OAI Repository Cataloging Procedures and Guidelines

Prepared by Lucas W. K. Mak (August 13, 2005)

Introduction

The purpose of this document is to provide recommended policies, guidelines and hints for creating collection-level description of Digital Library Federation (DLF) member institutions' digital repositories. This document should not be seen at final and definitive as these recommendations are based on issues faced and practices employed by the person at the time of preparing this document. New issues may constitute changes in practices laid out in this document.

This document is in three parts: a) general cataloging policies, b) procedures and recommended practices for individual elements in the metadata schema, c) factors affecting cataloging difficulty, and d) discussions on cataloging policies and framework.

General Cataloging Policies

The purpose of creating collection-level description of digital repositories is to provide a systematic overview of the contents and structure of digital repositories, and legal and administrative responsibilities associated with the digital repositories exposed through OAI-PMH.

Main Source of Information

In principle, the main source of information used for cataloging these repositories would be XML records from OAI-PMH queries. Cataloger can also consult the repository's web interface in providing description if the main source of information (i.e. XML records) does not provide sufficient information. Cataloger should consult guidelines of individual elements for specific source(s) of information.

Hierarchical Presentation of Sub-collections in a Repository

Since the collection-level description also illustrates the hierarchical structure of the repository, each sub-collection (presented as "set" in XML record from "ListSets" query) should be described separately, nested in the collection-level description's XML file according to their position in the repository, and linked back to the repository and other sub-collections through "Super-Collection", "Sub-Collection", and "Associated Collection" elements.

Cataloger should truthfully present the hierarchical relationship without change unless: a) a sub-collection on any level in the hierarchy contains only one sub-collection in its immediate lower level (i.e. sub-sub-collection) and contains identical records as its lower-level sub-collection, or b) ALL items under the same immediate higher level sub-collection/repository are identical in nature, and administrative & legal responsibility, but being grouped into different sub-collections/sub-sub-collections due to difference(s) in subject area or type.

In the first case, cataloger should skip describing the sub-collection and directly create a description for the sub-sub-collection and linked the sub-sub-collection directly back to the repository. This practice means the total omission of the description of that

particular sub-collection. Although it collapses the hierarchy, this practice minimizes redundancy of information.

In the second case, those sub-collections/sub-sub-collections should be combined into a larger single unit which will acquire a higher hierarchical level that equals to the immediate higher level of the original sub-collections/sub-sub-collections. The new larger single unit formed should, a) be described as a whole, and b) take up the identifier of the entity originally in the immediate higher level. To put it simple, cataloger should create a single description for all sub-collections/sub-sub-collections involved, instead of creating separate description for each sub-collection/sub-sub-collection.

On the contrary, cataloger cannot divide a larger unit of items into a number of smaller units under any circumstances. This restriction is due to the fact that cataloger cannot create identifiers for those newly created units as identifier(s) of the unit(s) below the repository level is taken from the "setSpec" which is decided by OAI data provider.

Redundancy of Elements in Hierarchical Description

Since each sub-collection has its own description and all descriptions use the same metadata schema, redundancy of information would be very serious if cataloger did not exclude any element(s) which contain the same value(s) across individual sub-collection-level descriptions and only describe that element(s) in repository-level description.

In general, if all sub-collection descriptions contain any element(s) with the same value(s), cataloger should describe that element(s) in the description of the collection on the immediate higher level of those sub-collections and skip describing that element(s) in individual sub-collection descriptions. If value of a particular element is different across the sub-collection descriptions which are under the same immediate higher collection and on the same level in the hierarchy, that element should be repeated in each sub-collection description but should be skipped in the description of the collection on their immediate higher level.

Elements describing technical aspects (e.g. metadata schema) of the repository/sub-collections tend to be having the same value(s) in descriptions on different levels. Also, if sub-collections within a repository share the same web interface or employ the same technology (e.g. DSpace), elements describing technical aspects in those sub-collection descriptions would probably be having the same value(s).

Input of Special Characters and Letters in Languages other than English

Special characters and letters in languages other than English sometimes exist in URLs, titles of collections, and personal or corporate names. Unfortunately, UIUC OAI Registry database does not support any XML file containing those characters. As a result, cataloger has to translate those special characters into their Unicode source code according to the "Character Map" available in WindowsXP when recording the information.

Authority Control

It is strongly recommended that cataloger check all personal and corporate headings against *Library of Congress Authority File* (http://authorities.loc.gov/). If not found, cataloger should create the heading(s) according to AACR2.

Some elements in the metadata schema require the use of controlled vocabularies, cataloger, in principle, should adhere to that requirement and clearly indicate the encoding scheme used in the XML tag (e.g. xsi:type=dcterms:"TGN").

Cataloging Procedures and Recommended Practices for Individual Elements

Collection Identifier / Sub-collection <dc:identifier>

In repository-level description, "RepoID" assigned by the UIUC OAI Registry is used as the "identifier" for the repository. In order to make it clear the number is the identification number of the repository in the UIUC OAI Registry, prefix "RepoID=" is added in front of the assigned number (e.g. the RepoID of "DSpace at The University of Washington" is 778, so the repository's identifier is "RepoID=778).

If there is more than one collection found in a repository (i.e. two or more sub-collections), each collection should be assigned a unique identifier. Identifier for sub-collection is the "setSpec" assigned by the OAI data provider. The "setSpec" can be found in individual repository's UIUC OAI registry record.

NOTE(S):

- Since "setSpec" is assigned by the OAI data provider, it may change from time to time, periodical update is strongly recommended.
- Cataloger cannot and should not create identifiers; otherwise, there will be mismatch of collection description and contents exposed by OAI data provider.

