTIPLE IDENTIFIERS

Multiple identifiers should be included if they will assist a service provider or an end-user in locating the resource described. However, in the case of digital objects, if the identifiers resolve to multiple versions of the resource, it is important to identify a single primary identifier that a service provider can label or use as the primary link to the resource. For example, only one `<dc:identifier>` element should be included with an actionable identifier (i.e., a URL). Additional `<dc:identifier>` elements might be included with a local identifier if it is not actionable (i.e., an end-user cannot click on the identifier to arrive at the resource). Again see Section 4.8 for a discussion of best practices related to linking to a resource.

[52] See **http://www.ietf.org/rfc/rfc2396.txt** and **http://www.ietf.org/rfc/rfc2732.txt**

[53] See **http://www.doi.org/handbook_2000/enumeration.html#2.2**

[54] **http://dspace.mit.edu/dspace-oai/request?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:dspace.mit.edu:1721.1/1560**

4.10.8. RIGHTS STATEMENTS ABOUT RESOURCES IN SHARED RECORDS

- Include rights information about a resource in the most granular format possible.
- State rights information in plain language intended for the end-user of a resource.
- Supply rights over the metadata record only in a metadata element specifically designed for this purpose.

Information about rights associated with resources, particularly digital resources, is an essential component of a shared metadata record. Users will use rights information to make decisions about which resources they can use and what they can do with the resources. Service providers may use rights information to filter in or out restricted material.

It is a best practice is to include rights information concerning the use of the resource at the most granular level possible, preferably in the appropriate field in the metadata record itself, for example, the `<dc:rights>` element in unqualified Dublin Core. Other possible locations are within a set description or a repository description. However, there is no guarantee that a service provider will make use of this information; thus, it may be hidden from an end-user.

Please note that the `<rights>` container available in the OAI PMH is not for rights pertaining to the resource, but for rights pertaining to the metadata. See Section 3.11.1, "Best Practices for Expressing Rights over Metadata."

In general, the audience for rights statements associated with resources is the end-user. Therefore, the information provided should be as free of legalese and technical jargon as possible. State clearly any restrictions on usage of the digital object, including explicitly mentioning lack of copyright restrictions when the digital object is in the public domain. Also provide contact information for end-users who wish to pursue required permissions for publication, exhibit, or other types of dissemination.

For example,

```
<dc:rights>Materials digitized for The Making of Modern Michigan are either
in the public domain, according to U.S. copyright law, or permission has been
obtained from rights owners. The digital version and supplementary materials
are available for all educational uses worldwide. For further information
contact the MMM project staff (mmm@mail.lib.msu.edu).</dc:rights>
```

If you maintain rights information relating to specific digital objects on a Web site or use machine-processable rights information (such as a Creative Commons license), you may wish to provide a URL in lieu of a textual rights statement. When doing so, you should provide enough textual explanation, along with the URL, to make the purpose of the URL clear to end-users.

For example,

```
<dc:title>Activity Patterns of Wild Rabbit (Oryctolagus cuniculus, L.1758),
under Semi-Freedom Conditions, during Autumn and Winter</dc:title>

<dc:identifier>http://www.socpvs.org/wbp/index.php/wbp/article/view/10.2461-
wbp.2005.1.6</dc:identifier>

<dc:rights>Authors who publish research articles in WBP retain copyright over
their work. This secures their "moral right" to safeguard the integrity of
their work and to have the full work referenced whenever all or part of it is
reproduced. By publishing their research in WBP, authors agree to allow free
and unrestricted non-commercial use of the work by others under the terms of
the http://creativecommons.org/licenses/by/2.5/ Creative Commons Attribution
license.</dc:rights>[55]
```

[55] **http://www.socpvs.org/wbp/index.php/wbp/oai/?verb=GetRecord&metadataPrefix=oai_dc&identifier=oaiwbp.socpvs.org:article/7**

4.10.9. TYPES OF RESOURCES IN SHARED RECORDS

- Present format and type and/or genre information in all shared records.
- Choose type values from relevant controlled vocabularies consistently and explicitly.
- Express multiple type terms in repeated fields.

Information about the type and the format of the described resource can be used for searching, presenting, and sorting materials. Using values from an identified controlled vocabulary greatly increases this potential, allowing service providers to offer improved search functionality.

There are three basic divisions within this category:

- Format – Specifies the physical medium or material of the analog or digital object. For example, the Internet Media (MIME) Type[56] provides standard encodings for different digital formats.
- Type – Specifies the characteristics and general type of content of the resource. For example, the DCMI Type Vocabulary[57] provides a list of general types to describe resources.
- Genre – Specifies a particular style, form, or content, such as artistic, musical, literary composition, etc. Genre terms are generally highly structured and specific, like the Thesaurus for Graphic Materials II: Genre and Physical Characteristic Terms (TGM II).[58] More genre sources are listed at the Library of Congress Source Codes for Genre site.[59]

There is generally some overlap between type and genre; genre may be thought of as a more specific type.

