James Robertson
Manager, DigitalNZ Systems
National Library of New Zealand

www.DigitalNZ.org
www.NatLib.govt.nz
DigitalNZ

Launched 2008
Find, Share, and Use NZ digital content
Website & API
200 Content Partners
30 million photos, audio, video, newspapers & more
10 million queries a month

Powers National Library of NZ search ...
He Tohu is a permanent exhibition of three iconic constitutional documents that shape Aotearoa New Zealand.

Visit He Tohu at the National Library

Use the Library

Researchers
Start your research at the Alexander Turnbull Library

Librarians
Use Te Puna and National Library services

Publishers & Authors
Get an ISBN, a record, and advice about Legal Deposit

Schools
Support for school libraries and learning

Closed Monday
October 23

The Library will be closed for Labour Day on Monday, October 23.

We open again on Tuesday. Have a good weekend!

Collection availability

Due to the Kaikoura earthquakes, some of our collections were not available for research.

On now at the Turnbull: Humble
The life of 100 everyday objects
Other API Consumers
WELCOME TO THE NEW DIGITALNZ

We're thrilled to have you here, and hope you'll enjoy the fresh look and added functionality. Want to know more? Read in detail about our new site. And, we'd love your feedback – hit the "feedback" button at the bottom each page to contact us.

STORIES NEW

Collect the items you're interested in and build a story around them. Here are a few of our recent favourites.
Supplejack

View at:
NatureWatch NZ

From:
NatureWatch NZ

Date:
2016/10/18 11:17 AM NZDT

Description:
Impressive encircling tendrils of a supplejack vine from the base of a rowanow to about 5m up the trunk where the leaves were sprouting.

Usage:
Share, Modify

Category:
Images

Subjects:
Plantae
Supplejack

Metadata Aggregation, Discovery and Management

Open-Source (Ruby on Rails, MongoDB, Solr, etc.)

Winner ‘Government’ 2016 NZOSS Awards

Runs the DigitalNZ API

- github.com/DigitalNZ (download & support)
- digitalnz.github.io/Supplejack (documentation)
- groups.google.com/forum/#!forum/supplejack (discussion)
Find, Share and Use

Aggregation
- Customizable Metadata Schema - permissive or strict
- Scripted Metadata ‘Harvesting’ using Ruby and Sj DSL
- Sources: XML, OAI, RSS, HTML, Sitemap, Structured Text, etc.

Discovery
- API Endpoints: Search, Records, Sets/Stories
- API Formats: JSON, XML, RSS
... and Manage the process

Parser Script Development
○ Version control
○ Shared code

Harvester Process Management
○ Job scheduling
○ Active, Finished, Failed, Stopped jobs
○ Multiple Environments (. i.e Production, Staging, UAT)

Manage Individual Records/Collections
Link Checking
Supplejack ‘Manager’
Supplejack Dashboard

Get started by identifying data sources to bring in to Supplejack. Then extract data from the sources by creating and running parser scripts. Complete the process by scheduling and monitoring regular harvest jobs.

Staging Server Statistics

<table>
<thead>
<tr>
<th>Harvest Jobs</th>
<th>0 currently active</th>
<th>1 finished</th>
<th>0 failed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last 24 hours</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Link Checking         | n/a reactivated    | n/a suppressed | n/a deleted |
| Last 24 hours         |                    |              |           |
### Recently Edited Parsers

<table>
<thead>
<tr>
<th>Parser Name</th>
<th>Data Format</th>
<th>Contributor</th>
<th>Last Edited</th>
<th>Last Editor</th>
<th>Parser Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLNZcat Alma OAI XML</td>
<td>xml</td>
<td>National Library of New Zealand</td>
<td>about 24 hours ago</td>
<td>Lisa Meehan</td>
<td>record</td>
</tr>
<tr>
<td>Bfm Rss</td>
<td>rss</td>
<td>95bFM</td>
<td>about 24 hours ago</td>
<td>Ting</td>
<td>record</td>
</tr>
<tr>
<td>Test Parser</td>
<td>xml</td>
<td>test</td>
<td>about 24 hours ago</td>
<td>Dan</td>
<td>record</td>
</tr>
<tr>
<td>Te papa api</td>
<td>json</td>
<td>Museum of New Zealand Te Papa Tongarewa</td>
<td>7 days ago</td>
<td>Ting</td>
<td>record</td>
</tr>
<tr>
<td>EMu</td>
<td>xml</td>
<td>Alexander Turnbull Library</td>
<td>9 days ago</td>
<td>Dan</td>
<td>record</td>
</tr>
</tbody>
</table>

### Next Scheduled Jobs

<table>
<thead>
<tr>
<th>Parser Script</th>
<th>Next Job</th>
<th>Frequency</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tvnz Headlines Xml</td>
<td>October 20, 2017 11:45:00</td>
<td>daily at 11:45 (45 11 ** *)</td>
<td>Full And Flush</td>
</tr>
<tr>
<td>RadioNZ News</td>
<td>October 20, 2017 13:50:00</td>
<td>daily at 13:50 (50 13 ** *)</td>
<td>Normal</td>
</tr>
</tbody>
</table>
Schema (API/MongoDB)
Schema

Adaptable
- Ruby Domain Specific Language (DSL)
- mix metadata standards though namespaces
- field-types: string, integer, datetime, boolean

API/Search Options
- Field Aliasing
- Field Grouping
Example Record Schema

class RecordSchema

    include SupplejackApi::SupplejackSchema

    # Namespaces
    namespace :sj, url: ''

    # Fields
    string :record_id, store: false,
    namespace: :sj
    string :title, search_boost: 10, search_as: [:filter, :fulltext],
    namespace: :dc
    string :description, search_boost: 2, search_as: [:filter, :fulltext],
    namespace: :dc
    string :display_content_partner,
Harvesting (Parser Scripting)
Harvesting Process

Websites
- HTML

Data Service
- XML, OAI-PMH, RSS, JSON...