Title <dc:title>

In repository-level description, information in <repositoryName> of the XML record from an "Identify" query is used as the title of the repository. If there is more than one sub-collection found in a repository, each sub-collection should be assigned a title in sub-collection-level description. Title for sub-collection is the "setName" assigned by the OAI data provider.

NOTE(S):

- Cataloging rules on capitalization, punctuation, and distinction between title proper and other title information are recommended when recording title information.
- Since "setName" is assigned by the OAI data provider, it may change from time to time, periodical update is strongly recommended.

Alternate title <dc:title>

Alternate title is needed when there is a discrepancy between the title information of the repository found in the XML record and title provided by OAIster's description/ title found on the repository website. Also, alternate title should be created if title information contains abbreviation, or is itself an abbreviation. If there is a translation of title information, that translation should be recorded as alternate title. Same considerations also apply to "setName" (i.e. the title of sub-collection).

Description <dcterms:abstract>

Description of a repository can usually be found in the OAIster's collection description if there is any. When there is no OAIster's collection description in UIUC

OAI Registry record, cataloger should consult the "About" page or similar page(s) of the repository website.

Description of individual sub-collections within a repository can usually be found in the "About" page of the sub-collection website.

NOTE(S):

- Cataloger should be aware of possible obsoleteness and errors (e.g. number of records and names of sub-collections) of OAIster's abstract and "About" page.

Physical Characteristics <dc:format xsi:type="dcterms:IMT">

The format(s) of the digital items can be found in <dc:format> of the XML records from the "ListRecords" query. Besides, format(s) can also be checked by browsing the collection through its web interface. Internet MIME type can be found in *IANA : MIME Media Types* (http://www.iana.org/assignments/media-types/) and *Digital Formats for Library of Congress Collections* (http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml).

NOTE(S):

- Cataloger should be aware of the inaccurate or imprecise MIME type shown in "ListRecords" query (e.g. an institution might assign "application/octet-stream" to a ZIP file instead of "application/zip").
- Cataloger should be as specific as possible (e.g. image/jpeg) when describing file format, instead of copying the general format type (e.g. image) in "ListRecords" XML records provided by the OAI data provider.
- If the MIME type starts with a "x" after the forward slash, cataloger should indicate the required software for opening that file in <dcterms:require>.

Physical Characteristics <dc:format>

This simple DC element should only be used when cataloger is describing physical objects which have no digital manifestation in the collection, or when a specific MIME type cannot be found for the digital object(s).

Size <dcterms:extent>

Size of the repository/ sub-collection can be found in the "Record Counts" of individual repository's UIUC OAI Registry record.

NOTE(S):

- If the repository/sub-collection contains only one type of objects (e.g. images), cataloger can specify the type (e.g. 1000 images); however, if it contains a mixture of object types, a general term "records" is preferred to avoid inaccurate description or time-consuming counting/estimation of quantity of individual object types.
- Cataloger should also be aware of the fact that the number shown in "Record Counts" may not be the same as the real number of digital objects available in the repository. This discrepancy may be caused by incomplete/selective expose of repository content through OAI protocol, or multiple objects within a single OAI record. Also, the size of the repository may not be equal to the summation of its sub-collections' content due to duplication of records in different sub-collections and other unknown reasons.

- On the other hand, inclusion of "Size" in repository-level description can be used as a quick check for any update (addition/deletion of objects) of the repository performed by OAI data provider.

Language <dc:language xsi:type="dcterms:ISO639-2">

This element refers to the language in individual items, but not that of the repository web interface. Language can be found in <dc:language> of the XML record from the "ListRecords" query. Cataloger should use the language codes listed in Library of Congress' website (http://www.loc.gov/standards/iso639-2/langcodes.html).

NOTE(S):

- Cataloger should be aware of the existence of "<dc:language> in XML records of collection of graphic materials (esp. photos) from "ListRecords" query.
- Since native languages are not listed in the LC's list, cataloger should use simple DC element <dc:language> when entering names of native languages.

Type

Type can be identified a) in XML records from "ListRecords" query, b) by browsing the collection through its web interface, or c) in "About" page of the collection.

NOTE(S):

- Cataloger should be aware of the possible discrepancy of available item types between web interface and OAI data.

Type <dc:type xsi:type="imlsdcc:Type">

Cataloger should consult the list provided in *IMLS Digital Collections Registry entry form* (http://imlsdcc2.grainger.uiuc.edu/colltest/). If the collection contains two-dimensional graphic materials (esp. photos), cataloger should also consult *Library of Congress' Thesaurus fro Graphic Materials II* (http://www.loc.gov/lexico/servlet/lexico?usr=pub&op=sessioncheck&db=TGM_II) and use appropriate genre & physical characteristic term(s) to provide a more specific description of "Type" information.

NOTE(S):

- Some repositories (e.g. Library of Congress) may include TGMII terms in item-level records. Since repository usually will talk about the nature of the collection, cataloger can consult the "About" page of individual collections if no such information is provided in item-level records.

Type <dc:type xsi:type="cld:CLDT">

Cataloger should consult *Dublin Core Collection Type (CLDType) Vocabulary* (http://www.ukoln.ac.uk/metadata/dcmi/collection-type/) for appropriate terms. If sub-collections exist, "CollectionCollection" should be used for the repository-level description; and terms describing content of items (e.g. CollectionStillImage) are used for the sub-collection-level description only.

Type <dc:type xsi:type=dcterms:"DCMIType">

Cataloger should consult *DCMI Type Vocabulary* (http://dublincore.org/documents/dcmi-type-vocabulary/) for appropriate terms.