Generally metadata formats have distinct elements for the format (sometimes called physical characteristics) and type. It is a best practice to provide all three elements when possible and to use a controlled vocabulary whenever possible.

Not all metadata formats have a specific element for genre; notably, in the OAI PMH context, Dublin Core does not. In these cases it is a best practice to use the type element to record the genre of the resource. For example:

```
<dc:type>image</dc:type>

<dc:type>still image</dc:type>

<dc:type>Panoramic photographs.</dc:type>

<dc:type>Gelatin silver prints.</dc:type>
```
[60]

It is a best practice to include the type and genre information that is most likely to aid a user in finding and selecting a resource of interest. Like dates, type elements, in particular, are often misused in shared records to record information about multiple versions of a resource. For example,

```
<dc:title>Letter from Joseph Gales to John McDonogh, March 5, 1830</dc:title>

<dc:description>One-page letter from Gales thanking McDonogh for his check on
behalf of Gerrit Smith.</dc:description>

<dc:date>1830-03-05</dc:date>

<dc:language>en</dc:language>

<dc:type>image</dc:type>

<dc:format>jpeg</dc:format>
```
[61]

[56] **http://www.iana.org/assignment/media-types/**

[57] **http://dublincore.org/documents/dcmi-type-vocabulary/**

[58] **http://www.loc.gov/rr/print/tgm2/**

[59] **http://www.loc.gov/marc/sourcecode/genre/genresource.html**

[60] **http://memory.loc.gov/cgi-bin/oai2_0?verb=GetRecord&metadataPrefix=oai_dc&identifier=oai:loca1.loc.gov:loc.pnp/pan.6a27910**

[61] The links to the records used as confusing or bad examples are nor provided because the working group did not want to single out particular organizations, but readers should know that these are real examples.

In this example, while the resource described is a textual object, the type is given as image. This is most likely because the type is describing the digital object, which is a jpeg, not the physical resource itself. However, if a user, searching for handwritten letters from a particular period, had narrowed their search to "text," this record would not have appeared in the search results.

Format elements are also sometimes used to record administrative information that is not useful in a shared record. For example,

```
<dc:title>Idaho & Washington Northern Railroad and Connections. (1910)
</dc:title>

<dc:type>Maps</dc:type>

<dc:format>Original maps were scanned in color at 600 dpi on a Microtek 9600XL
scanner and saved as TIFF files. The TIFF files were converted into the MrSID
format at a compression ratio of 12 to 1 using LizardTech's Geospatial Encoder
1.5 software. These MrSID files were then uploaded into the CONTENTdm database
at the Washington State University Libraries.</dc:format>

<dc:format>image/jpeg</dc:format>
```
[62]

It is a best practice not to include such administrative information in a shared record.

When using a metadata format that allows repeating elements, place each value in its own element rather than placing multiple values within a single element.

4.10.10. BIBLIOGRAPHIC CITATIONS IN SHARED RECORDS

- Provide a bibliographic citation either to identify the work described in a metadata record or to reference a related work.
- Format the bibliographic citation using standards applied in your community of practice.

There are two common instances where a metadata record might include a bibliographic citation:

- As an identifier to the work itself
- As a reference to another work (which may indicate the place of a work in a hierarchy or may be an unspecified relationship)

A bibliographic citation, particularly in the context of E-print servers and institutional repositories, can provide a user enough context, for example, to link a prepublication version of a paper to the published version of the paper. Journals that make metadata about articles available via the OAI PMH should also include a bibliographic citation within each record to provide important identifying information about the resource.

4.10.10.1. IDENTIFYING THE WORK ITSELF

Many metadata formats have elements to include citation information. In qualified Dublin Core, for example, `<bibliographicCitation>` refines the Identifier element. In this case, when attempting to include a citation to the described work itself, the citation should be in either the `<bibliographicCitation>` refinement or the Identifier element, depending on whether the record is simple or qualified Dublin Core. For instance, if the resource is an article for a journal, it is appropriate to include very specific information about the article, even page references, if such information is used to cite the article in a standard format for reference by other resources, even if the article being described is in a digital format.

Because Dublin Core elements and refinements are repeatable, it is possible to have a URL and a bibliographic citation in the same record, each in a distinct Identifier element.

Following are examples in qualified Dublin Core:

```
<dcterms:bibliographicCitation>ESOP, v.2, no. 1, Apr. 2003, p. 5-8
<dcterms:bibliographicCitation>

<dcterms:bibliographicCitation>Nature, v.87, p. 200
<dcterms:bibliographicCitation>
```

[62] The links to the records used as confusing or bad examples are nor provided because the working group did not want to single out particular organizations, but readers should know that these are real examples.

