CANADIANA

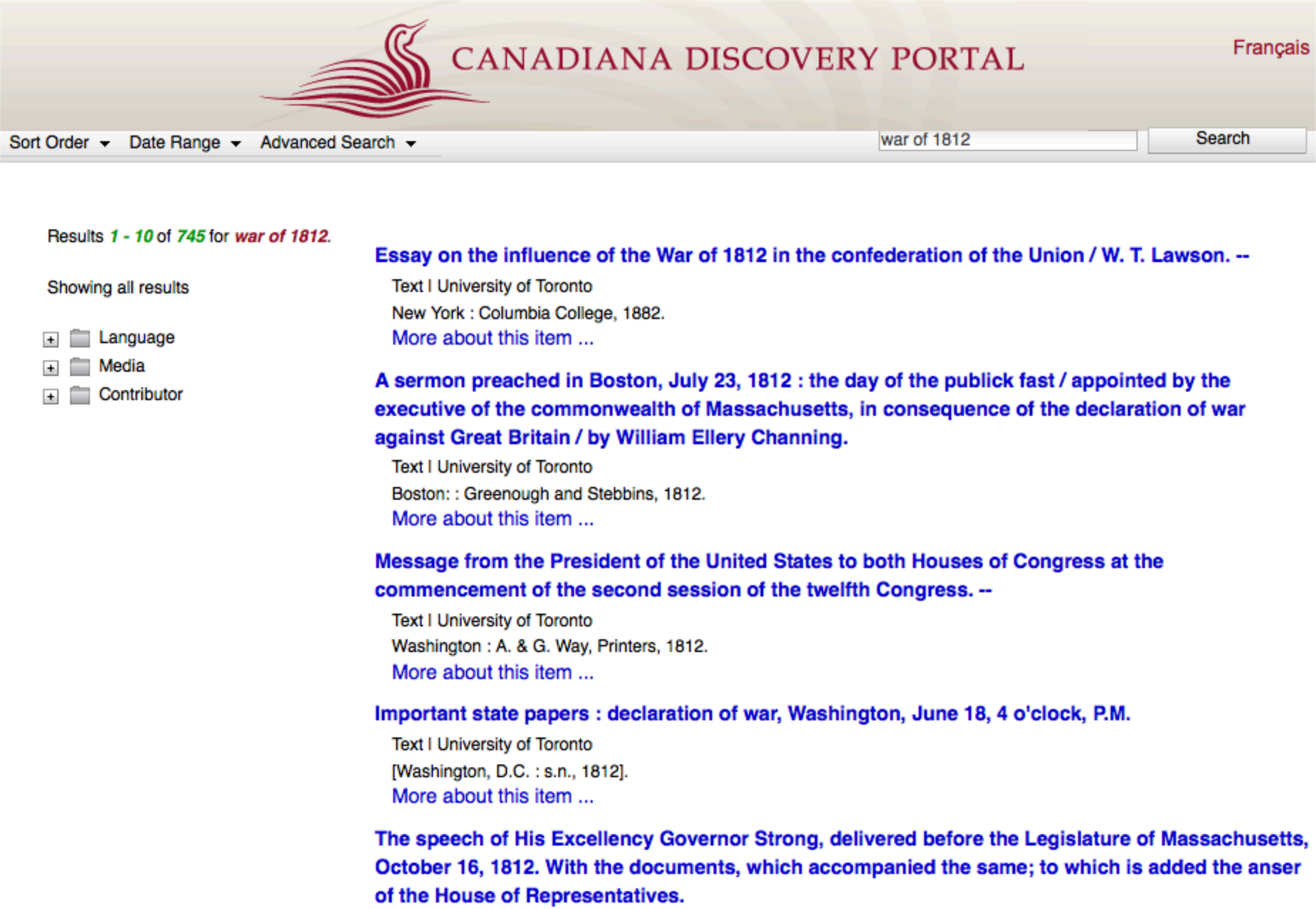
Discovery Portal Metadata and API

CANADIANA

The Canadiana Discovery Portal

- Goal: make Canada’s digital collections searchable in one place
- Contributed metadata
- Online since spring 2010
- Successor to Alouette portal

CANADIANA



Results *1 - 10* of *745* for *war of 1812*.