Description of this simple DC element is essentially the duplication of <dc:type xsi:type="cld:CLDT"> excluding the prefix "Collection" (e.g. CollectionStillImage → StillImage).

NOTE(S):

- To minimize redundancy, the broader term "Image" is not used simultaneously with "StillImage" or "MovingImage" in cases concerning the latter two although inclusion of the broader term is required by DCMI guidelines.

Rights <dc:rgiths>

Rights information usually can be found in <dc:rights> of item-level records from "ListRecords" query, or "Copyright" page of the repository/sub-collection website. Cataloger can copy the "rights information" or create a pointer to the webpage which contains "rights information". A qualified DC <dc:rights xsi:type="dcterms:URI">should be used when entering URL.

NOTE(S):

- If "in public domain" or "no known rights" is explicitly stated, cataloger should include this information in the <dc:rights> description.
- Since materials in a collection can be contributed and owned by more than one entity, more than one copyrights statement is possible in some repositories/ sub-collections.

Access Rights <dcterms:accessRights>

Access rights information can usually be found in item-level records from "ListRecords" query, or "About"/ "Use permission" page in the repository/sub-collection web interface. "Restricted to institution affiliated personnel" is the most common restriction in access rights.

NOTE(S):

- Since access rights of different sub-collections may be different even they are under the same repository due to institutional policy, conditions placed by donors of collections and other reasons, cataloger should examine each sub-collection carefully.

Accrual Method <cld:accrualMethod xsi:type="cld:DCCDAccrualMethod">

Information about accrual method can be found in "About" page of the repository/sub-collections. Accrual method also can be deduced from submission policy of the repository, or any statement of change in ownership or custody of the collection (i.e. Provenance). Cataloger should consult the suggested vocabularies listed under http://www.ukoln.ac.uk/metadata/dcmi/collection-DCCDAccrualMethod/.

NOTE(S):

- For electronic theses, dissertations, or scholastic articles, the accrual method normally is "deposit" (except the repository/sub-collection is a academic journal).
- Also, certain repository types have a specific kind of accrual method, e.g. the accrual method of repositories using DSpace framework is usually "deposit".
- Cataloger should be aware of possible mixture of accrual methods (e.g. donated by someone and developed through purchase after the donation) involved in a collection.

Accrual Periodicity <cld:accrualPeriodicity xsi:type="cld:DCCDAccrualPeriodicity">

Information about accrual periodicity may be found in "Policy" page of the repository. Cataloger should consult the suggested vocabularies list under http://www.ukoln.ac.uk/metadata/dcmi/collection-DCCDAccrualPeriodicity/

NOTE(S):

- Certain types of repository may have a specific accrual periodicity depending on digital item submission policy. For example, deposit of digital items into DSpace is initiated by individual creators but not the hosting institution; as a result, the accrual periodicity would possibly be "Completely irregular".
- Also, types of materials also determine the accrual periodicity (e.g. for theses and dissertations, the accrual periodicity would possibly be "Semiannual" or "Three times a year").

Accrual Policy <cld:accrualPolicy xsi:type="cld:DCCDAccrualPolicy">

Usually there is no direct information found about accrual policy, either in metadata from OAI data provider or website of the repository. However, accrual policy can be deduced form "accrual method" (e.g. "deposit" implies a "passive" accrual policy). Cataloger should consult the suggested vocabularies listed under http://www.ukoln.ac.uk/metadata/dcmi/collection-DCCDAccrualPolicy/

Custodial History <dcterms:provenance>

Information about change of ownership or custody of a collection can be found in "About" page of individual collections/sub-collections. Statement of "Provenance" should match the idea represented by the term used in "Accrual Method".

Audience <dcterms:audience xsi:type="imlsdcc:Audience">

There will normally be no explicit statement about the target audience of a collection, if purpose of the repository/sub-collection is not stated. Cataloger should consult controlled vocabularies listed under the *IMLS Digital Collections Registry entry form* (http://imlsdcc2.grainger.uiuc.edu/colltest/), or use other terms as appropriate.

NOTE(S):

- If there is no statement about the target audience, cataloger should try to deduce it from the nature(s) and topic(s) of items in that collection, or from the availability of certain types of supplemental materials. For example, if it is a repository for theses, dissertations and research papers, the target audience will be "Scholars/ Researchers/ Graduate students"; or if the collection comes with a lesson plan (i.e. teacher and student resources), that collection will be probably oriented to "K-12 teachers and administrator" and "K-12 students".

Logo <cld:logo xsi:type="dcterms:URI">

There are three kinds of logo associated with the repository/sub-collection. A repository may have logo(s) showing the technology employed, the project on which it is based on, and/or the hosting/contributing institution(s)/department(s). These logos can be found in the web interface of the repository/sub-collection.

NOTE(S):

- Logo should be a graphic icon, rather than a photograph of something.

Subject

Cataloger can deduce the "aboutness" of the collection by browsing the "About" page, title (esp. textual materials, audios, videos), content (esp. graphic materials) of individual items in the collection, and categorization of items in the collection. (Usually the hosting institution calls the "categorization" as "subject".) Though item-level XML records usually contain "subject(s)" of individual items, they are too specific to be used as collection-level descriptors (unless the repository/sub-collection is a very specialized and focused one).