Following is an example in MODS:

```
<titleInfo>
      <title>Non-subject-matter Outcomes of Schooling</title>
</titleInfo>
<part>
      <detail type="volume">
            <number>99</number>
      </detail>
      <detail type="issue">
            <number>5</number>
            <caption>no.</caption>
      </detail>
      <extent unit="page">
            <start>131</start>
            <end>146</end>
      </extent>
      <date encoding="w3cdtf">1999</date>
</part>
```

4.10.10.2. IDENTIFYING RELATED WORKS

Bibliographic citations to related works should be in elements that allow the relation of the described work to another work. When including a citation to a higher level in a hierarchy, make sure the citation is only to that level. For instance, when the citation is to a journal in which the article appeared, the citation may not include the entire citation to the article (although it may cite to the issue).

Following are examples in qualified Dublin Core:

```
<dcterms:isPartOf>ESOP, v.2, no. 1, Apr. 2003</dcterms:isPartOf>
<dcterms:references>Nature, v.87, p. 200<dcterms:references>
```

4.10.10.3. FORMAT OF BIBLIOGRAPHIC CITATIONS

In all cases, best practice is to use a standard citation format, appropriate to the community that will be using the information. If using Dublin Core, see Apps (2005) for guidelines. Use of particular citation formats can be documented at the repository or set level—it is not possible to document at the element level. See Section 4.9 for further discussion of documentation.

4.11 BEST PRACTICES FOR TECHNICAL ASPECTS OF METADATA

This section contains information on using XML schemas and namespaces and character encoding issues.

4.11.1. XML SCHEMAS AND NAMESPACES

- Use XML schemas endorsed by relevant communities.
- Use XML namespaces when required to validate XML metadata against a given XML schema.

As pointed out in Section 2, "General Areas of Competency," working knowledge of XML, XML namespaces, and XML schemas is fundamental to being an OAI PMH data provider. Following are some basic pieces of information important to providing shareable XML metadata. The NSDL XML FAQ from the *Metadata Primer* (NSDL 2005), the NSDL's XML, Namespaces, and Schemas FAQ,[63] and the Wikipedia entry on XML[64] are helpful resources for beginners.

4.11.1.1. XML SCHEMAS

XML schemas are a way to indicate the expected structure of XML documents. Using a machine-readable grammar; they allow for machine validation of the contents of an XML file.

The OAI PMH requires that every OAI PMH response validates against the OAI PMH XML schema at **http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd**. Thus, service providers can anticipate the format of the information they will harvest, and both data and service providers can automatically validate responses against the XML schema to ensure consistency. Note that the OAI PMH schema deliberately nests multiple XML schemas: a single OAI `ListRecords` response uses the OAI PMH XML schema for the OAI response elements, an XML schema for the metadata format of the records, and additionally, XML schemas for any `<about>` containers associated with the records.

To reiterate: every OAI response, including every metadata record you provide, must be XML-schema valid. If a particular metadata record has an element not allowed by the XML schema for its metadata, then that particular record will not be valid. If a particular record has a bad value according to the XML schema for its metadata, then that record will not be valid. For example, metadata in the oai_dc format must validate according to the XML schema provided by OAI at **http://www.openarchives.org/OAI/2.0/oai_dc.xsd**, which allows only the fifteen unqualified Dublin Core elements. Additional metadata formats served by an OAI PMH repository must validate to the XML schema indicated both in the `<ListRecords>` responses and in the `<ListMetadataFormats>` response. More information about this is included below.

When choosing which XML schema to use for a given metadata format, best practice is to use XML schemas that have been officially vetted by specific communities, governing agencies, etc. A benefit is that the schema will have been thoroughly tested for completeness, errors, and compliance with related standards. This is demonstrable through the widespread use of the schema by other users in a given community. See Section 4.5, "Potential Metadata Formats for Use with OAI PMH" for some metadata formats with official XML schemas.

In instances where no officially vetted XML schema exists, providers may opt to generate their own. Of course, the XML schema itself must validate. It is also a requirement for the XML schema creator to make the schema persistent and accessible online so that documents bound to the specification can be effectively validated. Any new version of the schema should replace the older version at the same location and be backwards compatible whenever possible. The older version should be archived. It is also advisable for schema creators to generate crosswalks that can effectively orient other users to given concepts and data elements extant in related metadata formats.

As far as possible, the data provider should utilize an existing XML schema. If the data provider needs additional elements, it should develop a schema. This new XML schema should refer to appropriate namespaces when using concepts from existing schemas. For example, if the schema is a profile of the Dublin Core element set with two additional elements, all elements referring to the Dublin Core concepts should be labeled with the appropriate Dublin Core namespaces.

