DARE use of OAI-PMH: guidelines

Version 1.2, March 2007

Metadata for this document

Title	DARE use of OAI-PMH
Creator	Feijen, Martin; Vanderfeesten, Maurice
Subject	DARE repositories; OAI-PMH; harvesting
Description	Guidelines for the use of OAI-PMH within the Dare Programme
Publisher	Stichting SURF
Date	2006-03-22;
Type	Internal report
Format	Text/richtext
Identifier	SURF OZ
Language	Eng
Rights	Copyright Stichting SURF. The text of this document may be used freely, without permission of Stichting SURF.

Document history

Version	Remarks
July 2006	First internal version presented to project manangers
1.1 2006-08-22	2006-08-21: Additional agreements (to metadataprefix naming, datestamp format, set naming, deleted records, resumption token life span, harvest batch size, service window properties, adminEmail for error logging feedback, Prefix & namespace declaration, XML validation, Communication for Repository modification)
1.2 2007-03-22	2007-03-22: Small changes; URL of the dare_didl schema changed to the KNAW Set naming is not manditory, but preferred. Resumption-token life span is set from at least 5 to at least 24 hours Service window example has been removed, HTTP Error 503 will suffice to indicate a repository is down. This nice to have feature can be worked-out in the future. In the namespace section a line has to be changed.

Remark: The examples used for DIDL; do NOT use them literally! For the precise use of the DIDL document see the current version of the DIDL document specification. That document will overrule all DIDL examples mentioned here.

Acknowledgements This document is largely based on discussions between repository managers and SURF. They have offered their experience and suggestions to create the guidelines as presented in this document.

Source material The guidelines are based on and refer to the Open Archives Initiative Protocol for Metadata Harvesting, Protocol version 2.0. See: http://www.openarchives.org/OAI/openarchivesprotocol.html
The order of presentation of the guidelines is the same as in the protocol text.
When useful, the protocol text is quoted. When the text has been changed, e.g. bold added to highlight some part of the text, this has been indicated between brackets.