NOTE(S):

- Cataloger should be aware of the discrepancy between subject(s) available throguh the web interface and OAI protocol. Since the purpose of creating collection description is to describe what available through OAI-PMH, cataloger should exclude subject(s) that is not available through OAI and base the final decision of subject analysis on metadata exposed by OAI-PMH.
- If a repository contains journal(s), cataloger can copy its subject heading(s) from library catalog.
- Though collection-level subject description tends to be more general than item-level description, specificity of subject term(s) assigned depends on the scope and focus of the collection. In general, subject terms assigned for a collection of theses and dissertations would be more general than those for a collection of photographs about an event or an area.
- Browsing through individual titles of a collection of theses and dissertations is usually impractical and extremely time-consuming. It is also almost impossible to identify the subject(s) of each thesis or dissertation due to its scholastic nature and limited expertise of cataloger in certain academic fields. When cataloging a collection of theses and dissertations, cataloger can look for names of academic departments/research areas and used them as subject headings.
- Normally there would be no dominant subject(s) in a general thesis and dissertation repository. Usually a single subject area would not contribute more than 10-15% of the content in a thesis collection. It is the cataloger's discretion to decide on what and how many subject terms to be included in the description.
- Subject terms shown in the browsing interface of collection website usually are "keywords" instead of controlled vocabularies of any kind. Cataloger should translate the usable "keywords" into GEM and LCSH controlled vocabularies.
- Also, those "keywords", especially for graphic materials, may be describing objects depicted in the item. Cataloger should avoid describing those "specific objects" but the "general aboutness" represented by the collection.

Subject <dc:subject xsi:type="imlsdcc:GEM">

Cataloger should consult the controlled vocabularies listed under *Subject Element GEM Controlled Vocabulary* (http://raven.ischool.washington.edu/help/about/documentation/gem-controlled-vocabularies/vocabulary-subject). The hierarchical arrangement of vocabularies of GEM is different from that of the LCSH. Terms of GEM are more general than LCSH.

Subject <dc:subject xsi:type="dcterms:LCSH">

Cataloger can use *Classification Web* (http://classificationweb.net/) to build LC subject headings.

NOTE(S):

- Though the absence of subdivision delimiter may create confusion on interpretation of the form subdivision terms (e.g. "Greenbrier (Tenn.)—Photographs" denotes the collection is about photographs of Greenbrier in Tennessee, or is itself a collection of photographs of that place?), cataloger should include form subdivision in subject headings when addition of such information can make the heading more accurate and specific.
- Cataloger should be aware of the use of Thesaurus of Graphic Materials I (TGMI) in some graphic material collections (esp. those of Library of Congress) by OAI data provider. Though TGMI terms look almost identical to LCSH, cataloger should replace them with suitable LCSHs if those TGMI terms are deemed suitable for subject entry.

Spatial Coverage <dcterms:spatial xsi:type="imlsdcc:GeographicName">

Spatial coverage information can be found in <dc:coverage> of item-level XML records from "ListRecords" query, "About" page of the collection, categories shown in web browsing interface of the collection, or title of the collection. Cataloger should consult *LCSH Authority File* (http://authorities.loc.gov/) or *Getty Thesaurus of Geographical Names (TGN)* (http://www.getty.edu/research/conducting_research/vocabularies/tgn/) for controlled vocabularies.

NOTE(S):

- Cataloger should look for geographical terms in LCSH first, since LC terms provide more information about the location the place concerned than TGN.
- To minimize redundancy, cataloger should not repeat a geographical term from one scheme with an equivalent term from another scheme.
- Cataloger should change the parameters in "xsi:type" according to the encoding scheme (i.e. LCSH or TGN) used.
- Though entries in "spatial coverage" may possibly repeat "geographical subdivision" terms in LCSH, such repetition should be tolerated under current description schema.

Temporal coverage <dcterms:temporal xsi:type="imlsdcc:TimePeriod">

Temporal coverage information can be found in <dc:coverage> or <dc:date> of item-level XML records from "ListRecords" query, "About" page of the collection, categories shown in collection's web browsing interface, or title of the collection.

NOTE(S):

- If possible, cataloger should use exact time period, instead of an approximate period, in describing temporal coverage.
- Cataloger should also avoid using name of the period.
- "Temporal coverage" of a photographic collection should normally be the same of its "Contents Date Range <cld:dateContentsCreated>" unless they are photographs of historical object(s).

Accumulation Date Range <dcterms:created>

Accumulate Date Range information may be found in "About"/"History of the collection" page of the collection, or in biographical information about the creator(s) of items in the collection.

NOTE(S):

- "Accumulation Date Range" would normally be identical or similar to "Contents Date Range" if the collection being described is photographs and the photographer is the "collector" of the collection. (However, it is debatable from when and did the photographer treat those photographers as a collection? Also from a theoretical perspective, can a "creator" or items in a collection, at the same time, be treated as the "collector" of the same set of items?)

Contents Date Range <cld:dateContentsCreated>

It is the creation/publication date(s) of the original physical item if the digital collection is originated from physical items. Contents Date Range information can be found in <dc:date> of item-level XML records from "ListRecords" query, "About" page of the collection, or bibliographical information of the creator(s) of items in the collection.

NOTE(S):

- Since most OAI data providers won't specify whether the <dc:date> in XML records is about the "creation date of the item" (i.e. Contents Date Range), or the "Temporal coverage" of the item, cataloger should carefully examine other background information of the item/collection (e.g. date of birth and death of the creator) to decide what the <dc:date> is talking about. If the aboutness of <dc:date> is undetermined and no other information can be found, leave the "Content Date Range" element out from the description.

Collector <dc:creator>

Collector would normally be the same as the "Hosting institution" <dc:publicher>; however, collector can also be the original creator(s) of items in the collection (though this perspective is debatable as stated in "Accumulation Date Range").

Owner <marcrel:own>

Ownership information can be found in rights information and provenance statement in the web interface of the repository/collection or in item-level XML records.