More information on XML schemas can be found from these general resources:

- The XML Schema Primer – **http://www.w3.org/TR/xmlschema-0/**
- The XML Schema Specification – **http://www.w3.org/XML/Schema**

[63] **http://metamanagement.comm.nsdlib.org/NSDL_XML_FAQ.html**

[64] **http://en.wikipedia.org/wiki/Xml**

The XML schema validation of an XML document (including an XML schema itself) can be tested with a variety of XML schema tools, including the following:

- The online W3C Schema validator – XSV- **http://www.w3.org/2001/03/webdata/xsv W3C Schema validator**
- A variety of tools are listed at the XML Schema page – **http://www.w3.org/XML/Schema**

4.11.1.2. XML NAMESPACES

XML namespaces serve as mechanisms to contextualize or scope information in XML instance documents. For example, in a registrar's office, `<pass>` in an XML document may mean a student succeeded in a course, while `<pass>` in an XML document about soccer may mean one player has sent the ball to another player. XML namespaces prevent "name collision"—when a single name has an ambiguous meaning because it means different things in different contexts. XML namespaces are used to disambiguate XML information, such as XML element or attribute names.

XML namespaces must be a valid URI. However, there is no requirement that a namespace URI be resolvable: many XML namespace URIs will return nothing if used as URLs in a Web browser, even though they use the http: URI scheme.

XML namespace declarations assign URIs to XML namespaces. Let's examine the following XML document:

```
<oai_dc:dc
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/">

      <dc:title>NSDL Metadata Primer</dc:title>

</oai_dc:dc>
```

There are two namespaces and two elements in the example. The two namespaces are declared with the "`xmlns:oai_dc`" and "`xmlns:dc`" attributes of the outer element. "`xmlns`" indicates an XML namespace declaration ("XML namespace" = xmlns). The attribute name characters after "`xmlns:`" are the XML namespace prefix that will be used to indicate this namespace in qualified names in the XML document. The value of the xmlns attribute is the URI for the namespace. Note that the oai_dc namespace URI, **http://www.openarchives.org/OAI/2.0/oai_dc/**, is a URL, but does not resolve to anything if it is entered into a Web browser. However, this URL is clearly in the domain of the Open Archives Initiative, which provides the definition for the XML elements in the namespace. Thus, the namespace is associated with the organization responsible for setting the scope of the namespace, but in this case, it does not resolve to anything. The namespace declaration `xmlns:dc="http://purl.org/dc/elements/1.1/"` assigns the URI **http://purl.org/dc/elements/1.1/** to the "dc" namespace prefix.

XML namespace prefixes are used in a qualified name: the characters before the colon in a qualified name are the element's "namespace prefix," while the characters after the colon are the "local name" for the element. In our example, the outer element's qualified name is "oai_dc:dc"; it has local name "dc" and is scoped to the namespace URI indicated by the namespace prefix "oai_dc". The inner element's qualified name is "dc:title"; it has local name "title" and is scoped to the namespace URI indicated by the namespace prefix "dc." Note that the "dc" before the colon in "dc:title" refers to a namespace URI, while the "dc" in "oai_dc:dc", since it is after the colon, is the local name.

XML requires that you have an XML namespace declaration for each namespace prefix you use in your XML. The OAI PMH requires the use of namespaces and, hence, their declaration in your served XML. Further, most XSLT engines require strict adherence to XML namespaces.

There can be any number of XML schemas for a single namespace. One of the reasons it is suggested that namespaces be assigned in a domain controlled by the issuing organization is to encourage all schemas written for that format to adhere to the same concept of the format. For example, the Dublin Core Metadata Initiative provides documentation on usage for simple and for qualified Dublin Core at **http://dublincore.org/documents/usageguide/**. It also provides sample XML schemas for each, but organizations may use their own XML schemas for qualified Dublin Core, as does the National Science Digital Library, for example.

Default namespace declarations have a null namespace prefix. These namespace declarations look like this:

```
<dc xmlns="http://www.openarchives.org/OAI/2.0/oai_dc/" >
      <title xmlns="http://purl.org/dc/elements/1.1/" >NSDL Metadata Primer
      </title>
</dc>
```

Note that the two examples above are semantically the same: they have the same locally named elements, scoped to the same namespace URIs, and contain the same values. The fact that different namespace prefixes are used in the second example is NOT a semantic difference in XML.

If you use the null namespace prefix, the OAI PMH requires that you must have a default XML namespace declaration to indicate the appropriate namespace URI.

The second example above declares a second default namespace declaration in the `<title>` element. This illustrates that XML namespace declarations have a scope within the XML document: each XML namespace declaration (default or not) pertains to the element in which it is declared and all that element's children, unless it is superseded by a namespace declaration for the same prefix in one of its descendants. In our second example, the default namespace URI is **http://www.openarchives.org/OAI/2.0/oai_dc/** for the outer element, but the `<title>` element overrides the default namespace, declaring it to be URI "**http://purl.org/dc/elements/1.1/**"—which is true for the `<title>` element and all of the `<title>` element's children. However, the closing tag `</dc>` is again using the default namespace URI of **http://www.openarchives.org/OAI/2.0/oai_dc/** because we are no longer in the `<title>` child element or any of its descendants. Thus, the `</dc>` tag is correctly parsed as the closing tag for the first `<dc>` tag.