Showing all results

Essay on the influence of the War of 1812 in the confederation of the Union / W. T. Lawson. --

Text | University of Toronto

New York : Columbia College, 1882.

More about this item ...

A sermon preached in Boston, July 23, 1812 : the day of the publick fast / appointed by the executive of the commonwealth of Massachusetts, in consequence of the declaration of war against Great Britain / by William Ellery Channing.

Text | University of Toronto

Boston: : Greenough and Stebbins, 1812.

More about this item ...

Message from the President of the United States to both Houses of Congress at the commencement of the second session of the twelfth Congress. --

Text | University of Toronto

Washington : A. & G. Way, Printers, 1812.

More about this item ...

Important state papers : declaration of war, Washington, June 18, 4 o'clock, P.M.

Text | University of Toronto

[Washington, D.C. : s.n., 1812].

More about this item ...

The speech of His Excellency Governor Strong, delivered before the Legislature of Massachusetts, October 16, 1812. With the documents, which accompanied the same; to which is added the anser of the House of Representatives.

Stuff we used

- Perl/Catalyst
- Template Toolkit
- MySQL
- Solr
- JavaScript/Dojo
- XML
- JSON
- Apache
- FastCGI
- Linux

Search capabilities

- Keywords, phrases, wildcards, Boolean
- Field-limited searching (subject, title, etc.)
- Sub-searching issues and pages
- Facets
- Date range searching
- Date and relevance sorting

CANADIANA

Full command-line for advanced users

```
queenston su:"war of 1812" | ti:"war of 1812"
-media:text
```

Contents

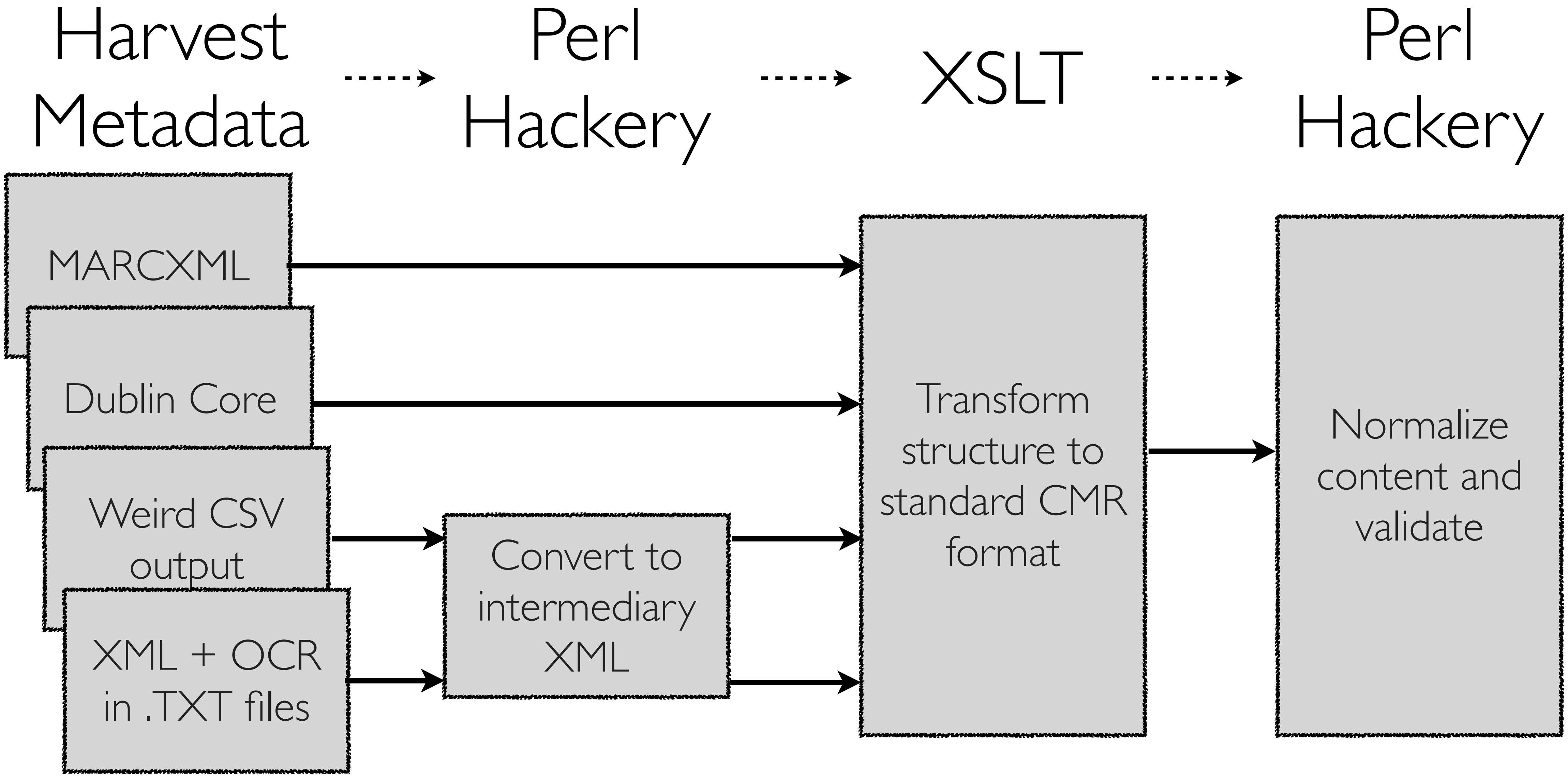
- 1 million items, 3+ million pages of indexed full text, 300,000 images, some A/V
- Page, document, series level indexing
- Any museum, library, archive can contribute
- Simplified subset of metadata: a finding aid, not union catalogue