NOTE(S):

- Ownership information should match the statement in <dcterms:provenance> and <cld:accrualMethod>.
- If scholastic communications (e.g. theses, dissertations, research papers, etc.) are involved, cataloger can check the "submission policy" or similar policy of the repository to determine "ownership" information.

Is Located At <gen:isLocatedAt xsi:type="dcterms:URI">

On repository level, "Is Located At" is the homepage of the repository; on sub-collection level, it is the homepage of the sub-collection concerned.

NOTE(S):

- Cataloger should be aware of alternative access point(s) (i.e. homepage) for a repository or sub-collection. If such a homepage is found, it should be included in the "Is Located At" information.
- If the sub-collection has no homepage, use the URL of its "Search" page as substitute.
- If the repository/ sub-collection has no web interface, leave the "Is Located At" element out from the description.

Is Accessed Via <gen:isAccessedVia xsi:type="dcterms:URI">

BaseURL can be found in UIUC OAI Registry records or XML record from "Identify" query.

NOTE(S):

- Since it is the baseURL of the repository, "Is Accessed Via" should only be included in repository-level description.

Sub-Collection <dcterms:hasPart>

Normally, sub-collections would be the same as those seen in the XML record from "ListSets" query, or set record in individual repository's UIUC OAI Registry record. Cataloger should input "setSpec", instead of "setName" into sub-collection description since "setSpec" is the unique identifier within a collection and has no other variant forms.

NOTE(S):

- Cataloger should be aware of the flattened relationship between "setSpec" (sub-collections and smaller units) in repository's UIUC OAI Registry record. Cataloger should examine the hierarchical implications behind naming of "setSpec" by OAI data providers and look for metadata about hierarchical arrangement in XML records especially from "ListSets" query.
- In most cases, description of this DC term would be made on repository level only; however, if a sub-collection contains sub-sub-collections, sub-collection-level description should include <dcterms:hasPart>.
- In order to clearly represent the hierarchical relationship between a repository, its immediate sub-collections, and other lower-level sub-collections in the hierarchy, "sub-collection" description in any level of the hierarchy should not go further than its next lower level (e.g. "Sub-collection" element on the repository-level description should only contains "setSpec" of sub-collections, but not those of sub-sub-collections).
- However, when a sub-collection contains only one sub-sub-collection and records in the sub-collection are the same as those in the sub-sub-collection, cataloger should skip that sub-collection and directly input the "setSpec" of the sub-sub-collection in the repository-level description (with omission of the whole collection description for that sub-collection simultaneously). Although this "skipping" would possibly create a description that describes two levels of collection at the same time, this practice can minimize redundancy.
- In some cases, smaller collections under the same larger unit should be combined into a single larger unit if items across smaller collections are identical in their nature and administrative & legal information, but are categorized into different sub-collections due to differences in topic or type. The combined collection would be described as a

unit (i.e. a single collection description) and using the identifier of their original common next larger unit as their single identifier.

Super-Collection <dcterms:isPartOf>

Information about super-collection can be deduced from same types of information as in "Sub-collection".

NOTE(S):

- Super-collection is not limited to intra-repository relationship. A repository can be a super-collection of a number of smaller repositories if those smaller repositories are each assigned a "setSpec" and listed as "sets" in the larger repository's UIUC OAI Registry record.
- Name and BaseURL of the larger repository should only be included in the "Super-Collection" element in repository-level description.
- Cataloger should check repositories under "Is Friend of" list of a repository's Registry record, and see if the above situation applies.

Catalog or Description <dc:description>

Catalog or description information may be found in "History" page, related resources, suggested readings, or bibliographies in the collection website.

NOTE(S):

- Digital collections that based on a well-established physical collection (esp. archives) may have such catalog or finding aids in existence.

Associated Collection <dc:relation>

Associated collection(s) can be found in "Has Friends" list in the repository's UIUC OAI Registry record. Name(s) and BaseURL(s) of "Friends" are included in the "Associated Collection" element in repository-level description only.

In some cases, associated collection(s) could also be sub-collections in the same repository if their contents are related. These related-collections can be on the same or different hierarchical level(s). Cataloger's discretion is needed in determining such intra-repository relationship.

NOTE(S):

- If another repository/sub-collection has a relationship with the repository/sub-collection currently being described as identified in "Super-Collection", "Sub-Collection" or "Source", that another repository/sub-collection should be excluded from the "Associated Collection" element.
- Wihtin a repository-level or sub-collection-level description, a repository/sub-collection should have either "Super-Collection", "Sub-Collection", "Associated Collection", or "Source" relationship with the repository/sub-collection currently being described.

Source <dc:source>

Source collection may be identified in "About" or "History" page of the collection website. Also, information may be found in <dc:description> of item-level XML records from "ListRecords" query.

NOTE(S):

- The purpose of <dc:source> is to identify the existing physical collection(s) which the current digital collection is based on; therefore, cataloger should not substitute "collection" with titles of individual items which are pooled together to form the digital collection but did not form an existing physical collection beforehand.

Associated Publication <dcterms:isReferencedBy>

Titles of associated publication can be found in "History", related resources, suggested readings, or bibliographies page of the repository/sub-collection.

NOTE(S):

- Since the publication should be "based on the use, study, or analysis of the collection", cataloger should not include news articles or press release about the launching of the repository which are commonly found in "History" page of the repository.
- Publications found in "Bibliographies" page are usually related to the content of the repository; whereas publications found in "History" and "Related resources" page tend to be related to technical aspect of the repository and are usually journal articles.