More information on XML namespaces can be found from these general resources:

- xml.com's "XML Namespaces by Example" – **http://www.xml.com/pub/a/1999/01/namespaces.html**
- Ronald Bourret's XML Namespaces FAQ – **http://www.rpbourret.com/xml/NamespacesFAQ.htm**
- Wikipedia entry on XML namespaces – **http://en.wikipedia.org/wiki/XML_namespace**
- World Wide Web Consortium's "Namespaces in XML" – **http://www.w3.org/TR/REC-xml-names/**
- Jenni Tennison's "Handling Namespaces" for XSLT – **http://www.jenitennison.com/xslt/namespaces.html**
- XML Namespaces by James Clark – **http://www.jclark.com/xml/xmlns.htm**

4.11.1.3. BINDING XML SCHEMAS TO NAMESPACES

XML namespaces are additionally used as part of the mechanism to bind XML instance documents to particular XML schemas. Note that there is no XML schema indicated for either of the two examples above.

XML schemas are actual documents, not abstract concepts (such as namespace URIs), so XML schema locations are indicated with URLs that resolve to an actual XML schema document. The location for an XML schema is indicated with the `schemaLocation` attribute, which resides in the schema instance namespace (see **http://www.w3.org/TR/xmlschema-0/#ref40**). It is conventional to use "xsi" as the namespace prefix for the schema instance namespace, so the qualified name of the attribute is `xsi:schemaLocation`. Do not forget to properly declare the namespace URI for the "xsi" prefix with a namespace declaration: `xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"`. The value of `xsi:schemaLocation` should be the namespace URI followed by a blank, followed by the URL for the appropriate XML schema for the indicated namespace. The following is one way to correctly indicate the OAI PMH schema in an OAI PMH response:

```
<OAI-PMH
xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.
openarchives.org/OAI/2.0/OAI-PMH.xsd" >
      xml body
</OAI-PMH>
```

Following is an example taken from the OAI PMH specification. In this example, the namespace is **http://www.openarchives.org/OAI/2.0/oai_dc/**, and the XML schema's URL is **http://www.openarchives.org/OAI/2.0/oai_dc.xsd**.

```
<oai_dc:dc
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
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" >

      <dc:publisher>Los Alamos arXiv</dc:publisher>

      <dc:rights>Metadata may be used without restrictions as long as the oai
      identifier remains attached to it.</dc:rights>

</oai_dc:dc>
```

As another example, the XML schema at **http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.02.xsd** for the namespace "**http://ns.nsdl.org/nsdl_dc_v1.02/**" would be indicated like this:

```
<nsdl_dc:nsdl_dc
xmlns:nsdl_dc="http://ns.nsdl.org/nsdl_dc_v1.02/
xsi:schemaLocation="http://ns.nsdl.org/nsdl_dc_v1.02/
http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.02.xsd">
```

4.11.1.4. HOW TO INDICATE XML NAMESPACES AND SCHEMAS IN OAI REQUESTS AND RESPONSES

OAI Metadata Prefix

A namespace prefix in XML associates a local name with the appropriate namespace declaration and therefore the URI for the namespace scoping the name. For example, `<oai_dc:dc>` has "oai_dc" as a namespace prefix, which indicates "dc" is scoped to the namespace URI indicated in the namespace declaration "`xmlns:oai_dc=...`" on the nearest ancestor element.

A metadata prefix in the OAI PMH is the string used to uniquely identify a particular metadata format for an OAI repository. `metadataPrefix` is a required argument for `ListRecords`, `GetRecord`, and `ListIdentifiers` requests. An OAI repository's mappings from metadata prefixes to metadata namespace URIs and their XML schemas are exposed via `ListMetadataFormats`. This is explained in Section 3.4 of the OAI PMH specification.[65]

While the OAI PMH reserves "oai_dc" as a metadata prefix, no XML namespace prefixes are dictated in OAI PMH. In fact, the OAI PMH metadata prefix and the XML namespace prefix in the OAI response may differ; however, it is strongly recommended that the same characters be used in both contexts.

