Canadiana Metadata Repository Format (CMR)

Version 1.0 (2010-08-20)

Introduction

Records ingested into the Canadiana Metadata Repository are mapped to a common standard: the Canadiana Metadata Repository (CMR) format. CMR is intended to be a relatively stable intermediary format which will be used to derive application-specific formats (such as a Solr database schema) which are likely to be more volatile. The purpose of CMR is to create a repository of normalized metadata from the diverse collections submitted by multiple contributors, which can then be used to create application specific derivatives on demand. CMR's design goals are to:

- be reasonably simple to understand
- implement strict validity controls which will allow for a simple XSLT transformation to produce derivative records (such as Solr database schemas) without having to modify the text or attribute content of elements
- preserve hierarchical and sibling relationships among records
- avoid imposing an overly arbitrary structure on source record types or relationships
- allow for reasonable and meaningful mappings from most metadata formats
- be capable of storing rich metadata but also able to store sparse records without excessive overhead or the use of dummy/placeholder data.

CMR is therefore meant to be a medium-complexity record, capable of representing more complex relationships and strongly-typed data than a simple format such as Dublin Core, but still easy to understand and manage.

Contributors to the Canadiana Metadata Repository may submit metadata in a variety of formats, which Canadiana will map to CMR as best as possible. Contributors are encouraged to submit their metadata directly in CMR, thereby ensuring the richest and most direct possible mapping of their metadata, taking advantage of the knowledge that contributors have about their own collections and metadata. This in turn will translate into richer, more consistent, and more discoverable metadata within the repository.

This document describes the CMR format and provides some best practice guidelines for contributors wishing to export their records in CMR for use in the Canadiana Metadata Repository. Notes are also provided outlining how the Canadiana Discovery Portal will or is expected to treat certain fields and interpret certain data. The term "should" is used in this document to describe practices which are not required by the standard but are recommended as best practices.

Versioning

Each version of the CMR schema will have a major and minor version number. Minor versions must be backward compatible: a record that validates under the 1.0 schema must also validate under the 1.1 schema. This means that a minor revision may include new non-required elements, add non-required attributes, increase the maximum number of occurrences for a given element, and add to or expand allowable field values. Changes which break backward compatibility will be assigned a new major version.

Schema Version 1.0

Structure

A recordset is a container for one or more records. It is a convenience which allows multiple related records to be kept together, but related records may also be stored separately without affecting the ability to reconstruct those relationships. Contributors should strike a balance between keeping related records together for convenience and keeping file sizes small for ease of processing. In general, all of the records associated with a single title (e.g. a monograph and its pages) should be put into a single file, unless doing so would create an excessively large file. The majority of CMR files should be kept under a few megabytes in size for ease and speed of processing, but larger files are permitted where warranted.

Each record describes a single object, chosen from a list of supported types. Each type implies a certain type of behaviour and will be treated differently by the Discovery Portal.

Recommended best practice is to create records at the level for which both significant metadata exists and direct access can be provided. For example, for a digitized book where a URI can be provided to access each individual page, and for which there is full text should be created using multiple records: a monograph record linking to the item as a whole and, for each page, a child page record linking to that page. If access can only be provided at the book level, or if there is no full text or other significant metadata for the pages or other sub-parts, a single record should be created, linking to the item as a whole.

A record consists of a set of control fields (some required and some optional) a description section and a resource section which may contain both links as well as references to local resources such as master or downloadable files. The latter are generally not applicable in the context of the Discovery Portal, but the same metadata can be used to power content-serving portals such as Early Canadiana Online. One resource: canonicalUri is mandatory. All other description and resource elements are

optional. Since the discoverability of a record is directly related to the quantity and quality of descriptive and control information, it is recommended that as much relevant detail as possible be included, but apart from the basic control information, no single piece of information is required for each record.

Ordering of fields within a record is strict: they must occur in the order described below.

Control Fields

The first four control fields are required for every record and must appear only once per record. They must appear in the following order: `type`, `contributor`, `key`, `label`.

