National Digital Collections Index

Discussion Paper: 17 April 2009 (DRAFT 1)

Overview

The purpose of the National Digital Collections Index (DCI) is to create an aggregate index of Canadian digital content and to provide a complete toolkit for memory institutions to load their index records. This is an evolution of the national index goal of the Alouette project and the Digital Collection Builder toolkit, scaling up function and capacity to support the growing demand by Canadian memory institutions to make their collections accessible.

The key features of the DCI will be:

- the use of documented and open standards for ingestion and export;
- records may be accepted from any collection which meets established criteria;
- records can be ingested, updated, and deleted;
- records can be normalized and accepted or rejected based on adherence to a set of standards and practices;
- the index should be able to link a record to multiple instances of the object it describes, and be able to validate the availability of these objects from time to time;
- the index is available to the public at large to harvest (in whole or in part) and to be re-purposed for a variety of purposes.

The DCI is not meant to directly extend or build upon the existing portal or index, nor is it meant to replace either one. It is meant to be a new project but one which, to the extent possible, leverages and builds upon existing tools, knowledge and content, whether from the Alouette project or from other work.

Discovery and Access

Although Canadiana.org may opt to create one or more discovery tools to provide access to the index, the intent is to leave most of the discovery tool work in the hands of third parties who will take all or part of the index and develop access tools which cater to the needs of a particular constituency. Canadiana.org may provide some or all of the tools to facilitate this, such as the Digital Collections Builder (DCB).

The DCI should be application agnostic, being able to supply its content to anyone via an open and standardized protocol, and not tied to or optimized for any one particular use case or software application.

Digital Collections Index: Key Components

Digital Collections Index

A database or repository containing the normalized, ingested metadata from all contributors.

Automated Harvester

A crawler which periodically queries and incorporates updates from contributors whose metadata is accessible via a standardized protocol. This is the preferred ingestion method for larger collections. The harvester may accept a variety of record formats, and will validate each incoming record against a set of minimum criteria before accepting it.

Submission/Update API

An API which accepts contributor-initiated update requests. It consists of a batch input API and may optionally include a Web front end for manual one-record-at-a-time updates. Otherwise similar to the automated harvester.

Automated Normalization

Automatically normalizes data to the extent possible, including standardizing formats (dates, etc.) and performing automated authority and subject mappings before transferring them to the index. Records can also be flagged for manual normalization based on certain criteria, including contributor identity as well as record content.

Manual Normalization

Records can be flagged and set aside as requiring or optionally benefiting from manual normalization or inspection before being placed in the DCI.

Link Validator

A service that periodically checks links to the digital objects provided by ingested metadata and attempts to flag any broken links or missing objects.

Export API

An API that allows for batch selection and export of records in one or more standardized formats. This API is publicly available to any client that implements the protocol. May optionally include a web interface for manual querying and retrieval of individual records and/or a web API for similar querying of individual records.

Portals and Discovery Layers

The DCI itself does not contain an integrated discovery layer and is application agnostic. Access to DCI content is via the export API, which can be used to download and mirror the index, either in whole or a custom-scoped view. This exported data can serve as the basis for any number of different access and discovery tools.

Canadiana.org may build a discovery portal of its own, and will also build and provide tools to assist third parties in harvesting from the DCI and in creating their own portals. Third parties may also elect to implement their own harvesting and discovery mechanisms independently of Canadiana.org.

Partnerships

The DCI will work with a combination of formal partners, informat partners, and anonymous users:

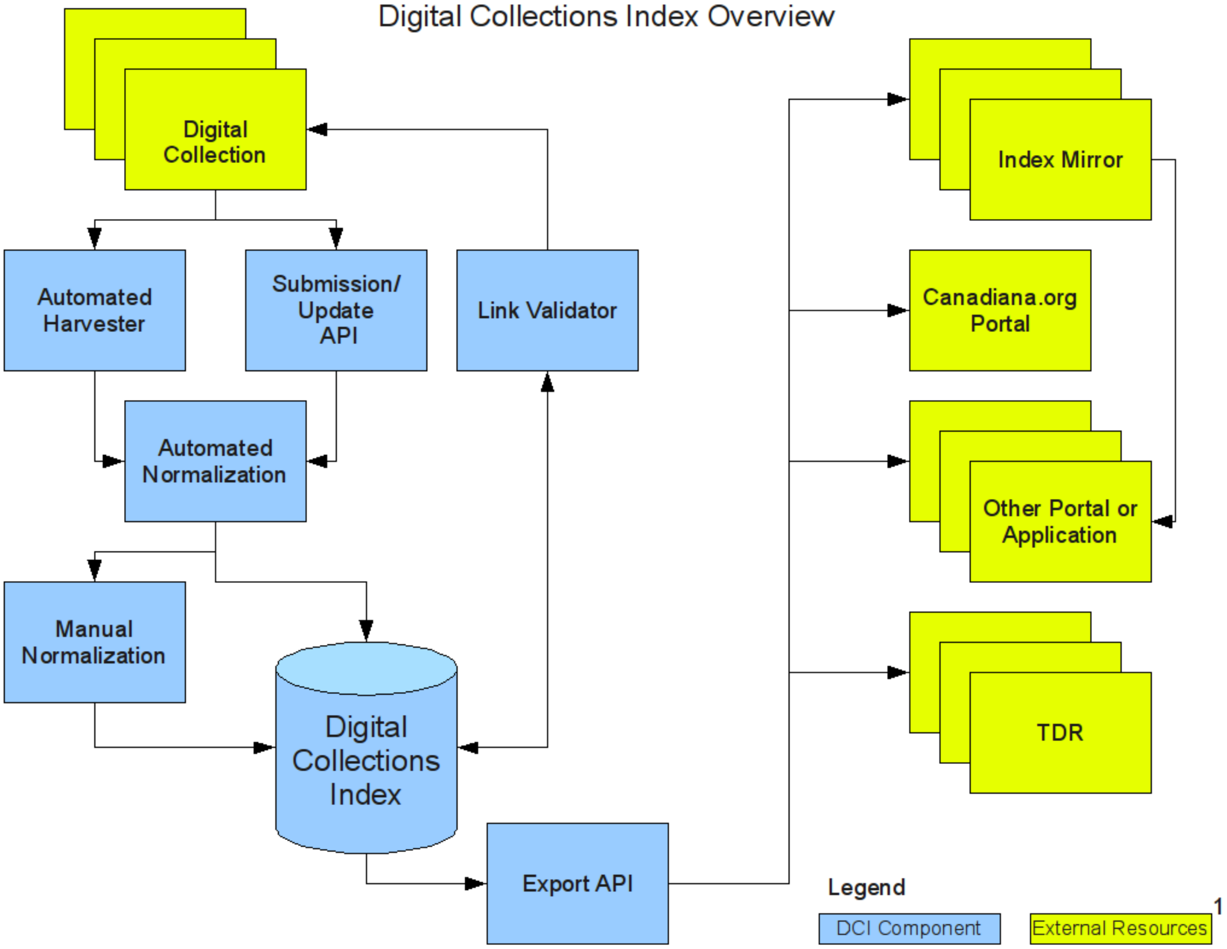
- contributors to the DCI must establish a relationship with Canadiana.org, be assigned an identity to uniquely tag their collections, and meet certain eligibility criteria;
- those who wish to contribute but lack the resources to meet these requirements on their own may still contribute through a trusted intermediary which does have an established relationship with Canadiana.org;
- Canadiana.org may establish partnerships for the purpose of mirroring the DCI, but the open and public nature of the API means third parties might also mirror the index for their own purposes;
- likewise, portal developers might work with Canadiana.org, but they may also act independently and on their own initiative, as the data from the index will be publicly available for anyone to repurpose;
- the establishment of a series of Trusted Digital Repositories (Tars) is part of the long-term objectives of Canadiana.org and its partners; these TDRs will store and safeguard the digital objects themselves on behalf of their owners, but could also act as mirrors or backups of the DCI.

Functional Requirements

Discussion of functional requirements should address the following issues:

- ingestion methods (including automated harvesting and contributor-pushed records)
- export/harvesting interface
- contributor authentication and identification
- unique ID generation
- index backup and replication
- revision and deletion of existing records
- digital object verification (e.g. via a link crawler)
- storage, server, and network requirements
- normalization of:
    - container format
    - vocabulary and schema mapping
    - best practices and semantic conventions
- automatic data normalization
- flagging, set-aside, and manual normalization provisions

Digital Collections Index Overview


Digital Collection
Automated Harvester
Submission/ Update API
Link Validator
Automated Normalization
Manual Normalization
Digital Collections Index
Export API
Index Mirror
Canadiana.org Portal
Other Portal or Application
TDR
Legend
DCI Component
External Resources
1