ListMetadataFormats

Every metadata format served by an OAI repository must have a namespace URI and an XML schema. These are exposed with `ListMetadataFormats` responses. This is explained in Section 4.4 of the OAI PMH specification.[66]

What follows is an example response for an OAI server providing two metadata formats: the required oai_dc, and the NSDL's version of qualified Dublin Core.

```
<OAI-PMH
xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">

      <responseDate>2006-02-08T14:27:19Z</responseDate>
```

[65] **http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces**
[66] **http://www.openarchives.org/OAI/openarchivesprotocol.html#ListMetadataFormats**

```
<request verb="ListMetadataFormats">http://services.nsdl.org:8080/
nsdloai/OAI</request>
<ListMetadataFormats>
      <metadataFormat>
            <metadataPrefix>oai_dc</metadataPrefix>
            <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd
            </schema>
            <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/
            </metadataNamespace>
      </metadataFormat>
      <metadataFormat>
            <metadataPrefix>nsdl_dc</metadataPrefix>
            <schema>http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.02.xsd
            </schema>
            <metadataNamespace>http://ns.nsdl.org/nsdl_dc_v1.02/
            </metadataNamespace>
      </metadataFormat>
</ListMetadataFormats>
</OAI-PMH>
```

The <metadataPrefix> element nested within each <metadataFormat> element must contain the metadata prefix string used to identify this format in OAI requests. It is strongly recommended that this string be the same as the XML namespace prefix used for the namespace URI in XML metadata records.

The <schema> element nested within each <metadataFormat> container must contain the URL for the schema document to be used to validate the XML metadata records, and the <metadataNamespace> element must contain the namespace URI for the metadata format.

4.11.2. CHARACTER ENCODING ISSUES

- Specify the UTF-8 character encoding within the XML declaration.
- Ensure all encodings within an OAI record are valid UTF-8.
- Use hexadecimal or decimal numerical entities instead of named entities (except for & < > " ').
- Escape required characters within URLs.

Proper character encoding is essential in XML. Because XML is meant as an interchange format, a given XML document (such as an OAI PMH metadata record) may be opened in many different environments, platforms, etc., each of which may have its own distinct preferences for character encoding. If a default character encoding is not specified by an XML instance, the intended encoding may be lost when used in other environments, often rendering the instance impossible to validate. In less extreme situations, diacritics or references to non-Latin characters may be corrupted.

4.11.2.1. COMMON PITFALLS – XML DECLARATION

Perhaps the most common pitfall with character encoding within XML is not correctly specifying the encoding with the XML declaration. The XML declaration should be the first line of any XML instance document, and it has the standard form, `<?...?>`, of an XML processing instruction. The XML declaration offers an encoding attribute that allows specification of many different Unicode, standard, and proprietary encoding or codepage values. Following are examples:

- `<?xml version="1.0" encoding="UTF-8"?>`
  Used for a variable-length 8-bit Unicode encoding
- `<?xml version="1.0" encoding="ISO-8859-1"?>`
  Used for a Latin-1 single 8-bit encoding
- `<?xml version="1.0" encoding="windows-1252"?>`
  Used for an MS Windows-specific encoding

The OAI PMH requires the use of UTF-8 encoding for all instances, and so all OAI responses must start with the UTF-8 declaration above. UTF-8 is the default encoding for XML, and most modern XML authoring tools will automatically encode new XML instances in UTF-8. However, this is not always the case, and it is not uncommon for an encoder to erroneously use a different encoding scheme when entering diacritics, symbols, and the like because the software is utilizing a different encoding. Always be sure that the XML encoding is set to UTF-8.

There are good links to further documentation for UTF-8 and the Unicode standard at **http://www.utf-8.com/**.

4.11.2.2. COMMON PITFALLS – USE OF CHARACTER ENTITIES

Characters (or code points in Unicode terminology) outside the simple ASCII range 32-127 (to &#x7F;) must either be encoded as multibyte UTF-8 sequences or use numerical entities. In environments that do not natively support UTF-8, it is often easier to use numerical entities. This choice should make no difference to harvesters and service providers because XML parsers accept the two forms interchangeably.

Another common pitfall is for encoders to insert characters such as diacritics and symbols into XML instances using named entities. For example, to use an "a" with an acute accent (á) in an instance, XML allows the use of

- hexadecimal numerical entities (á for a-acute)
- decimal numerical entities (á for a-acute)
- named entities (á for a-acute)—not recommended

The use of named entities will not work for XML schema validated instances used within the OAI PMH. This is because XML schema do not allow for DTD-style entity specifications internal or external to an XML instance. The exceptions are the five named entities that are recognized by XML authoring tools and parsers:

- & = & (ampersand)
- < = < (left angle bracket, less-than sign)
- > = > (right angle bracket, greater-than sign)
- " = " (quotation mark)
- ' = ' (apostrophe)

There are some subtleties as to when the characters > and ' need to be encoded as entities. However, in situations other than in XML markup, it is always safe to encode all five characters using the named entities, and this practice is recommended. For gruesome details, see: **http://www.w3.org/TR/2004/REC-xml-20040204/#syntax**.

It is advisable to use hexadecimal numerical character entities only when the necessary characters cannot be graphically represented within UTF-8 encoding by authoring software. Most operating systems and file systems will support UTF-8 encoding, and most characters and common scientific symbols from any language or discipline can be supported in modern systems and software. Therefore, in reference to the above example using the a-acute, it is recommended to simply enter a "What You See Is What You Get" UTF-8 encoded a-acute directly into the instance using an editor that is UTF-8 aware. However, beware of problems transferring data across operating systems and programs, which may corrupt encoding. We recommend checking the actual XML responses served by your data provider for proper encoding of non-ASCII characters.