`type`

Indicates the basic record type. A record's type has a variety of effects on the way it is treated by the Discovery Portal. Allowable values include:

- collection
- monograph
- serial
- issue
- page

collection: use to denote a thematic or other collection of multiple items. Any child records should be of one of the following types: collection, monograph, serial. There is no limit to how big or small a collection can be or how many levels of sub-collections can be specified. For example, a collection could encompass an entire digital holdings, a monographic series of three volumes, or anything in between.

monograph: use to denote a standalone item such as a book, photograph, sound recording, or artifact. Any child items should be of the page type. In general, this is the standard, default record type: use this for any record that does not clearly belong in one of the other categories.

serial: use to denote a serially published series, such as a newspaper, journal, or radio program. All children of a serial should be of the issue type. A serial that does not have individual issue records (i.e. there is only one record for the serial as a whole) is functionally equivalent to a monograph.

issue: use to denote an issue that belongs to a serial (e.g. a particular newspaper or journal issue, or a particular episode of a television show). An issue should have a parent record of type serial, and any child records should be of the page type.

page: use to denote a page or similar sub-part of a monograph or issue.

`contributor`

A code which uniquely identifies the contributor. Contributor codes are assigned by Canadiana. Every record must include the contributor's code; contributors must not use any codes but their own.

The purpose of the contributor code is to identify the source of each record, and to create a protected namespace for identifiers: all identifiers (key, pkey, gkey) will be prefixed with the contributor code and a dot (e.g.: oocihm.12345) to create globally-unique identifiers within the repository.

`key`

A unique (to the record contributor) code which identifies this record. Keys must be between 1 and 127 characters in length and can only consist of the characters: `A-Za-z0-9_.-`. This is to minimize issues with query string length and encoding.

Any key can be used as long as it meets the format and length requirements. An effective, though not required strategy is to make each key equal to its parent key with additional information appended. For example:

> a top-level collection: <key>eco</key>
> a monograph within that collection: <key>eco.12345</key>
> a page within that monograph: <key>eco.12345.8</key>

`label`

The default name or label to use when describing the item, for example, in search results or as a main title. Normally, this will be a title, title/statement of responsibility combination, caption, brief description, or similar material.

For issue and page records, label should be specified within the context of its parent item. The Discovery Portal may catenate a record's label with that of its ancestor(s) in order to display it in the proper context. For example:

> a serial: <label>The monthly example</label>
> a child issue: <label>vol. 1, issue 3 (March 1920)</label>
> a page within that issue: <label>p. 18</label>

When displaying the page within, for example, a search result, the Discovery Portal may use a combined label such as:

> The monthly example : vol. 1, issue 3 (March 1920), p. 18

When displaying the same information within the context of the serial, the label may be simply constructed as:

> vol. 1, issue 3 (March 1920), p.18

Context-Specific Control Fields

The remaining control fields are not required. They should be included when contextually appropriate. The gkey, media and lang fields may each occur multiple times. The remaining fields may occur a maximum of once each. The fields must appear in the specified order: `pkey`, `gkey`, `seq`, `pubdate`, `lang`, `media`.

`pkey`

The key value of this item's parent item. E.g., a page should have a pkey equal to the book it belongs to; an issue's pkey should equal the key of the serial it belongs to, and so forth.

Because contributor IDs are implicitly prepended to all keys, it is not possible to specify a parent child (or group: see gkey below) relationship between objects supplied by different contributors.

Circular references should be avoided. In addition, improper parent-child relationships (e.g. a monograph record with a pkey that points to a serial record) should also be avoided. In both cases, the Discovery Portal will detect and ignore such relationships.

A pkey may be safely specified for a record that does not yet exist, but should not be specified for records that are not likely to ever exist.

`gkey`

A group key describing an item other than the immediate parent that this record is a sub-part of. Examples: a page in an issue might have a gkey equal to the key of the containing serial, and all items might have a gkey equal to the collection to which they belong (or the collection may be set the direct parent, in some cases). Multiple gkeys are allowed.

The recommended use of gkey is for a child record to have a gkey linking to every ancestor apart from the immediate parent. A gkey may occasionally also be used specify a non-ancestor collection record which is used to describe a thematic set that crosses normal hierarchical boundaries, but care must be taken to avoid circular references.

`seq`

A positive integer describing the relative ordering of sibling records (records with the same pkey). For example, the pages in a book can be ordered using this key.

The use of sequential values is suggested but not required: three items with seq values of 10, 20 and 30 are functionally equivalent to the same items with values of 1, 2, and 3. No two records should share both the same parent and seq values. Doing so may result

in the random suppression or ordering of the conflicting records within the Discovery Portal, or other undefined behaviour.

`pubdate`

Describes the range of possible publication dates for the item. Dates must be full ISO 8601 dates, e.g. `1800-12-25T12:00:00.000Z`. The primary use of pubdate is to allow for sorting based on date and faceting based on date range.