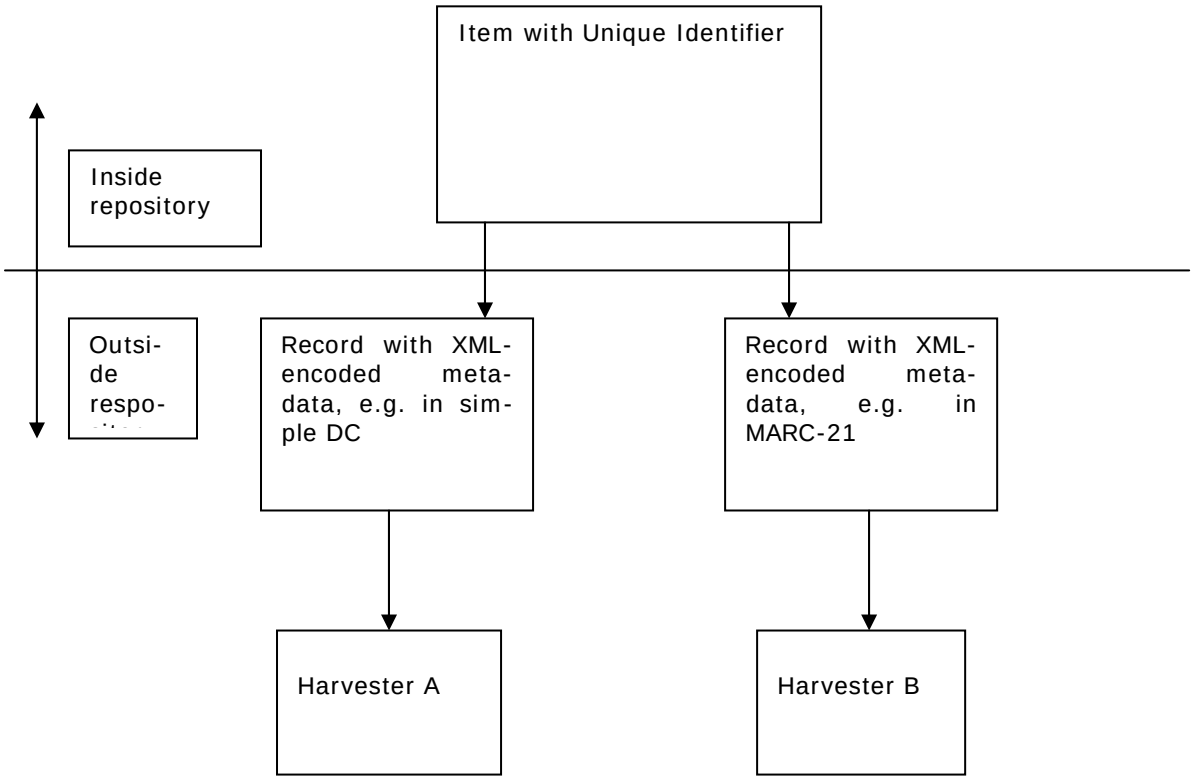
Table of Content

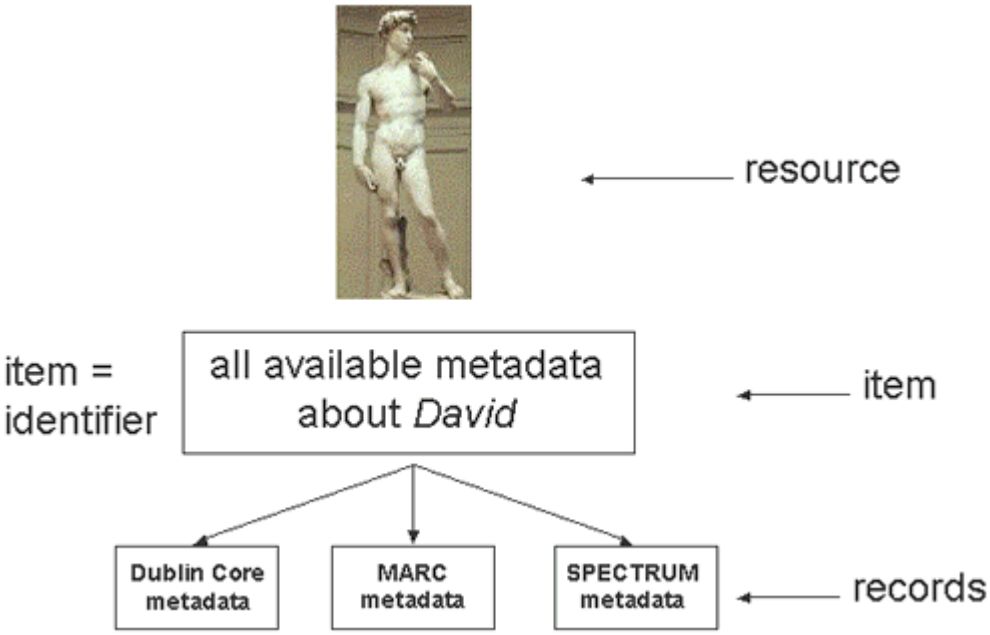
Definitions and concepts: item, record and unique identifier