4.11.2.3. COMMON PITFALLS – ENCODING URLS INSIDE XML INSTANCES

HTTP (Hypertext Transfer Protocol) URLs make heavy use of characters that often prove problematic for inclusion in XML without escaping. Characters such as ampersands, which HTTP GET URLs use to separate parameter values sent to a Web server, should be escaped in URLs referenced by XML metadata instances. For instance, consider the following URL used inside a Dublin Core Identifier element:

```
<dc:identifier> http://www.ideals.uiuc.edu/dspace-oai/request?verb=ListRecords&
metadataPrefix=oai_dc </dc:identifier>
```

Running this fragment through an XML parser would result in an error, because the parser would see:

```
&metadataPrefix=oai_dc
```

as a malformed entity reference. (Notice the beginning ampersand and the lack of a trailing semicolon.) Remember, XML entity references have the syntax &{variousCharacters}; . The above fragment begins with an ampersand, thus the parser expects an entity. Because the parser thinks the entity reference is lacking the closing semicolon, it returns an error. To get around this, the ampersand must be represented as the entity & as in the following example:

```
<dc:identifier> http://ideals.uiuc.edu/dspace-oai/request=ListRecords&amp;metad
ataPrefix=oai_dc </dc:identifier>
```

When the XML parser receives the & entity, it translates it back into an ampersand (&) in the text of the `dc:identifier`, which will successfully parse and can be used to issue the HTTP request. Note that it is not correct to URL-encode the ampersand as %26; although this will pass through XML parsers correctly, it will invalidate the HTTP GET request. The verb parameter in the example above would be interpreted to have the value `ListRecords&metadataPrefix=oai_dc` instead of there being two parameters, verb, and metadata prefix.

Escaping the characters & < > " and ' must be performed in addition to any URL escaping necessary to build a valid URL (e.g., replacing spaces with %20, or ampersands that do not separate parameters with %26). This is sometimes referred to as "double escaping," but, in fact, all URL-escape characters (which have the form %XX, where X is a hexadecimal digit) will be unaffected by subsequent XML escaping.

4.11.2.4. AVOIDING COMMON PITFALLS

Based on the discussion above, there are several things that data providers can do to avoid character encoding problems:

- Use an editor that supports UTF-8 encoding.
- Make sure you know that it is using UTF-8.
- Use the UTF-8 encoding attribute in your XML documents.
- Use UTF-8 character representations versus character entities when possible.
- Escape problematic characters in complex URLs.

4.12. FINAL PREPARATIONS

Once you have implemented an OAI data provider, it is useful to test the shareability of your metadata—in terms of both its content and technical fitness:

- Look at the metadata that is exposed via your OAI PMH data provider. Issue a `ListRecords` request for each of your sets or for your repository. This can be done through a Web browser. Is the harvest successful? Sometimes harvests will fail because of a character encoding or XML-related issue.
- If the harvest is successful, look at both the oai_dc records and records in other metadata formats. Does the record appropriately describe the resource and give enough context for a user not familiar with it to know what is being described?
- Ask someone not familiar with the resource or the collection to tell you what the resource is, based on the metadata.
- Run the harvested metadata through an XML-validating service. This process should highlight technical problems with the metadata.
- Look at service providers' sites that you would like to have harvest your metadata. Service providers often will have documentation about what they would like to see included in the metadata they harvest and what they need to support the functionality of their service.
- Ask if a service provider will conduct a test harvest on your site before you submit it to the Open Archives Initiative registration service.[67] Some service providers will do a test harvest on your site and give you feedback about technical problems with metadata (validation issues in particular).

Data providers that do these checks and adhere as much as possible to the best practices presented here should have good quality, shareable metadata that will be useful to service providers and ultimately end-users.

[67] **http://www.openarchives.org/data/registerasprovider.html**

5. References

Apps, Ann. 2005. Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata, Dublin Core Metadata Initiative. **http://www.dublincore.org/documents/dc-citation-guidelines/**.

Brogan, Martha L. 2006. *Contexts and Contributions: Building the Distributed Library*. Washington, DC: Digital Library Federation. **http://www.diglib.org/pubs/dlf106/**.

Bruce, Thomas R., and Diane I. Hillmann. 2004. The Continuum of Metadata Quality. In *Metadata in Practice*, ed. Diane I. Hillmann and Elaine L. Westbrooks, 238–256. Chicago: ALA Editions.

Caplan, Priscilla. 2003. *Metadata Fundamentals for All Librarians*. Chicago: ALA Editions.

