Final Report of the BC Digitization Portal Working Group

Overview

The BC Digitization Portal Working Group was formed by the BC Digitization Coalition in the spring of 2010. Its initial role was to evaluate alternatives to the existing software supporting the West Beyond the West search portal (westbeyondthewest.ca), and make a recommendation about which alternative is the best candidate to replace the existing software. The Working Group’s discussions revealed that the needs of the BC “memory organization” community were likely larger than just the search portal. This broader approach was confirmed in a meeting with the Coalition in October 2010.

This report represents the Working Group’s chief deliverable. It makes recommendations regarding the creation and implementation of three services to support the mission of the BC Digitization Coalition, along with additional recommendations for a permanent Technical Committee and for the formalization of the Coalition’s relationship with Canadiana.org.

Three Services

In addition to a search engine for BC digital collections, the Working Group has identified two closely related needs. First, some memory organizations have a need for assistance with hosting digital collections. These organizations, in particular small archives, libraries, and museums, lack the technical expertise necessary to set up tools such as the Digital Collection Builder on their own. Second, some organizations which already maintain their own digital collection software need assistance with extracting the metadata so that it can be easily loaded in the Coalition's search engine.

Therefore this report envisions three separate services to meet the three core needs identified. Recommendations specific to each service are discussed in the following three sections of the report.

To summarize, the services and the need they meet are as follows:

Service	**Need**
Search portal	Enable end users to search across many BC digital collections
Digital collection hosting service	Enable institutions to easily host collections online without setting up their own servers
Metadata transformation service (aka "metadata wrangling")	Make possible the easy exporting of their metadata into the search engine, for institutions which operate their own digital collection hosting software

I. Search portal

A. Overview of the Service

The Working Group considered a number of possible "architectures" for hosting the search portal and integrating it with other services such as the metadata wrangling function. In the end it was felt the best approach was to take advantage of the opportunities offered by the new Canadiana.org national index.

Recommendation 1: We recommend that that BC search portal be implemented as a scoped search of the Canadiana national index, using the Canadiana API; and that the Coalition identify funds per the budget outlined below and hire a developer to build the BC search portal.

The Canadiana index includes an API (Application Programming Interface), in essence a gateway which allows other websites to pass it a search query and receive back the results. The BC search portal will be a locally built and hosted website which uses the API to communicate with Canadiana and display search results to end users.

B. Implementation and Budget

Developing a simple search interface using a web application framework is a straightforward task. Working Group members discussed whether they felt this could be done as a volunteer task, an “off the side of the desk” job by a developer at one of the large institutions participating in the Coalition. It was felt that this was not the best approach because the job would be a low priority. Since it will not be expensive to build, and since launching a new West Beyond the West search engine is a high priority, it was felt the best approach was to identify funds and hire a developer for a short contract.

Budget for Search Portal Implementation

Software development [1]	$5,000 - $8,000
Web design, graphics [2]	$1,000
TOTAL	$6,000 - $9,000

Notes:
[1] The range of costs represents the difference between a basic implementation and a full implementation with all desired features. The functional spec would assign priorities to features, so that the developer completes the essential features first and then moves on to “would be nice” features as funds permit. The cost calculation is based on 1 month of time for a developer earning approximately $60k/year.
[2] West Beyond the West already has a very attractive logo. However, the design exists only as a small and low-quality web image. It may prove necessary to purchase or recreate the high-quality original, and/or carry out other graphic design work on the interface.

Project management could be carried out by the Coalition’s Digitization Coordinator position. A small group of 3-4 people including members of the current Working Group could meet with the Coordinator and software developer on a bi-weekly basis to assist in developing a functional specification and supervising the work of the developer. This group is further discussed in Recommendation 4, the Technical Sub-Committee.

Appendix A of this document lists desired functionality for the search portal, and would form the basis of a functional specification.

Hosting could be on a server at the Irving K. Barber Learning Centre, since it already hosts the Coalition website.

It is anticipated the search portal could be launched in approximately 3 months, including 1 month of developer time plus additional time for planning, testing, etc.

One area of concern for the Working Group is that the Canadiana API is currently a beta and may change in the future, necessitating updates to the software which runs the BC search portal. This could be mitigated to some extent through requiring the developer to create a flexible back-end to the system. This risk could also be mitigated through a more formal agreement with Canadiana; see section V for more details.

C. Additional Needs for the Search Portal

Once the portal is launched, it would benefit from a promotion plan to inform the wider BC community of users about the search engine's availability. This presumably is work the Coalition would undertake.