CANADIANA

Discovery Portal workflow

- Harvest metadata from contributors
- Convert to Canadiana Metadata Repository (CMR) format
- Index into Solr
- Public access via Web site or API

Ingestion and conversion



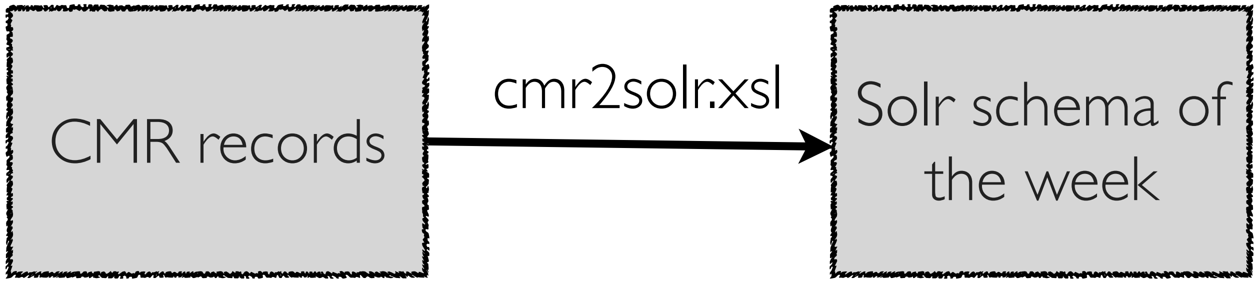
CANADIANA

From CMR to Solr index

- CMR is a (relatively) stable intermediary format

- Solr schema updated regularly to accommodate application changes

- Simple stylesheet converts CMR to current Solr schema



CMR functional requirements

- Map metadata from diverse sources to a common set of fields
- Normalize sortable/facetable data
- Simple to convert to Solr schema
- Manage and link parent-child relationships (series, document, page)
- Extensible