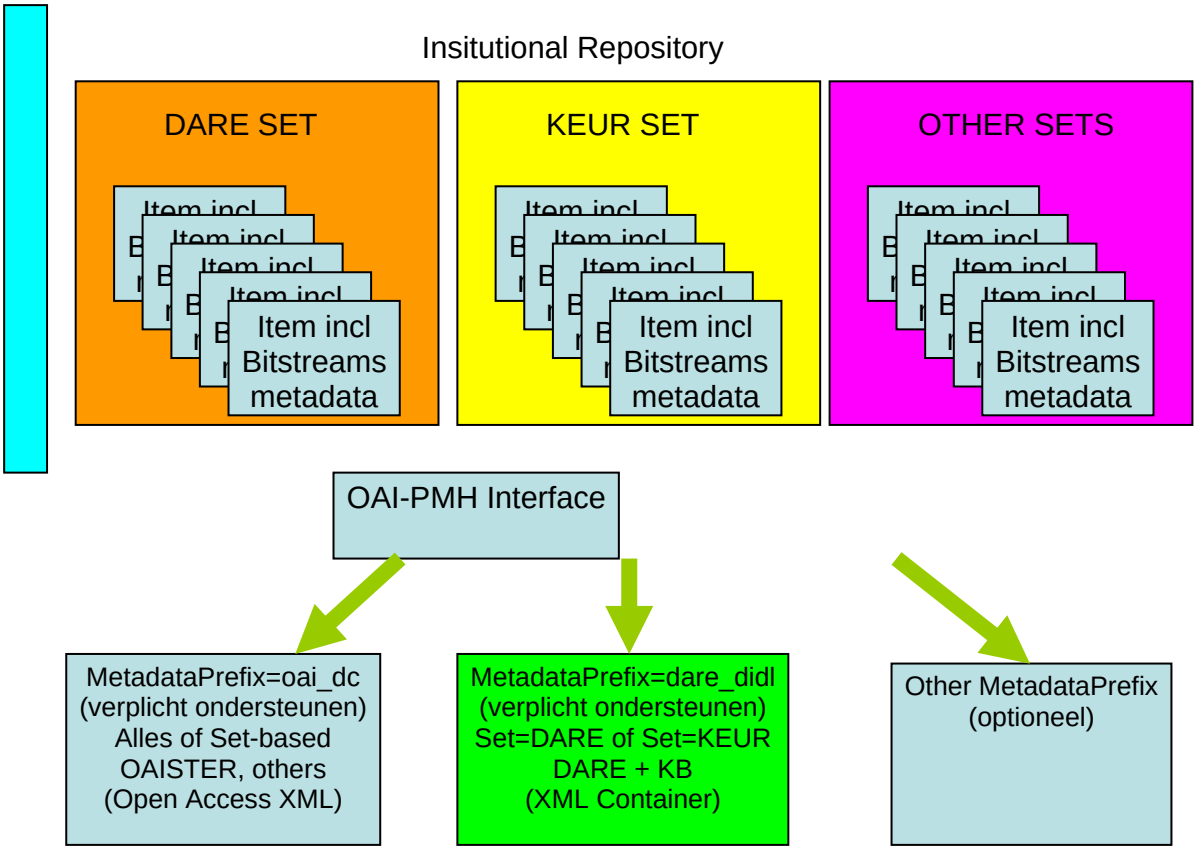
It is important to make a distinction between Item and Record. The protocol text states:
"...An item is conceptually a container that stores or dynamically generates metadata about a single resource in **multiple** formats, each of which can be harvested as records via the OAI-PMH [...]A record is metadata expressed in a **single** format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item...[bold added by MF]

Within DARE, the XML-encoded stream is constructed according to the XML-Container specifications. These specifications are given below.

The unique identifier identifies an item within a repository. Do not confuse this identifier with the element dc:identifier in Dublin Core. The OAI identifier has a different function: it is used to extract metadata, whereas the DC identifier is used to extract the resource.
Schematically:




resource
item =
identifier
all available metadata
about David
item
Dublin Core
metadata
MARC
metadata
SPECTRUM
metadata
records


Insitutional Repository
DARE SET
KEUR SET
OTHER SETS
Item incl
Bitstreams
metadata
OAI-PMH Interface
MetadataPrefix=oai_dc
(verplicht ondersteunen)
Alles of Set-based
OAISTER, others
(Open Access XML)
MetadataPrefix=dare_didl
(verplicht ondersteunen)
Set=DARE of Set=KEUR
DARE + KB
(XML Container)
Other MetadataPrefix
(optioneel)

Additional agreements and recommendations

These additional agreements are based on the Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0 found at http://www.openarchives.org/OAI/openarchivesprotocol.html.