As well, in order to provide a "BC search" of Canadiana's metadata, it will be necessary to identify collections relevant to BC within the larger Canadiana holdings. Initially this would be a manual process carried out by communicating with the Canadiana technical staff; the Digitization Coordinator could lead the work.

II. Digital Collection Hosting Service

A. Overview of the Service

The Working Group considered several software options and possible hosting arrangements for a digital collection hosting service. In the end it was unanimously agreed that the best option would be to ask Artefactual Systems, a local company, to submit a proposal for hosting an instance of the Digital Collection Builder (DCB) software.

The DCB is a free, open source application which the BC Digitization Coalition is already supporting through offering training. It is a lightweight software with a sufficient feature set to support Coalition needs, and is in many ways superior to expensive commercial alternatives. A third-party commercial hosting arrangement, while having a direct financial cost, will likely be a more straightforward and quicker hosting arrangement than the alternatives (such as asking a major BC public institution to host it). Artefactual Systems is the lead developer for the archival version of the software, and therefore best positioned to add any required or desired features. As well, Artefactual already hosts the BC Archival Association's MemoryBC service on a similar arrangement.

To confirm that Artefactual and the DCB could meet the needs of the Coalition, the Working Group developed a list of functional requirements for the hosting service. These are detailed in Appendix B of this report. Three members of the Working Group then met with Peter van Garderen, principal of Artefactual, to interview him regarding the requirements and his proposal for hosting the DCB.

Recommendation 2: We recommend the Coalition identify funds per the budgets below and sign a contract with Artefactual to (a) support an instance of the Digital Collection Builder as the Coalition's digital collection hosting service and (b) add certain essential features required by the Coalition.

Hosting the DCB will have an annual cost. As well, the requirements process identified a number of essential features that are not yet present in the software and will have to be developed. The following sections spell out those costs.

B. Implementation and Budget, Part 1: Annual Hosting Fee

Peter supplied an estimate of $1,500 per year as the annual hosting cost. This includes backups and release upgrades. There would be an additional per-gigabyte data charge if the total data collection size exceeded 100 G.

Budget for Hosting Service - Annual Fee

Hosting Fee	$1,500 / year

The Working Group feels that some of this cost could be recovered by charging a fee to participating institutions. The fee should be low enough so as not to be a barrier to small institutions. It might start at $25 or $50/year and scale up with institution and collection size.

C. Implementation and Budget, Part 2: Missing Features

The DCB largely met the functional requirements list. However, a number of desired features were missing, and Artefactual was asked to supply quotes for adding these features. With the the per-feature cost known, the list was then divided into essential features the Coalition would require in order to launch the service, and "would be nice" features which could be added at a later date. The essential-but-missing features are not unexpected; they are related to the need to make the DCB more flexible in a multi-institutional setting, and to meet the needs of our partner Canadiana.

Budget for Hosting Service - Essential Features

CMR XML Export This is the ability to export metadata in the Canadiana CMR format, for ease of ingesting into the Canadiana search engine	$5,000
Institution-specific branding of repository page This allows institutions to brand and customize their "homepage" within the DCB.	$5,000 - $7,500 [1]

Ability to set data standard schema at the collection level This allows greater metadata flexibility for participating institutions by allowing them to select a different metadata schema (e.g. Dublin Core, MODS, etc.) for each separate collection.	$5,000
TOTAL	$15,000 - $17,500

Notes:
[1] Artefactual quoted a figure of $7,500. However we feel that a lower cost could be achieved by a tight scoping of the desired level of customization.

The “would be nice” features list and costs are included in Appendix C.

Again the Digitization Coordinator could lead the crafting of a contract with Artefactual and a detailed specification for the three features to be added, drawing on the support of the Technical Sub-Committee (see Recommendation 4) and members of the Coalition.

Note that by going with the DCB, an open source solution, the Coalition will automatically and freely benefit as other user groups pay for the addition of other features to the core software. As well, in the future the Coalition may be able to identify partners with whom to share the cost of developing new features where there is a mutual benefit.

Artefactual indicated their development pipeline is full for the next few months, but they could likely launch this service around September of 2011.

D. Additional Needs for Hosting Service

The digital collection hosting service will have a number of additional needs as it is rolled out to the BC community:

- promotion of the service
- a sys admin role creating and managing institutional accounts
- training for staff at participating organizations
- creating collections policies and metadata standards
- setting up procedures for automatic harvesting of metadata by Canadiana
- invoicing of participating organizations

Again it is assumed the Coalition’s Digitization Coordinator would take on much of this work.