CEN. 2003. *CWA 14855: Dublin Core Application Profile Guidelines*. Brussels: European Committee for Standardization. **http://www.cen.eu/cenorm/businessdomains/businessdomains/isss/cwa/cwa14855.asp**.

Deutsche Initiative für Netzwerkinformation (DINI). 2003. *Electronic Publishing in Higher Education: How to Design OAI Interfaces*. **http://www.dini.de/documents/OAI-Empfehlungen-Okt2003-en.pdf**.

Digital Library Federation Aquifer Metadata Working Group. 2006. *DLF Aquifer Implementation Guidelines for Shareable MODS Records*. Washington, DC: Digital Library Federation. **http://www.diglib.org/aquifer/dlfmodsimplementationguidelines_finalnov2006.pdf**.

Dublin Core Metadata Initiative. 2006. Dublin Core Metadata Element Set, Version 1.1. **http://dublincore.org/documents/dces/**.

Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, and Sarah L. Shreeves. 2005. Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata. In *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, 32–41. New York: ACM Press.

Foulonneau, Muriel, and Sarah L. Shreeves. 2005. *Describing OAI Sets: A Discussion Paper*. **http://hdl.handle.net/2142/80**.

Gill, Tony, Anne J. Gilliland, and Mary S. Woodley. n.d. *Introduction to Metadata: Pathways to Digital Information*. Online edition, version 2.1, ed. Murtha Baca. Los Angeles, CA: Getty Research Institute. **http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html**.

Habing, Thomas G. 2006. OAI Data Providers. Presented at DLF OAI Implementers' Workshop, August 25, 2006, at Stanford University. **http://hdl.handle.net/2142/147**.

Heery, Rachel, and Manjula Patel. 2000. Application Profiles: Mixing and Matching Metadata Schemas. *Ariadne* 25, **http://www.ariadne.ac.uk/issue25/app-profiles/intro.html**.

Hillmann, Diane I. 2005. *Using Dublin Core*. Dublin Core Metadata Initiative. **http://dublincore.org/documents/usageguide/**.

Hillmann, Diane I., and Elaine Westbrooks, eds. 2004. *Metadata in Practice*. Chicago: ALA Editions.

Hochstenbach, Patrick, Henry Jerez, and Herbert Van de Sompel. 2003. The OAI-PMH Static Repository and Static Repository Gateway. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries*, ed. Catherine C. Marshall, Geneva Henry, and Lois M. L. Delcambre, 210–217. New York: ACM Press.

Lagoze, Carl. 2001. Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description? *D-Lib Magazine*. 7(1). **http://www.dlib.org/dlib/january01/lagoze/01lagoze.html**.

Lagoze, Carl, and Herbert Van de Sompel. 2001. The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*, ed. Edward A. Fox and Christine L. Borgman, 54–62. New York: ACM Press.

National Science Digital Library. 2005. *Metadata Primer*. **http://metamanagement.comm.nsdlib.org/outline.html**.

Open Archives Forum. 2003. *OAI for Beginners: The Open Archives Forum Online Tutorial*. **http://www.oaforum.org/tutorial/**.

Open Archives Initiative. 2002a. *Frequently Asked Questions*. **http://www.openarchives.org/documents/FAQ.html**.

Open Archives Initiative. 2002b. *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting*. **http://www.openarchives.org/OAI/2.0/guidelines.htm**.

Open Archives Initiative. 2002c. *Open Archives Initiative Protocol for Metadata Harvesting*. **http://www.openarchives.org/OAI/openarchivesprotocol.html**.

Powell, Andy, Mikael Nilsson, Ambjörn Naeve, and Pete Johnston. 2005. DCMI *Abstract Model*. Dublin Core Metadata Initiative. **http://dublincore.org/documents/abstract-model/**.

RLG. 2005. *Descriptive Metadata Guidelines for RLG Cultural Materials*. Mountain View, CA: RLG. **http://www.rlg.org/en/pdfs/RLG_desc_metadata.pdf**.

Shreeves, Sarah L., and Christine M. Kirkham. 2004. Experiences of Educators Using a Portal of Aggregated Metadata. *Journal of Digital Information* 5 (September 9), no. 3. **http://jodi.ecs.soton.ac.uk/Articles/v05/i03/Shreeves/**.

Shreeves, Sarah L. 2005. The basics of the Open Archives Initiative Protocol for Metadata Harvesting. In *Technology for the Rest of Us: A Primer on Computer Technologies for the Low-Tech Librarian*, ed. Nancy Courtney, 85–107. Westport, CT: Libraries Unlimited.

Tennant, Roy. n.d. *Bitter Harvest: Problems and Suggested Solutions for OAI-PMH Data and Service Providers*. **http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html**.

Wendler, Robin. 2004. The Eye of the Beholder: Challenges of Image Description and Access at Harvard. In *Metadata in Practice*, ed. Diane I. Hillmann and Elaine Westbrooks, 51–69. Chicago: ALA Press.