CANADIANA

CMR record structure

```
<?xml version="1.0" encoding="UTF-8"?>
<recordset version="1.1">
  <record>
    <!-- control & facet fields ... -->
    <description>
      <!-- bibliographic & full text fields ... -->
    </description>
    <resource>
      <!-- URLs, filenames, resource pointers ... -->
    </resource>
  </record>
</recordset>
```

Control and facet fields

```
<type>document</type>
<contributor>oocihm</contributor>
<key>8_04218_10</key>
<pkey>8_04218</pkey>
<label>[Vol. 1, no. 10 (May 1869)]</label>
<seq>10</seq>
<lang>eng</lang>
<media>text</media>
<pubdate min=”1869-01-01T00:00:00.000Z”
max=”1869-12-31T23:59:59.000Z”/>
```

Metadata normalization

- Purpose: enable sorting, faceting, subsetting
- Focus on low-effort, high-utility information
- 3 steps: identify, decode, normalize

Identifying date of publication

Obvious

```
<datafield tag="260" ind1="0" ind2=" ">
  <subfield code="a">Vancouver,</subfield>
  <subfield code="c">1907-12.</subfield>
</datafield>
```

260$c: "Date of publication, distribution, etc."

Identifying date of publication

Not quite as obvious

```
<dc:date>2010-04-13T15:28:13Z</dc:date>
<dc:date>2010-04-13T15:28:13Z</dc:date>
<dc:date>1911</dc:date>
<dc:date>1916</dc:date>
```

dc:date: "A point or period of time associated with an event in the lifecycle of the resource."

Normalizing date of publication

- [18---1910]
- August 1986
- 02/07/1983
- 2-6-98
- l9ll
- Nov-77
- 1990-01
- '84
- 1920s
- c. 1920’s
- ca. 192
- Early 20th Century

Identifying & normalizing language

- en, fr
- eng, fre
- engfre
- English;French

CANADIANA

Identifying media types: text

- announcement
- article
- atlas
- binding
- book(let)?
- broadside
- correspondence
- document
- finding aid
- journal
- letter
- magazine
- manuscript
- minutes
- news(letter|paper|print)s?
- notebooks?
- pamphlet
- pdf
- periodical
- poems?
- report
- rtf
- scrapbooks?
- short story
- texts?
- textual \w+
- thes[ie]s
- typescript
- yearbooks?

Normalization strategy

- Heuristics based on previously-encountered material
- Not 100% accurate
- Most fields are optional
- Erring on the side of exclusion increases relevance; encourages better metadata