III. Metadata Transformation Service

A. Overview of the Service

Large academic and public libraries who host digital collections usually have the ability to export directly to Canadiana.org, but smaller institutions (such as many of the recipients of the Irvin K. Barber BC History Digitization Program) do not. These institutions are unable to have their collections’ metadata harvested by Canadiana.org for technical and/or staff resource reasons. A

"metadata transformation" service that provided assistance in getting these institutions' metadata into the Canadiana.org index would greatly increase the amount of BC content aggregated in the search portal.

Recommendation 3: We recommend the Coalition develop a Metadata Transformation Service by identifying funds per the budget outlined below and hire a metadata transformation specialist.

In effect this transformation service is a single individual (the "metadata wrangler") hired on a part-time contract to work on institutional data migration as and when target collections are identified.

B. Metadata Transformation Specialist Job Description & Budget

The metadata transformation specialist will use technology and tools *to the extent possible* to convert each hosting institution's metadata into a format that can be imported directly into the Canadiana.org index. It is likely that a considerable amount of manual or semiautomated work with tools listed below will be involved in many cases. Due to the wide variety of platforms used to host content, no single solution to exporting metadata (such as OAI-PMH) exists. For example, some collections may be hosted in systems that use a relational database to store metadata, while others may simply embed their metadata in static HTML web pages; even within these two general types of platform, the specialist will need to solve a variety of unique problems. Therefore, this position will require the use of custom scripting in languages such as Perl or Python, the use of spreadsheets and other common data management tools, and the use of specialized tools such as XML editors to transform the incoming data into a standardized output format. The specialist will also document the processes involved with each collection's metadata so that ongoing work can be more easily performed in the future or in some cases redistributed to the hosting institutions, and, where possible, develop scripted approaches to common or general transformation tasks.

Possible qualifications for the metadata transformation specialist will be a professional librarian or archivist who has technical proficiency in the tools listed above and in XML, or a programmer who has experience manipulating library or archival metadata. Knowledge of a variety of metadata standards such as MARC, Dublin Core, and controlled vocabularies would be desirable.

It is likely a suitable candidate could be found in the Lower Mainland, possibly at SLAIS or a new graduate with technical skills.

Budget for Metadata Transformation Service

Metadata Transformation Specialist (6 month contract)	$9,000

Notes: The calculation is based on a salary of $30/hour x 15 collections x 20 hours/collection to migrate metadata. Those hours would be spread across the six month period as fits the individual's schedule.

Supervision of the wrangler could be carried out by the Digitization Coordinator. Support for the Coordinator to identify candidates, carry out the hiring process, and supervise the wrangler's work could come from the Technical Sub-Committee in Recommendation 4.

As suggested above, a sensible place for the metadata transformation specialist to start would be with the collections funded by the BC History Digitization Program, since this list of collections is well defined and some are thought to be "at risk". Renewing the contract could be based on evaluation of needs as new collections requiring migration are identified.

C. Additional Needs for the Transformation Service

In addition to the technical work of handling metadata, the following tasks should be done in support of the service:

- Promote the service to eligible libraries, archives, and museums, and actively contact institutions known to be carrying out digitization work
- Work with institutions to prepare for use of the service, e.g., pre-export updating of collections, quality checks, data normalization, etc.
- Advise institutions on improved workflows for metadata creation and maintenance, with the goal of making future exports easier to transform into the Canadiana.org format

IV. Technical Sub-Committee

As indicated above, the BC Digitization Coalition's Coordinator can carry out much of the project management and coordination necessary to implement the search portal, hosting service, and metadata wrangling. However, for all three of these services the Coordinator would benefit from being able to draw on the knowledge and skills of a group of experienced digitizers drawn from the sectors and institutions represented on the Coalition.

Recommendation 4: We recommend the Coalition create a permanent Technical Sub-Committee to assist and advise the Digitization Coordinator in implementing the other recommendations in this report.

The sub-committee could be chaired by the Coordinator. Some members of the current Working Group are willing to serve on the new Technical Sub-Committee.

V. Relationship with Canadiana

The recommendations in this report, if implemented, will mean that the BC Digitization Coalition will be relying heavily on the infrastructure of Canadiana.org to support the search engine function and serve as a central metadata storage site. Because of this reliance, it would be advisable that the Coalition formalize this relationship with Canadiana.

Recommendation 5: We recommend the Coalition consider whether to pursue some type of formal arrangement with Canadiana such as a letter of agreement to solidify the relationship, create mutually understood expectations, and ensure stability in areas such as technical functionality.