These guidelines will provide additional agreements on the implementation of OAI-PHM for a smooth operation between repository and harvester in the DARE network

Metadataprefix naming, datestamp format, set naming, deleted records, resumption token life span, harvest batch size, service window properties, adminEmail for error logging feedback, Prefix & namespace declaration, XML validation, Communication for Repository modification.

MetadataPrefix naming

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces

OAI-PMH supports the dissemination of records in multiple metadata formats from a repository. The `ListMetadataFormats` request returns the list of all metadata formats.
metadataPrefix arguments are used in `ListRecords`, `ListIdentifiers`, and `GetRecord` requests to retrieve records, or the headers of records that include metadata in the format specified by the metadataPrefix.

For purposes of interoperability, repositories must disseminate Dublin Core, without any qualification. Therefore, the protocol reserves the metadataPrefix 'oai_dc', and the URL of a metadata schema for unqualified Dublin Core, which is http://www.openarchives.org/OAI/2.0/oai_dc.xsd. The corresponding XML namespace URI is http://www.openarchives.org/OAI/2.0/oai_dc/.

dare_didl

The DARE community supports the implementation of the metadataPrefix 'oai_dc' and the the metadataPrefix 'dare_didl'. The schema for the XMLcontainer dare_didl is located at **http://www.repository.knaw.nl/web/dare_didl.xsd**.
The corresponding XML namespace for dare_didl namespace URI currently is **http://www.repository.knaw.nl/web/dare_didl**.
Every DARE repository **must** support this 'dare_didl' metadata schema.
The specification of the dare_didl XMLcontainer can be found in the document XMLcontainer1.1.1.pdf. at the location: http://www.darenet.nl/upload.view/DARE_DIDL-XMLcontainter-Specification-v1.1.1.pdf

```
<OAI-PMH ...>
  <...>
    <record>
      <metadata>
       <didl:DIDL>
         <didl:Container>...</didl:Container>
```

Datestamp format

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Datestamp
, http://www.openarchives.org/OAI/openarchivesprotocol.html#Dates
and http://www.w3.org/TR/NOTE-datetime
The value of datestamps in both requests and responses must comply to the specifications for UTCdatetime in this document. The DARE agreement supports the use of optional granularity which involves the time with seconds **`YYYY-MM-DDThh:mm:ssZ`**.
This value complies with the specifications for the UTCdatetime in sections 3.3.1 in the OAI-PMH document. Datestamps are encoded using ISO8601 and are expressed in UTC.

```
<OAI-PMH ...>
  <...>
    <GetRecord>
      <record>
       <header>
         <datestamp>2001-12-14T12:01:45Z</datestamp>
```

A repository that supports YYYY-MM-DDThh:mm:ssZ should indicate so in the Identify response.

```
<OAI-PMH ...>
  <...>
  <Identify>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  <...>
```

Set naming

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Set

The OAI-PMH document says the folowing:
Repositories **may** organize items into sets. Set organization **may** be flat, i.e. a simple list, or hierarchical.

The DARE agreement is that DARE repositories support at least two type of sets. The 'dare' set and the 'keur' set. Both sets are **flat** and do not have any hierarchical structure.

The table below shows the highly **preferred** setName and setSpec that can be used for either set.