Hosting Institution <dc:publisher>

Hosting institution can be identified by the URL of the homepage of the repository, or Rights statement of the repository. Normally, the identified hosting institution is the sole entity responsible for the whole repository and its sub-collections if available.

NOTE(S):

- It is not uncommon that an institution hosts more than one repository through its libraries and academic units. As a result, cataloger should be more specific in identifying hosting institution, i.e. if the university library is the host, cataloger should name the "library" rather than the "university" as the hosting institution.
- Usually, "Hosting institution" would also be the "Collector" of the collection/repository.

Contributor <dc:contributor>

Contributor(s) of a repository/sub-collection can be identified in "development history", "statement of collection development responsibility" of the repository/sub-collection, or acknowledgement statement.

NOTE(S):

- Creator(s) holding principal responsibility over significant portion(s) of items in the collection is currently considered as "Contributor(s)" if that person(s) is not named as "Collector".
- If the collection is assembled under a joint project or agreement, other contributing institution(s) is currently identified as "Contributor(s)".

Administrator <imlsdcc:managedBy xsi:type="dcterms:URI">

Administrator(s) can be found in individual repository's UIUC OAI Registry record, or in <adminemail> of XML record from "Identify" query. Cataloger should input the e-mail(s) of the administrator(s) instead of the name(s).

NOTE(S):

- Normally, the administrator(s) is responsible for the whole repository and its sub-collections if available. As a result, the "Administrator" element should be included on repository-level description only to avoid redundancy; however, if individual sub-collection(s) is found to be administered by another entity, cataloger should identify that entity in corresponding sub-collection description.

Interaction with Digital Collection
<imlsdcc:interactivity xsi:type="imlsdcc:Interactivity">

Cataloger should consult suggested terms listed in *IMLS Digital Collections Registry entry form* (http://imlsdcc2.grainger.uiuc.edu/colltest/) and check with terms used in existing collection description. Cataloger can also supply new terms if no suitable term is found. Types of interaction can be identified by examining functions of the repository's/collection's web interface.

NOTE(S):

- If all sub-collections share the same interface (e.g.. using the same presentation software/ framework), type(s) of interaction will normally be the same; therefore, cataloger should include the "Interaction with Digtial Collection" element in repository-level description only to minimize redundancy.

Metadata Schema Used
<imlsdcc:metadataSchema xsi:type="imlsdcc:MetadataSchema">

Metadata schema used can be identified in repository's UIUC OAI Registry record or in XML record from "ListMetadataFormats" query. Cataloger should transcribe the code(s) of metadata schema shown, instead of its full name.

NOTE(S):

- Normally, all sub-collections would use the same set(s) of metadata schema; therefore, cataloger should include the "Metadata Schema Used" element in repository-level description only.
- In exceptional cases where sub-collections use different set(s) of metadata schema, cataloger should identify schema(s) used in every sub-collection description and skip this element in repository-level description.

Supplementary Materials <imlsdcc:supplement xsi:type="imlsdcc:Supplement">

Cataloger should identify any supplementary materials accompanied the collection according to terms suggested in the *IMLS Digital Collections Registry entry form* (http://imlsdcc2.grainger.uiuc.edu/colltest/). Supplementary materials usually have designated web pages in the collection website. "Related resources", "Suggested readings", "Background information", and "Bibliographies" are useful indicators.

Notes <imlsdcc:notes>

Note can be any information that the cataloger think is useful to collection description viewers but is not suitable to be put into other elements in this description framework. A note can indicate, but not limited to, the availability of full-text (not due to differences in access rights), digitization project information, or the size of the original physical collection (if it is different from the digitized collection).

Requirement <dcterms:require>

Cataloger have to provide the name(s) of software required for opening certain type(s) of digital item(s), if the Internet MIMIE type starts with an "x" after the forward slash.

Factors Affecting Cataloging Time and Difficulty

Time for cataloging a repository varies. Some "simple" repositories can be done in an hour, whereas some "complex" repositories can take more than a week to finish. Factors affecting the level of difficulty in cataloging repository include: a) hierarchical structure of the repository, b) availability of sub-collections, c) nature of sub-collections in the repository, d) availability of web interface for the repository/individual sub-collections, e) design of the repository's/sub-collection's web interface, and f) information in XML metadata records from OAI data provider. These factors determine the time used in finding information and difficulty of subject analysis.

Hierarchical Structure of the Repository

In general, a repository with a flattened structure normally implies that it is a relatively small repository. On the contrary, a more hierarchical structure usually implies a well-established and large repository. Cataloger has to spend more time in creating descriptions for sub-collections on intermediate level(s) in a hierarchical repository compared to a more flattened one.

Availability of Sub-collections

If a repository has many sub-collections, cataloger will have to spend more time in cataloging that repository because each sub-collection has to be described separately. Also, existence of a large number of sub-collections normally means the repository is a large one (unless each sub-collection contains only a few records, e.g. in some DSpace repositories). In principle, the bigger the repository, the more time-consuming it is to be cataloged.

On the contrary, organized division of contents into different sub-collections may also supply useful title and abstract information for subject analysis than a repository which has no sub-collection but a pool of digital objects on many different subject areas.

Nature of Sub-collections in the Repository

Collections containing some types of digital objects (e.g. photos) are less complex in subject analysis than collections containing other types of objects (e.g. theses and dissertations). Collection containing graphic materials usually comes with indicative "title" and "about" information of the collection.

The more focused the collection, the easier the subject analysis process. Collection developed from a project, or built by a small number of people (i.e. items in a collection are authored by a few people) usually is more focused than a collection built by deposition.