Indexing CMR records: cmr2solr.xsl

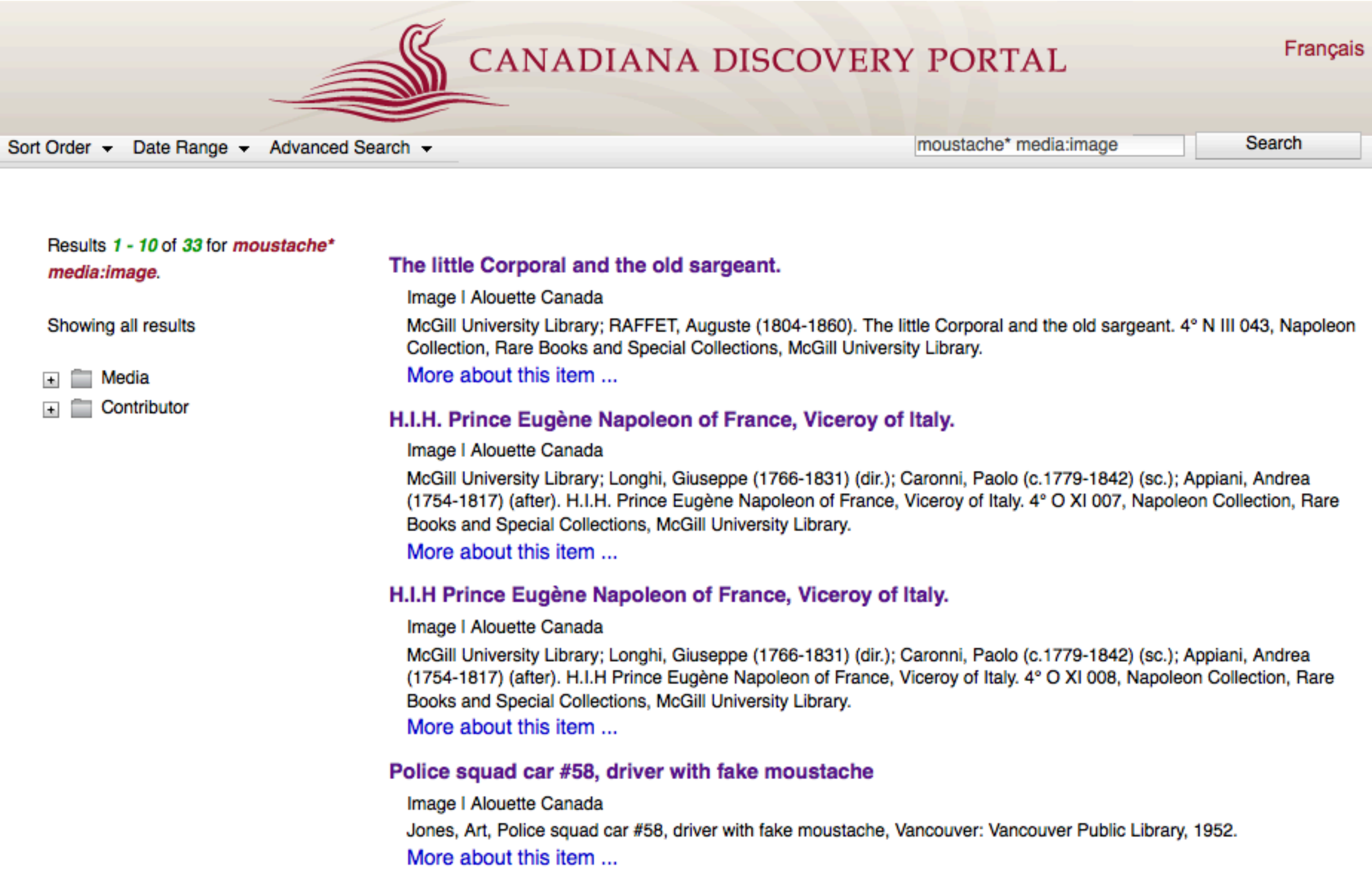
- Convert CMR to Solr documents with XSLT
- Structural manipulation only; no content hacking
- Separating storage and index/access formats makes changing either one much easier

CANADIANA

Web service API

- Discovery portal supports a simple web service API
- Extension of the interactive query syntax
- Original intent was for use as an AJAX interface for internal use

/search?q=moustache* media:image



Results *1 - 10* of *33* for ***moustache* media:image***.

Showing all results

The little Corporal and the old sargeant.

Image | Alouette Canada

McGill University Library; RAFFET, Auguste (1804-1860). The little Corporal and the old sargeant. 4° N III 043, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

H.I.H. Prince Eugène Napoleon of France, Viceroy of Italy.

Image | Alouette Canada

McGill University Library; Longhi, Giuseppe (1766-1831) (dir.); Caronni, Paolo (c.1779-1842) (sc.); Appiani, Andrea (1754-1817) (after). H.I.H. Prince Eugène Napoleon of France, Viceroy of Italy. 4° O XI 007, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

H.I.H Prince Eugène Napoleon of France, Viceroy of Italy.