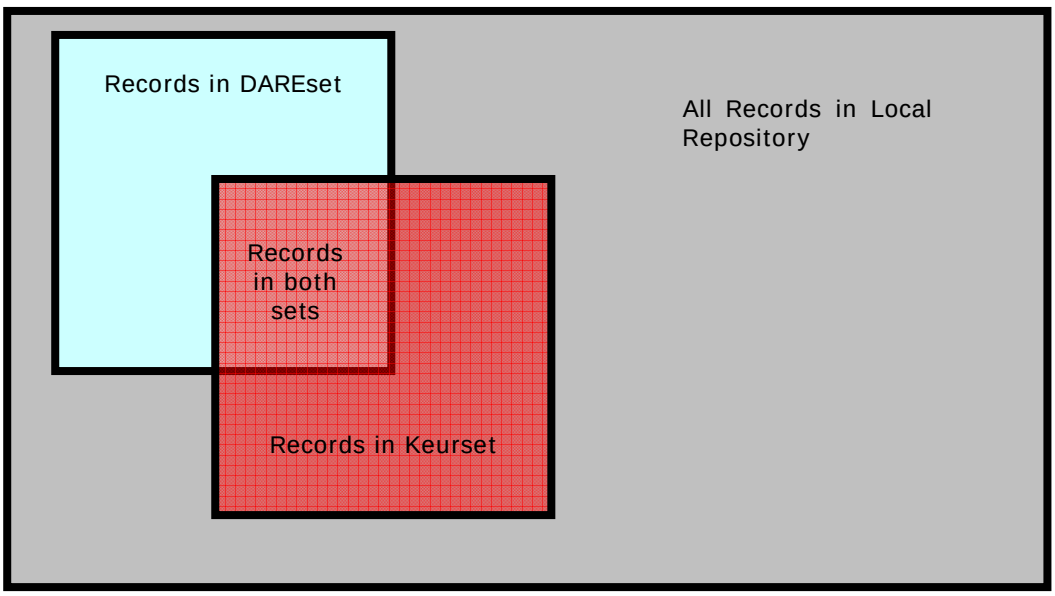
	setName	**setSpec ***
The DARE set	DAREset	dare
The Keur set	Keurset	keur

*A harvester only uses the setSpec request to perform selective harvesting. The letters must be in smallcaps.

Set Content

The specific content of the setSpec is determined at the local repository.
A DARE repository using these kind of sets must conform to the following rules when inserting a record into one/both of these sets. DARE uses this guideline to globally define the content of both sets:

- The *DAREset contains* records that must contain an object that is open accessable for a normal internet user. (What kind of objects/records is left to the local repository.)
- The *Keurset contains* records, which is a collection of all the published work of the institutions leading scientists. These records do not have to contain any object. When it contains an object it does not have to be open access, but it is highly recommended.



Set Location

The DAREset and the Keurset **can** each be located at a different location/baseURL.

Deleted records

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#DeletedRecords

If a record is no longer available then it is said to be deleted. Repositories must declare one of three levels of support for deleted records in the deletedRecord element of the Identify response:

- *no* - the repository does not maintain information about deletions. A repository that indicates this level of support must not reveal a deleted status in any response.
- *persistent* - the repository maintains information about deletions with no time limit. A repository that indicates this level of support must persistently keep track of the full history of deletions and consistently reveal the status of a deleted record over time.
- *transient* - the repository does not guarantee that a list of deletions is maintained persistently or consistently. A repository that indicates this level of support may reveal a deleted status for records.

The DARE agreement requests the DARE repositories to use the option '`transient`'. Also '`persistent`' can be used. This option makes the harvester do an easier job to detect deleted records.
Goals: {yet to be filled in}
Use of transient: {yet to be filled in} {verval tijd, in what time does a repository not have to tell a harvester a record is deleted.}

Resumption token life span

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Idempotency
Repositories that implement resumptionTokens **must** do so in a manner that allows harvesters to resume a sequence of requests for incomplete lists by re-issuing a list request with the most recent resumptionToken. The purpose of this is to allow harvesters to recover from network or other errors that would otherwise mean that the list request sequence would have to be started again.

The protocol does not mention the life span of a token. A token life span is the time a repository keeps the token stored in memory, along with the resume information. When the life span is to short, the repository does not give the harvester a reasonable time to return to complete the harvest. When this happens the repository does not comply to the protocol, see above: "*must do so in a manner that allows harvesters to resume...*".

Out of practice: a reasonable time for a token to be kept alive is *at least* 24 (twenty four) hours.
Along with this life span there is an optimal batch size: see section "Harvest batch size".