Availability of Web Interface

Normally, every repository/sub-collection would have its own web interface which contains background information, related resources of the repository/sub-

collection, as well as provides searching and browsing interface to access individual digital objects. Although XML record exposed through OAI-PMH is the main source of information for cataloging purpose, it is always easier to browse the digital contents in its web interface in doing subject analysis. This is specifically beneficial for collections of graphic materials. Moreover, some information exists in both XML records and web interface.

If web interface does not exist, cataloger has to flip through page after page of item-level metadata records to decide on the subject area(s). Time used in this process depends on how many records returned per resumption token.

Design of Web Interface

Some repositories have a unified interface for all of its sub-collections. This creates a predictable pattern which cataloger can easily adapt to and know where to find required information. Unified interface normally guarantees identical value(s) in many elements on technical aspects of the repository. This means time-saving of cataloging procedures by skipping elements in description on different levels and describing those elements in the repository-level description once.

On the other hand, unified interface may work against cataloger. Some interfaces (e.g. DSpace) work with textual objects better than graphic digital objects. If the same interface is used for both textual and graphic collections in the repository, cataloger may have difficulties in browsing those graphic materials when doing subject analysis. For example, cataloger has to rely on titles of individual photos in deciding subject headings since DSpace provides no thumb-nail images. Although DSpace provides subject heading(s) for each item, it is impractical to click into each item to look for that.

Information in XML metadata records provided by OAI data provider

In general, the richer the information provided in XML metadata records, the easier the process of finding information for individual elements in the metadata schema. However, some repositories did not tag the information appropriately in the XML records which may make the rich information become totally useless in creating collection description. As a result, cataloger may have to check against information available in web interface or throw those pieces of information away due to possible inaccuracy caused in final description. For example, some item-level records contain <date> information; however, some repositories did not indicate clearly whether that date is the date of "digitization", "creation of the original physical item", "adding that item into the digital collection" or any other events.

Discussions on Cataloging Policies and Framework

This part of the document is going to discuss some theoretical and practical issues which have to be resolved in order to develop a set of more consistent and theoretically sound cataloging practices.

Combining Sub-collections by Cataloger

It is recommended to combine some sub-collections to form a larger unit under certain circumstances according to the guideline in "Sub-Collection" element description.

This regrouping of sub-collections is an "All-or-nothing" practice restricted by the original arrangement of sub-collections by OAI data providers.

If a repository contains four sub-collections and only three of them are deemed suitable to be grouped together according on the reasons explained in the guideline, cataloger still cannot combine those three into one larger sub-collection. This failure in grouping similar sub-collections is due to the fact that cataloger cannot create an identifier for the newly consolidated sub-collection.

It is tempting to include all original identifiers of the sub-collections involved in this element description, since the maximum occurrence of the element "Identifier" can be "unbounded" according to the Dublin Core Collection Description Application Profile (http://www.ukoln.ac.uk/metadata/dcmi/collection-application-profile/#dcidentifier). However, the purpose of allowing multiple identifiers is believed for providing "alternative" identifier(s) for the WHOLE collection. Identifiers other than the first one should be viewed as "alternative" identifier for the WHOLE sub-collection, instead of pointing to its constituting smaller units. Therefore, any identifier used should represent the WHOLE combined sub-collection, rather than just part of it.

Subject of the Study: Physical vs. Digital

When creating description for digital collection that is derived from a physical collection, there is always confusion about what the subject for the description is: whether the cataloger is describing the physical collection (i.e. source) or the digital collection? It is especially problematic when describing "Accumulation date range", "Accrual method", "Accrual policy", and "Accrual periodicity".

If the cataloger sees the subject being cataloged is the physical collection, describing its "Accumulation date range", "Accrual policy", and "Accrual periodicity" usually will become meaningless. The purpose of describing the growth of the collection is to let service provider know whether there will be new items and how frequent they are being added to the collection. From this perspective, the description should be about the growth of the digital collection rather than the physical one. In many cases, the physical collection, which the digital one is derived from, is based on an already ended project or is a pool of items by a deceased author. "Closed" would normally be the value for "Accrual policy" as there is no more addition to the physical collection, though the digitization process may be still going on. In the same case, there is no need to describe "Accrual periodicity" as no more items will be added. Description of the growth of the physical collection shades no light on that of the digital collection, and means nothing to OAI service providers or repository's users who want to look for new digital items.

On the other hand, cataloger faces another problem if digital collection is decided to be the subject being described. Controlled vocabularies for the element "Accrual Method" usually do not fit in describing addition of "non-digitally-born" digital items. In many cases, the hosting institution digitizes individual items in the physical collection which the institution owns. As a result, neither the concept of "deposit", "donation", nor "purchase" applies in this situation as all of them are talking about the acquisition of the original physical collection. Moreover, the concept of "Loan" and "License" also do not apply since the institution owns the collection (both physical and digital). However, the remaining concept "Item Creation" does not fit the situation. Though the digitization

process can be seen as "creating" digital objects, it is arguable whether the process constitutes a "creation from nothing" since it is essentially a process of "reformatting".

Partial and Passive Accrual Policy

The list of *DCCD Accrual Policy Proposed Term* has no suitable controlled vocabulary to describe a policy that is "passively" adding items to a specific "part" of the collection. On the repository level, it is possible that a repository has some of its sub-collections (i.e. "partial") "passively" adding items, especially one is based on "deposition initiated by creators of individual items".