The value of min should be equal to or less than the value of max. Both are required. For normalization and interoperability reasons, The full ISO 8610 date and time must be specified in GMT. The date should reflect the full range of dates during which the item or items described by the record are known or are thought to have been published, or as close an approximation as is reasonably possible. Some examples:

A book published in 1856:
<pubdate min="1856-01-01T00:00:00.000Z" max="1856-12-31T23:59:59.999Z" />

A photograph taken sometime during the 1920s:
<pubdate min="1920-01-01T00:00:00.000Z" max="1929-12-31T23:59:59.999Z" />

A newscast broadcast at 1PM on July 1st, 1981 and running 30 minutes:
<pubdate min="1981-07-01T13:00:00.000Z" max="1981-07-01T13:30:00.000Z" />

A journal starting publication in April 1843 and running to August 1888:
<pubdate min="1843-04-01T00:00:00.000Z" max="1888-08-31T23:59:59.999Z" />

Date range faceting in the Discovery Portal will match any record whose pubdate range overlaps with the requested range. E.g.: a date range of 1885–1925 (which would expand to: 1885-01-01T00:00:00.000Z–1925-12-31T23:59:59.999Z) would match the example photograph and journal, but not the book or broadcast.

`lang`

A 3-letter ISO 639-3 language code relating to the content of the item. If a work is multilingual, this field can be specified multiple times. The primary use of this field is to allow for faceting based on language.

It is important to note that standards such as ISO 693-2 PartB and the MARC and CANMARC language codes are similar to ISO 693-3, but some individual codes differ. (E.g. "fra" and "deu" in ISO 693-3 versus "fre" and "ger" in ISO 693-2 Part2B and MARC.) It is important for maximum consistency and interoperability, to use the ISO 693-3 codes.

`media`

A code indicating the format(s) of available digital representation with respect to its method of use or consumption. If multiple formats are available, this element may be specified multiple times.

As with lang, the primary use of this element is to allow for faceted searching based on media type. Media type codes should be assigned based on formats which are actually available to the end user, rather than any source or intermediary formats which may exist. The type should be chosen based on the way in which a user would typically expect to consume or use the format.

Allowable types include:

- data
- image
- plaintext
- sound
- text
- video

data: use to indicate structured, machine actionable data. For example: TEI or other structured XML markup, CSV data, or binary data. Program source code (Perl, C, PHP, etc.) should also be tagged as data. Only structured formats which can be meaningfully parsed and manipulated by software should be given this designation: text that is simply marked up for presentation (such as HTML or PDF) rather than programmatic manipulation should be given the text designation.

image: use to indicate an image, regardless of file format (PNG, JPEG, PDF, etc.) other than an image of primarily textual material (e.g. a PDF of a digitized text). For example: a scanned photograph, poster, artwork, or artifact.

plaintext: use only to indicate unstructured encoded text (e.g. ASCII).

sound: use to indicate a sound recording in any format (MP3, AAC, etc.) regardless of its content (music, speech, audiobooks, etc.)

text: use to indicate any material that is primarily textual in nature. For example: a digitized book or page (whether in PDF, JPEG, PNG, etc.) or an HTML web page or other format that is readily consumable by human beings.

video: use to indicate any video recording in any format (MPEG, H.264, etc.) regardless of its content (motion picture, presentation, video footage, etc.)

Description

The description section consists of seven field types, each of which are optional and which may occur multiple times. They may occur in any sort of order within the description element.

The Discovery Portal uses these fields for two purposes: to provide keyword search access to the record and to display information about the record. The contents of each field should be entered the way they should be displayed on-screen to a user. (HTML or other markup should not be used.) Each field should contain one descriptive item. (E.g., if an item has multiple subjects, each should be put in a separate subject field.)

Description should be done at the level of the item being described, and should not be repeated in child records. For example, if a digitized book with OCR text is being entered as a single record, both the bibliographic description (title, author, subject, etc.) and the OCR text (text) should be included in the record. If the item is being entered as a monograph record and a series of child page records, the bibliographic description should be part of the monograph record and the OCR text of each page part of that particular page record.

Much of the information in the description section is being solicited in anticipation that it will eventually be useful, once a certain critical mass has been collected, even though not every field or attribute currently has a specific use in the Discovery Portal. All descriptive information is optional, but the less information provided for a record, the less discoverable it will be, and the less meaningful information the Discovery Portal will be able to provide to the user about the item.

Each descriptive field takes an optional lang attribute, which specifies the language of the field's content, using an ISO 693-3 language code. In the Discovery Portal, specifying English (eng) or French (fra) will affect the stemming and stop word filters applied to the field content on ingest.

Certain descriptive fields have an optional type attribute which can be used to further qualify the field's content using an enumerated list of types specific to that field. The descriptive fields and optional type qualifiers are listed below.