Harvest batch size

The batch size is the number of records a repository delivers to the harvester for one resumption token.
The agreement is that DARE repositories must set the batch size between 100 and 200 records.
Using this batch size for all DARE repositories will make the harvester operate at optimal performance.

Service window properties

Not a recommendation, but a nice to have feature is a Service window. A service window indicates when a repository is down, or schedulled to go down for maintenance in the `Identify` request. This will prevent the harvester to report unnessesary errors.
Also the service window provides information to the harvesters when it is the most appropriate time to harvest.

Currently the standard **mandatory** message to tell the harvesters the repository is down is with the **http 503 error code**. (Harvesters should nicely deal with this message)

In the future:
Information about the service window can be located in the <about> element in the `Identify` request. An appropriate format for the service window has not yet been developed. The DARE community is striving for to use International standards. At this time the most appropriate standard will be Calendar formats like vCal or iCal.

AdminEmail for error logging feedback

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Identify
The repository must provide an administrator e-mail address for the `Identify` request.
In the near future we want the harvester to give immidiate response to the Repository Adminitrator to inform about the errors this DARE repository is creating.

See table below for an example of usage to the administer e-mail address.

```
<OAI-PMH ...>
  <...>
  <Identify>
    <adminEmail>somebody@loc.gov</adminEmail>
    <adminEmail>anybody@loc.gov</adminEmail>
    <...>
```

The use of an `adminEmail` in the `Identify` request is mandatory, and is also dictated by the OAI-PMH protocol. See below:
“The `Identify` verb is used to retrieve information about a repository.”
“The response must include one or more instances of the following element:

- `adminEmail` : the e-mail address of an administrator of the repository.”

Prefix & namespace declaration

Look at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Record

namespace declarations -- the declarations of the namespaces used within the metadata part, each of which is prefixed with xmlns. Namespace declarations within the metadata part fall into two categories:

- metadata format specific namespace(s) - every metadata part **must** include *one or more xmlns prefixed attributes* that define the correspondence between a metadata format prefix -- e.g. dare_didl -- and the namespace URI (as defined by the XML namespace specification) of the respective metadata format. Some metadata formats employ tags from multiple namespaces, requiring multiple xmlns prefixed attributes -- in the example, there are declarations for both oai_dc and dc.
- xml schema namespace - every metadata part **must** include the attribute xmlns:xsi, the value of which must always be the URI shown in the example, which is the namespace URI for XML schema.
- xsi:schemaLocation -- the value of which is a URI, URL pair; the first is the namespace URI (as defined by the XML namespace specification) of the metadata that follows in this part, and the second is the URL of the XML schema for validation of the metadata that follows.

The recommended use of prefixes and namespaces is that these entities should be declared on the first element of that namespace. This prevents "operational diffiulties", as described in http://www.w3.org/TR/REC-xml-names/#ns-using .
"Using prefixes may lead to operational difficulties in the case where the namespace declaration attribute is provided, not directly in the XML document entity, but via a default attribute declared in an external entity."

Example of the recommended use of prefixes and namespaces.

```
<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"
>
  <...>
  <metadata>
    <didl:DIDL
      xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
      xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
      xmlns:dcterms="http://purl.org/dc/terms/"
      xsi:schemaLocation="
urn:mpeg:mpeg21:2002:02-DIDL-NS http://standards.iso.org/.../didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS  http://standards.iso.org/.../dii.xsd"
    >
      <...>
    </didl:DIDL>
  </metadata>
  </...>
<OAI-PMH>
```

According to the proclamation in the same document (http://www.w3.org/TR/REC-xml-names/#ns-using), the DARE agreement will be that it is also possible to declare prefixes and namespaces in the ancestors of the document.

"The namespace prefix, unless it is xml or xmlns, MUST have been declared in a namespace declaration attribute in either the start-tag of the element where the prefix is used or in an ancestor element (i.e. an element in whose content the prefixed markup occurs)."