However, this lacking of suitable vocabulary exerts no substantial effect on the description of accrual policy on repository-level. Since the "partial" concept in the above scenario implies that there are sub-collections adopting accrual policy other than "Partial & Passive". Repository-level description is, nonetheless, unable to truthfully reflect all differences in accrual policy among sub-collections. As a result, "Accrual policy" has to be described separately in each sub-collection description instead of collectively on the repository level. The lacking of vocabulary for "Partial and Passive" accrual policy will have effect only when it happens on the lowest level of the hierarchy where cataloger is not able to substitute the collective description with separate descriptions in smaller units. However, "partial" accrual policy in individual sub-collection is very hard to be identified unless it is stated explicitly by the hosting institution. Also, "partial" accrual policy would normally be used for collection that is constituted by a number of different components (e.g. a collection that contains items created by two persons with one person's part completed but the other one still developing).

Accrual Method for Inter-institutional Project

There is uncertainty in assigning term in "Accrual method" for collection built by a joint effort from different institutions under an agreement or project. Neither "Deposit", "Donation", nor "Purchase" suits this joint project since participating institutions usually retain the ownership of their contributions. Moreover, the "permanent status" of the agreement, hence the collection built, sometimes is in doubt. "License" nor "Item creation" would be the suitable term as this kind of projects normally does not involve any licensing fee, or initiate the creation of items in the collection. Normally, these items are archival records or items already held by participating institution.

It seems that "Loan" – the remaining controlled vocabulary in the list – is the most suitable term for collection based on a joint project since "Loan" implies no transfer of ownership nor involvement of financial payment. However, "Loan" may not be an appropriate term when considering the status of participating institutions in the project. The concept of "Loan" seems to imply that the hosting institution decides the collection content and the addition of items is initiated by a request from the hosting institution. Instead, individual contributing institutions usually retain the right to decide what items to contribute (not deposit, since no transfer of ownership) and initiate the contributions themselves. Strictly speaking, "Loan" is not the appropriate term describing the accrual method of collection initiated by an inter-institutional project.

Definition of "Collector"

"Collector" is "an entity who gathers (or gathered) the items in a collection together". Though the definition is very strict forward, three types of uncertainty are identified in the cataloging process.

Relationship between Creator and Collector

Can the principal creator(s) of items in a collection, at the same time, be the collector of that collection? In some cases, a hosting institution acquired the whole collection (e.g. a set of photographs, boxes of personal papers) directly from its creator(s) or their heir(s). If the collection is a set of archival records, it is the normal practice for an archive or library to keep the original arrangement used by the creator(s). Given the integrity and arrangement of that set of items at the time and after the time of acquisition, a creator was essentially gathering individual items in the collection at the time of creating them. Hence, the process of creation can be seen as the process of gathering items. Nonetheless, it is debatable whether the creator(s) treated those items as a collection "intentionally" and organized those creations "systematically".

On the other hand, it can be argued that "Collector" here is referred to the entity gathers or gathered "digital" items in a collection as opposed to "physical" items discussed above. However, this perspective is not applicable to digital collection based on an existing physical collection, since the "collector" is essentially "reformatting" items which had been gathered already. This perspective is only applicable to collection with digitally born items.

Collection Development after Transferal of Ownership

It is common that the hosting institution acquired a collection through transfer of ownership. If the hosting institution does not further develop the collection (i.e. adding new items) after the act of acquisition, the hosting institution does not assume the role of bringing items together (i.e. gathering) but only acts as an owner of the collection. The collector, therefore, is arguably to be the original collector who transferred the collection to the hosting institution.

On the contrary, if the hosting institution has added new items to the acquired collection, the hosting institution assumes the role of both "collector" and "owner". Now, the question is whether the original collector who transferred the collection to the hosting institution should be considered as "Collector", "Contributor", or none of the two. Should involvement of monetary payment in the transfer of ownership affect that consideration?

Nonetheless, it is hard to tell whether the hosting institution did add new items after acquiring the collection, unless it is explicitly stated. Most of the time, the institution would only mention the event of the acquisition and the approximate quantity of items involved.

Contributor vs. Collector

In some cases, a digital collection is a joint effort in which participating institutions contribute their own existing collections to form a larger collection. Individual institutions are essentially the "collectors" of components of the larger collection. Should participating institutions be considered as "Collector" rather than "Contributor"?

Type, Spatial Coverage, Temporal Coverage, and LCSH

LCSH consists of topical, chronological, geographical, and form elements. According to the current practice and the cataloging framework used, geographical information presented in LCSH is repeated in spatial coverage. Most of the time, values of the geographical element in LCSH are exactly the same as those found in "Spatial coverage". Repetition of chronological information in LCSH and Temporal coverage is less serious since rules for building LCSH limit the use and format of chronological information in the headings. Form/Genre information in the LCSH also repeats those described in the element Type.

However, if cataloger only includes topical information in LCSH and describes the temporal, spatial and form attributes in other elements used in the schema, it will break the linkages between those attributes and may cause misinterpretation. For example, if a collection is about Christianity in Europe and Buddhism in Asia, by putting topical and geographical information into different elements will cause confusion. This confusion and possible misinterpretation will be much more serious when number of subject area and type of information increases.

Repository Re-exposed by Aggregator

There are some DLF member institution's repositories not being harvested directly by UIUC or OAIster, they are being re-exposed by aggregator like languagearchive.org. The re-presentation of the repository by a third party aggregator may not contain accurate record count or set information. Moreover, the baseURL provided by the aggregator is not the original baseURL of the re-exposed repository. Above all, those repositories' status as OAI repositories and their existence are doubtful since the RegistryPop, GoaiglePop, nor OAIster can find their original baseURL. Should the cataloger create collection description for repositories only available through an aggregator?

Alternative Title

Should the "Alternative title" of the repository/collection be tagged by <dcterms:alternative> instead of <dc:title>?