Image | Alouette Canada

McGill University Library; Longhi, Giuseppe (1766-1831) (dir.); Caronni, Paolo (c.1779-1842) (sc.); Appiani, Andrea (1754-1817) (after). H.I.H Prince Eugène Napoleon of France, Viceroy of Italy. 4° O XI 008, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

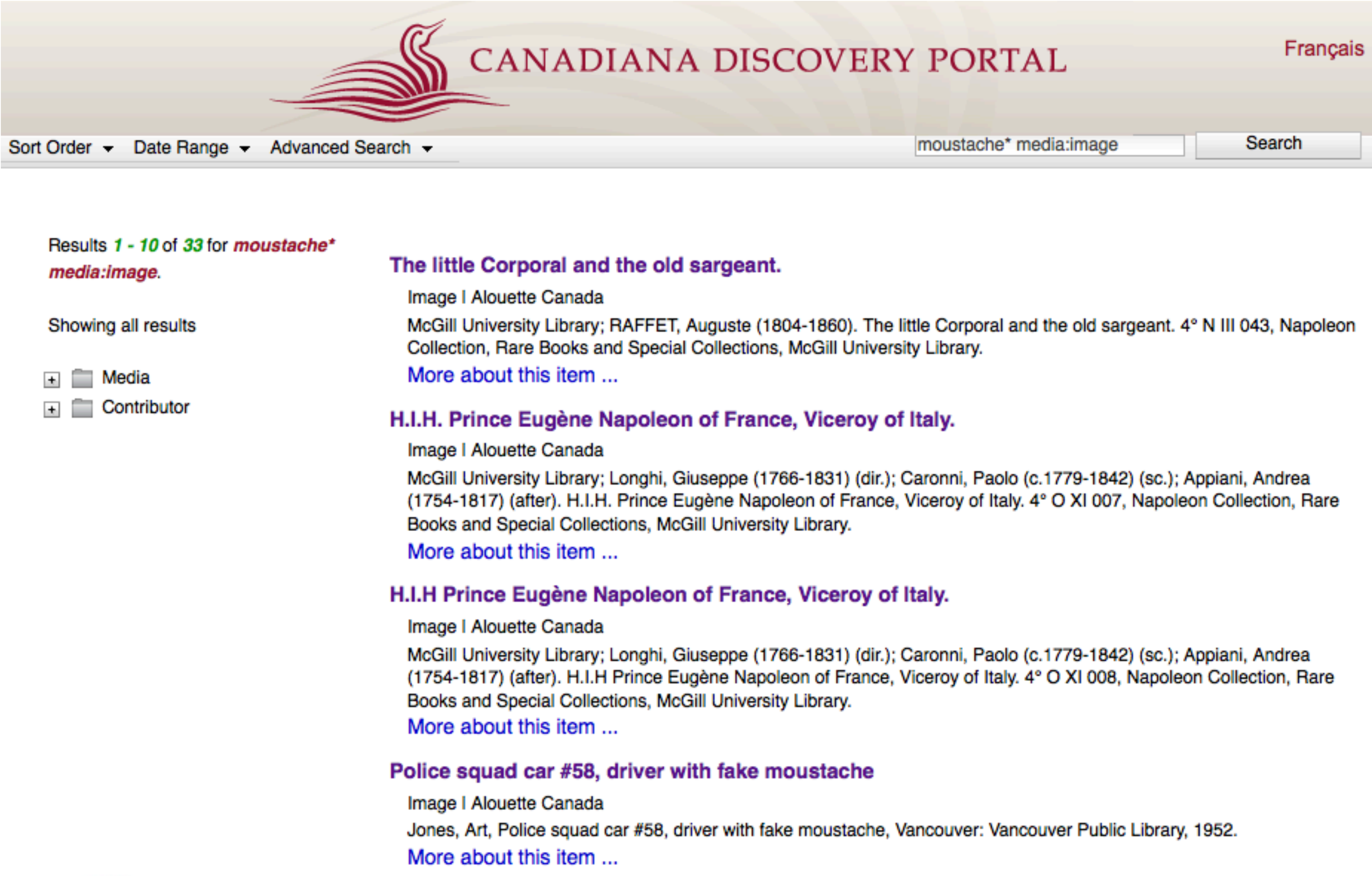
Police squad car #58, driver with fake moustache

Image | Alouette Canada

Jones, Art, Police squad car #58, driver with fake moustache, Vancouver: Vancouver Public Library, 1952.