Some aspects of the relationship which might be covered include:

- a statement of API functionality and a process through which changes would be communicated in advance to the Coalition
- a process for informing Canadiana of collections to be included in the BC search
- regular harvest of metadata from the BC hosted service to Canadiana
- a process for coordination with the metadata migration service so that new collections can be regularly loaded
- clarification of any Canadiana policies around desired metadata formats, data quality, and collections policies that may impact uploading of Coalition-hosted or migrated metadata

Appendix A: Desired Functionality for the BC Search Portal

This section lists desired functionality for the BC search portal. It is planned that the Coordinator and Technical Sub-Committee would use this section to create a detailed functional spec for the software developer.

It is recognized that the BC search portal may be limited by the capabilities of the Canadiana API. In other words, the API may not enable some of the functionality outlined here. This is a list of desirable features rather than absolute requirements. The approach of a scoped search of Canadiana has been selected for reasons related to content; we'll 'make do' with the level of functionality available to us. Presumably the API will increase in sophistication as it develops.

Desired Functionality

1. A user should be able to do a 'basic' keyword search. The search will match on terms contained in the title, author, subject heading, and description fields, and possibly other fields as appropriate.

1.1 If multiple search terms are entered, the search will default to Boolean AND of each term.

2. A user should be able to do an 'advanced' search allowing exact phrase searching and Boolean AND and OR.

2.1 A user should be able to limit an advanced search by the following fields:
- title
- author/creator
- subject heading
- media type (e.g. video, image, text, audio, etc.)
- date of creation of the original document/object
- geographic location
- contributing institution or collection

3. A user should be able to choose the number of hits displayed per page within the range 10-100.

3.1 A user should be able to sort by a variety of fields such as title, creator, and contributing institution.

4. A user should be able to browse by the following fields:
- author/creator

- subject heading
- media type (e.g. video, image, text, audio, etc.)
- contributing institution or collection

5. A user should be able to enter search terms including non-English characters such as accents.

6. A user should be able to comment on objects found in the portal. The comments field should be protected by a Captcha, and provide a means for Coalition staff to approve comments before they are visible to the public.

7. A user should be able to create a personal account on the portal, protected by a password s/he chooses. When logged in, the user will see an "add to favorites" button next to digital objects, and can create a folder of "favorite" objects for easy retrieval on a future visit.

8. A user should be able to share a link to a digital object through a "share" button which connects to at least the following: Facebook, Twitter, delicious, and Digg, or else forwards the link as an email.

Appendix B: Functional Requirements for the Digital Collection Hosting Service

The following requirements were used to interview Peter van Garderen of Artefactual to determine whether the functionality of the Digital Collection Builder (DCB) software would meet the Coalition's needs.

For the BC Digitization Coalition's hosting service, the Working Group envisions a system that will enable a memory institution to request and obtain an account for an institutional space. Within that space they will be able to create one or more digital collections, including uploading digital objects and attaching metadata.

The primary user roles within this system will be:

- Hosting Service Admin: a staff member of the BC Digitization Coalition who creates institutional accounts and manages metadata export to Canadiana
- Institutional Admin: one (or more?) staff members at each participating institution who manages Collection Creator accounts and controls any institution-level settings
- Institutional Collection Creator: one or more staff members at each participating institution who uploads digital objects and attaches metadata

The system will also have end-user search and browse functionality.

Coalition Staff Requirements

1. The Hosting Service Admin can create and manage an institutional space (and one or more associated Institutional Admin accounts) to house a hosted collection.

2. The Hosting Service Admin can expose the entire collection of metadata for harvesting by other parties in a variety of formats including OAI-PMH, MARC, and custom formats.

3. The Hosting Service Admin can expose the entire collection of metadata for crawling by search engines such as Google.

4. Please identify any additional service-level functionality you feel is relevant.

Institutional Requirements

5. The Institutional Admin can manage institutional-level settings including customizing the appearance and branding of their institution's space.

6. The Institutional Admin can create and manage one or more digital collection spaces.

7. The Institutional Admin can create and manage multiple Collection Creator accounts.

8. The Institutional Admin can manage metadata creation workflow. For example, one Collection Creator is able to create metadata and a second person approves it.

9. The Institutional Admin can choose a metadata template corresponding to common metadata schemas such as Dublin Core (unqualified and DC Terms), MODS, and EAD. The metadata schema can be selected at the collection level; in other words an institution with multiple collections could have a different metadata template for each collection.