Example of the optional , but not recommended uses of prefixes and namespaces.

```
<OAI-PMH
      xmlns="http://www.openarchives.org/OAI/2.0/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
      xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
      xmlns:dcterms="http://purl.org/dc/terms/"
      xsi:schemaLocation="
http://www.openarchives.org/OAI/2.0/
  http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd
urn:mpeg:mpeg21:2002:02-DIDL-NS http://standards.iso.org/.../didl.xsd
urn:mpeg:mpeg21:2002:01-DII-NS  http://standards.iso.org/.../dii.xsd "
>
  <...>
  <metadata>
    <didl:DIDL>
      <...>
    </didl:DIDL>
  </metadata>
  </...>
<OAI-PMH>
```

XML validation

The XML that the repository provides as output has to be initially validated by the DARE repository administrator, and by the DARE harvester administrator on the first harvest.
A DARE repository must provide valid XML according to the OAI-PMH schema, and also to the dare_didl schema.
Validation can be done by using an XML validator (e.g. from altova. www.altova.com) by saving the repository output as an xml document and opening it in the validator.

For a validator to validate an XML document, inside the document the xsi:schemaLocation(s) must be used.

For the <OAI-PMH> schema use:

```
<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"
>
```

For the <didl:DIDL> schema use:

```
<didl:DIDL
      diext:DIDcreated="YYYY-MM-DDThh:mm:ssZ"
            xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
            xmlns:dcterms="http://purl.org/dc/terms/"
            xmlns:diext="http://library.lanl.gov/2004-04/STB-RL/DIEXT"
            xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
            xsi:schemaLocation="
                        urn:mpeg:mpeg21:2002:02-DIDL-NS http://purl.lanl.gov/STB-
RL/schemas/2004-08/DIDL.xsd
                        urn:mpeg:mpeg21:2002:01-DII-NS http://purl.lanl.gov/STB-
RL/schemas/2003-09/DII.xsd
            http://library.lanl.gov/2004-04/STB-RL/DIEXT http://purl.lanl.gov/STB-
RL/schemas/2004-04/DIEXT.xsd"
>
```

For the <oai_dc:dc> schema use:

```
<oai_dc:dc
            xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
            xmlns:dc="http://purl.org/dc/elements/1.1/"
            xsi:schemaLocation="
                        http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd
                        http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd"
>
```

For the <dare_qdc:qdc> schema use:

```
<dare_qdc:qdc
            xmlns:dare_qdc="http://dare.nl/dare_qdc:/2.0"
            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
                        http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/simpledc20021212.xsd"
>
```

For the <qdc:dc> schema use (not tested):

```
<container_qdc:qualifieddc
            xmlns:container_qdc="urn:dc:qdc:container"
            xmlns:qdc="http://purl.org/dc/elements/1.1/"
            xmlns:dterms="http://purl.org/dc/terms/"
            xmlns:dcmitype="http://purl.org/dc/dcmitype/"
            xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
            xsi:schemaLocation="
                        urn:dc:qdc:container
http://dublincore.org/schemas/xmls/qdc/2003/04/02/qualifieddc.xsd

                        http://purl.org/dc/elements/1.1/
http://dublincore.org/schemas/xmls/qdc/2003/04/02/dc.xsd

                        http://purl.org/dc/terms/
http://dublincore.org/schemas/xmls/qdc/2003/04/02/dcterms.xsd

                        http://purl.org/dc/dcmitype/
http://dublincore.org/schemas/xmls/qdc/2003/04/02/dcmitype.xsd"
>
```

Communication for Repository modification

Modification to baseURL, setSpec metadataPrefix, or metadata schema's

When a DARE repository modifies either baseURL, setSpec, metadataPrefix or metadata schema's, that influences the DARE content cycle then: the concerning repository administrator **must** report this to the DARE community and the DARE harvester administrator in particular.

HTTP request format

No further additional guidelines necessary on the use of the HTTP request format for DARE repositories.