More about this item ...

/search?q=moustache* media:image&fmt=xml



Results *1 - 10* of *33* for ***moustache* media:image***.

Showing all results

The little Corporal and the old sargeant.

Image | Alouette Canada

McGill University Library; RAFFET, Auguste (1804-1860). The little Corporal and the old sargeant. 4° N III 043, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

H.I.H. Prince Eugène Napoleon of France, Viceroy of Italy.

Image | Alouette Canada

McGill University Library; Longhi, Giuseppe (1766-1831) (dir.); Caronni, Paolo (c.1779-1842) (sc.); Appiani, Andrea (1754-1817) (after). H.I.H. Prince Eugène Napoleon of France, Viceroy of Italy. 4° O XI 007, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

H.I.H Prince Eugène Napoleon of France, Viceroy of Italy.

Image | Alouette Canada

McGill University Library; Longhi, Giuseppe (1766-1831) (dir.); Caronni, Paolo (c.1779-1842) (sc.); Appiani, Andrea (1754-1817) (after). H.I.H Prince Eugène Napoleon of France, Viceroy of Italy. 4° O XI 008, Napoleon Collection, Rare Books and Special Collections, McGill University Library.

More about this item ...

Police squad car #58, driver with fake moustache

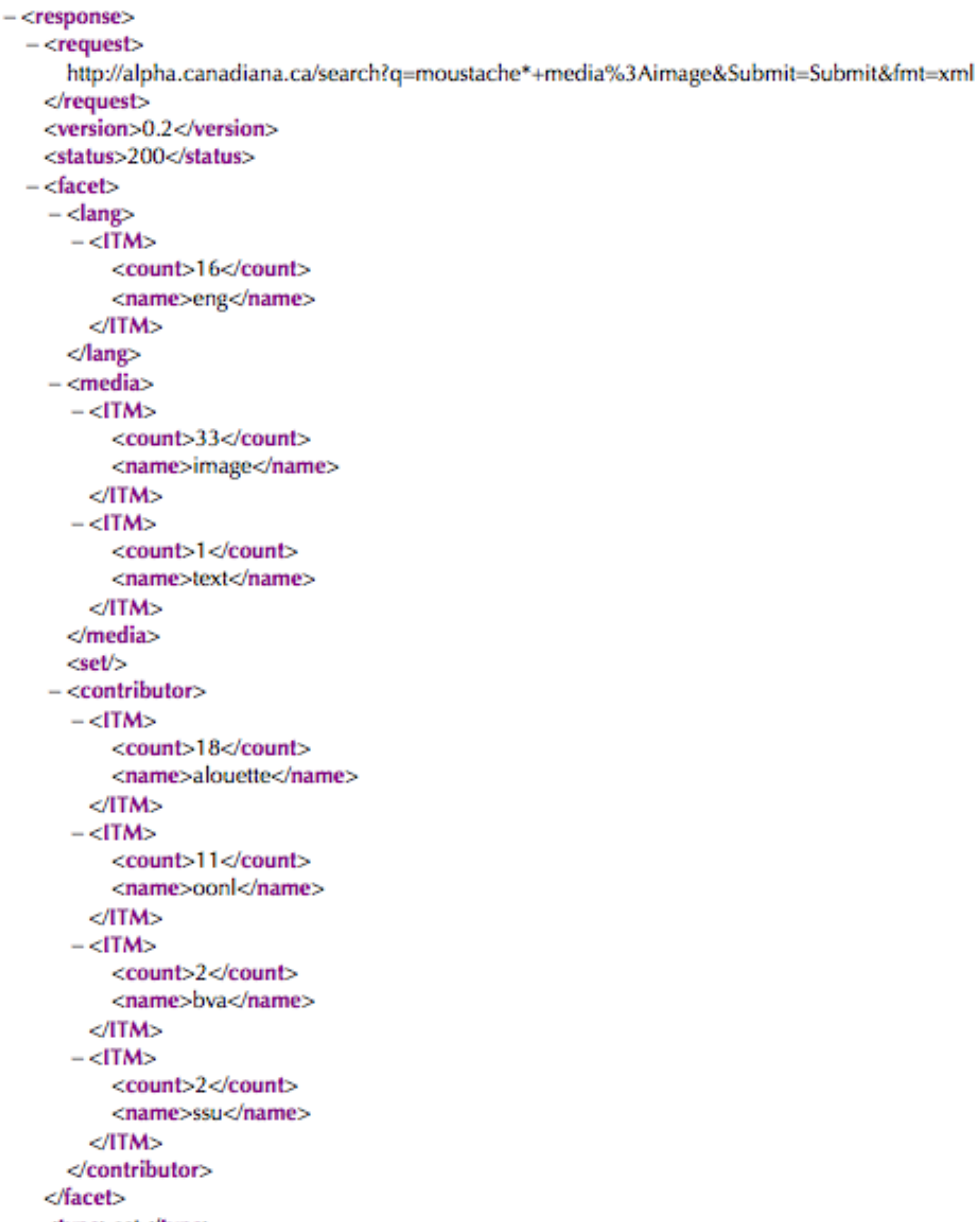
Image | Alouette Canada

Jones, Art, Police squad car #58, driver with fake moustache, Vancouver: Vancouver Public Library, 1952.

More about this item ...

CANADIANA

/search?q=moustache* media:image&fmt=xml



```
- <response>
  - <request>
      http://alpha.canadiana.ca/search?q=moustache*+media%3Aimage&Submit=Submit&fmt=xml
    </request>
    <version>0.2</version>
    <status>200</status>
  - <facet>
    - <lang>
      - <ITM>
          <count>16</count>
          <name>eng</name>
        </ITM>
      </lang>
    - <media>
      - <ITM>
          <count>33</count>
          <name>image</name>
        </ITM>
      - <ITM>
          <count>1</count>