9.1 The Institutional Admin can create a unique metadata schema for a collection.

10. A Collection Creator can create and manage a digital collection. This includes: uploading full-size digital files; uploading additional images such as thumbnails if necessary; and assigning metadata.

10.1 The Collection Creator can upload a variety of formats of digital objects within a collection including simple (e.g. photos) and complex (e.g. multi-page text, multi-part video).

11. Please identify any additional institution-level functionality you feel is relevant.

End-User Requirements

12. An end-user can do a 'basic' keyword search. The search will match on terms contained in the title, author, subject heading, and description fields, and possibly other fields as appropriate. The end-user can search across all collections, or limit the search to a single collection.

12.1 If multiple search terms are entered, the search will default to Boolean AND of each term.

13. An end-user can do an 'advanced' search allowing exact phrase searching and Boolean AND and OR.

13.1 An end-user can limit an advanced search by the following fields:
- title
- author/creator
- subject heading
- media type (e.g. video, image, text, audio, etc.)
- date of creation of the original document/object
- geographic location
- contributing institution or collection

14. An end-user can choose the number of hits displayed per page within the range 10-100.

14.1 An end-user can sort search results by a variety of fields such as title, creator, and contributing institution.

15. An end-user can browse institutions hosted by the service and collections within each institution. Within each collect an end-user can browse by:
- author/creator
- subject heading
- media type (e.g. video, image, text, audio, etc.)

16. Please identify any additional end-user functionality you feel is relevant.

Scale Requirements

17. The software should be able to provide all the functions listed above while storing up to 50,000 digital objects and their metadata.

Appendix C: Digital Collection Hosting Service - "Would Be Nice" Features

This table outlines functionality missing from the DCB which the Working Group identifies as "would be nice" for a future round of development.

Expose CMR XML via OAI harvesting This feature would allow for automation of the Canadiana harvesting. Note that there is a case for asking Canadiana to be a development partner in creating this feature, since they would	$7,500

benefit from staff time savings on their end.	
DC Terms Schema Support This would allow a mid-way metadata option between two of the current options, unqualified Dublin Core (perhaps too “dumbed down” for some needs) and MODS (perhaps too rich and complex).	$7,500
Browsing on subject/place/name/media type within a collection This feature would improve the browse options for end users.	$2,500
Site usage reports available via sys admin UI This feature would simplify DCB administration by providing reports on topics such as collection size.	$7,500

Appendix D: Summary of Recommendations and Budget Tables

Recommendation 1: We recommend that that BC search portal be implemented as a scoped search of the Canadiana national search portal, using the Canadiana API; and that the Coalition identify funds per the budget outlined below and hire a developer to build the search portal.

Budget for Search Portal Implementation

Software development Web design, graphics	$5,000 - $8,000 $1,000
TOTAL	$6,000 - $9,000

Recommendation 2: We recommend the Coalition identify funds per the budgets below and sign a contract with Artefactual to (a) support an instance of the Digital Collection Builder as the Coalition’s digital collection hosting service and (b) add certain essential features required by the Coalition.

Budget for Hosting Service - Annual Fee

Hosting Fee	$1,500 / year

Budget for Hosting Service - Essential Features

CMR XML Export This is the ability to export metadata in the Canadiana CMR format, for ease of ingesting into the Canadiana search engine	$5,000
Institution-specific branding of repository page This allows institutions to brand and customize their “homepage” within the DCB.	$5,000 - $7,500
Ability to set data standard schema at the collection level This allows greater metadata flexibility for participating institutions by allowing them to select a different metadata schema (e.g. Dublin Core, MODS, etc.) for each separate collection.	$5,000
TOTAL	$15,000 - $17,500

Recommendation 3: We recommend the Coalition develop a Metadata Transformation Service by identifying funds per the budget outlined below and hire a metadata transformation specialist.
Budget for Metadata Transformation Service

Metadata Transformation Specialist (6 month contract)	$9,000

Recommendation 4: We recommend the Coalition create a permanent Technical Sub-Committee to assist and advise the Digitization Coordinator in implementing the other recommendations in this report.

Recommendation 5: We recommend the Coalition consider whether to pursue some type of formal arrangement with Canadiana such as a letter of agreement to solidify the relationship, create mutually understood expectations, and ensure stability in areas such as technical functionality.

Appendix E: Working Group Members

- Gordon Coleman, BC Electronic Library Network (Chair)
- Bronwen Sprout, University of BC
- Cecily Walker, Vancouver Public Library
- Chris Mathieson, BC Museum Association
- John Durno, University of Victoria
- Lara Wilson, Archives Association of BC
- Mark Jordan, Simon Fraser University