`title`

Use for title or title statement(s), including any alternate, secondary, or parallel titles. Each title should be placed in a separate element. A title may be further qualified with a type attribute in the following cases:

main: designates a title as being the authoritative or standard title for the work. This value should be specified for at most one title element.

uniform: indicates a title is the title of a series, proceedings, conference name, etc., to which this item belongs.

`author`

Use to record the name of an author, composer, artist, editor, or other creator associated with the intellectual content of the work. An author may be further qualified with one of the following type attributes:

editor: indicates that the author's role is one of editor, aggregator, translator, or similar secondary role, rather than being responsible for the primary intellectual content of the work.

`publication`

Use to record a statement of publication, creation date, imprint, or other information related to the time and place of publication or creation. The publication field does not take a type attribute.

`subject`

Use to indicate a subject heading or description. Both controlled vocabularies (e.g. LCSH) and free-form values are allowed. Subject headings are usually more than simple keywords or tags (see descriptor below). The subject field does not take a type attribute.

`note`

Use to record general notes about the item. The optional type attribute further qualifies the note, and will typically be used to generate a more informative label to indicate the nature of the note:

continued: a title or work that this serial is continued by.

continues: a title or work that is continued by this serial.

extent: describes the physical extent (size, pages, running time, etc.) of the item.

frequency: frequency of publication.

missing: notes about missing issues, pages, or other incomplete aspects of the work.

rights: a copyright notice or other information about rights and usage.

source: notes regarding the owner or location of the original source copy on which the digital copy is made.

`descriptor`

A descriptor is a keyword or short (no more than about 3 words) phrase describing the content of the item. It may be used for faceting purposes as well as display. Descriptors should be used consistently, preferably taken from a controlled list, whether a published standard or a local list. A descriptor can further be qualified by a type which indicates the sort of thing the descriptor is:

corporate: a corporate name. Use this value when specifying the name of a country, city, etc. as a primarily political entity.

date: a date or time period.

location: a geographic place, area, or coordinate.

person: a personal name.

`text`

Use this field to record text. If no type attribute is specified, the type "content" will be assumed. Allowed type values include:

content: the content of the item itself, such as text from OCR or a transcription of speech from an audio recording.

description: a summary, abstract, or other description of the item or its content.

Resources

A record must have a `canonicalUri`. All other resource fields are optional. Future versions may incorporate support for additional resources, such as alternate copies, formats or versions. Order is important: fields must appear in the following sequence: `canonicalUri`, `canonicalPreviewUri`, `canonicalMaster`, `canonicalDownload`.

`canonicalUri`

The URI of the standard online representation of the item, which could be a digital object, a web page that links to the object or a record for the object, or something else. It does not have to be unique, but best practice is for each record to point to (or near) a usable representation of the item it describes, rather than a more generic parent item. (E.g., the canonicalUri for a page record should provide access to that page, not to its parent book or other container as a whole.)

`canonicalPreviewUri`

The URI of the standard representative image that could be used as a thumbnail or preview image. It should point to a small web-friendly image. It also does not have to be unique. The Discovery Portal will assume that any supplied value is suitable for use as the src attribute of an HTML img element.

`canonicalMaster`

The name of a local digital master object, from which derivative images can be produced. the MIME type must be specified and a size (in bytes) and MD5 hex digest are optional.

The list of enumerated MIME types is currently incomplete and will be augmented as needed.

The meaning of this element is application-specific and primarily intended for internal use by Canadiana for other applications. The Discovery Portal does not make use of this value, and will ignore it if it is specified. It can be safely omitted.

`canonicalDownload`

The name of a local digital representation which may be downloaded directly without further derivation or processing. It takes the same attributes as the canonicalMaster.

As with canonicalMaster, this element is not intended for use with the Discovery Portal and can be safely omitted.

Future Versions

CMR is intended to capture a richer set of metadata than can be meaningfully used in the immediate future in anticipation of future usefulness. The utility of metadata depends not only on structure and consistency, but also in having a certain critical mass of compliant content in order to make features such as faceting and sorting not only technically possible but also useful.

CMR may be extended in the future to capture additional metadata. Some possible control fields include:

- Subject date coverage: a date range describing the time period addressed by an item.
- Geographic coverage: a place name selected from a standardized list indicating a geographic region addressed by or relevant to the item. A second field may include geographic coordinates, such as a latitude, longitude and radius.
- Rights: a standardized set of fields encoding the copyright and use status of an item, e.g. using Creative Commons indicators.