          <name>text</name>
        </ITM>
      </media>
      <set/>
    - <contributor>
      - <ITM>
          <count>18</count>
          <name>alouette</name>
        </ITM>
      - <ITM>
          <count>11</count>
          <name>oonl</name>
        </ITM>
      - <ITM>
          <count>2</count>
          <name>bva</name>
        </ITM>
      - <ITM>
          <count>2</count>
          <name>ssu</name>
        </ITM>
      </contributor>
    </facet>
```

API calls

- Add `fmt=xml` or `fmt=json` query parameter to turn any query into a Web service call
- Identical information; JSON more convenient for machine parsing, XML more human-friendly

Query syntax

```
/search?ti=moustache&media=image&fmt=json
```

is equivalent to

```
/search?q=ti:moustache media:image&fmt=json
```

Response message

- Request status
- Paging information
- Facets
- Result set (documents)
  - Matching pages

CANADIANA

Paging information

```
<pubmin_year>1813</pubmin_year>
<pubmin>1813-01-01T00:00:00Z</pubmin>
<page>1</page>
<pubmax>1973-12-31T23:59:59.999Z</pubmax>
<next_page>2</next_page>
<hits>159</hits>
<hits_from>1</hits_from>
<hits_per_page>10</hits_per_page>
<hits_to>10</hits_to>
<prev_page/>
<pubmax_year>1973</pubmax_year>
<pages>16</pages>
```

Other query types

- Searches can also be run at the page level (`t=page`) or at the series level (`t=series`)
- Individual records can be retrieved:
  - `/view/RECORD_ID?fmt=xml`

CANADIANA

Index subsets

- Tagged sets identified by some criteria can be used to create searchable subsets
- Can be used to create a subset for custom query or application
- Example: **`set=bc`** limits searches to content contributed by British Columbia institutions

See also

- http://www.canadiana.ca
- http://search.canadiana.ca
- http://search.canadiana.ca/support/search
- http://search.canadiana.ca/support/api
- William@Canadiana.ca