Local Files
- Plain text, CSV, XML, JSON

Supplejack Harvester

Supplejack Database & Search Indexes
- MongoDB + Solr
Partners & Collections (Sources)

Archives NZ

Data Source
Archives NZ XML API

Parser Script

Archives NZ XML API

Data Source
Archives NZ YouTube

Parser Script

Archives NZ YouTube
Shared Code (Snippets)

Archives NZ

- Data Source: Archives NZ XML API
  - Parser Script
    - Archives NZ XML API

Tourism NZ

- Data Source: Tourism NZ YouTube
  - Parser Script
    - Tourism NZ YouTube

Snippet

YouTube Shared Parser Rules
<table>
<thead>
<tr>
<th>Harvest Name</th>
<th>Data Format</th>
<th>Partner</th>
<th>Source ID</th>
<th>Last Updated</th>
<th>Last Edited By</th>
<th>Parser Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCL Recollect</td>
<td>xml</td>
<td>Wellington City Libraries</td>
<td>WCL Recollect</td>
<td>2017-10-20 14:02:31 +1300</td>
<td>Ting</td>
<td>record</td>
</tr>
<tr>
<td>Bfm Rss</td>
<td>rss</td>
<td>95bFM</td>
<td>Bfm-rss</td>
<td>2017-10-19 11:28:36 +1300</td>
<td>Ting</td>
<td>record</td>
</tr>
<tr>
<td>Test Parser</td>
<td>xml</td>
<td>test</td>
<td>Test</td>
<td>2017-10-19 11:26:32 +1300</td>
<td>Dan</td>
<td>record</td>
</tr>
</tbody>
</table>
class BfmRss < SupplejackCommon::Rss::Base
  base_url "http://www.95bfm.co.nz/sm/podcasts.rss"
  include_snippet "Global validations"
  attribute :category, default: "Audio"
  attribute :language, default: "en"
  attributes :content_partner, :display_content_partner, :display_collection, :
  attributes :usage, :copyright, default: "All rights reserved"
  attributes :landing_url, xpath: "//link"
  attribute :title, xpath: "//title"
  attribute :description, xpath: "//description", truncate: 350
  attribute :date, xpath: "//pubDate", date: true
  attribute :object_url, xpath: "//enclosure[type='audio/mpeg']"
  attribute :type, xpath: "//enclosure/@type"

  attribute :display_date do
    fetch("//pubDate").mapping(/\d{4}-\d{2}-\d{2}\s\d{2}:\d{2}:\d{2}\z/ => 'l\').select(:first)
  end

  attribute :internal_identifier do
    get(:landing_url).downcase
  end

  attribute :collection do
    get(:title).mapping(/.*$/ => '') + "95bFM"
  end
Discovery (API)
API (v3)

Public: Search, Retrieve Records, Create Sets/Stories

Private: Metrics, User Accounts, Key Mgmt./Throttling

RESTful, many params, faceted, 3 formats

https://apidigita.nz.org/v3/records.json?
api_key={key_hash}&text=supplejack
Request parameters, optional

The following may be specified as request parameters:

- **and** - Restricts search to records matching all facet values. Example:
  ```
  ...&and[content_partner][]=Kete+Horowhenua&and[category][]=Images
  ```
- **or** - Restricts search to records matching any of the specified facet values. Example:
  ```
  ...
  &or[category][]=Image&or[category][]=Videos
  ```
- **without** - Restricts search to records that don't match any of the facet values. Example:
  ```
  ...
  &without[category][]=Newspapers
  ```
- **per_page** - The number of records to return per page of search results. Maximum of 100. Defaults to 20.
- **page** - Which page of results to return. Defaults to 1.
- **facets** - A list of facet fields to include in the output. See the note on facets below for more information. Example:
  ```
  ...
  &facets=year,category
  ```
- **facet_per_page** - The number of facets to return per page of facets fields. Defaults to 10.
- **facets_page** - Which page of facet fields to return. Defaults to 1.
- **sort** - The field upon which results are sorted. Defaults to relevance sorting. The sort field must be either **date** or **syndication_date**.
- **direction** - The order in which the results are sorted. Possible values: **asc** (ascending) or **desc** (decending).
- **geo_bbox** - A geographic bounding box scoping a search to a geographic region. Order of latitude-longitude coordinates is north, west, south, east. The following example searches the Wellington region:
  ```
  ...
  &geo_bbox=-41,174,-42,175
  ```

Response format
Job Scheduling
Job Schedules - Staging

One-off

<table>
<thead>
<tr>
<th>Parser Script</th>
<th>First Run at</th>
<th>Environment</th>
<th>Mode</th>
<th>Frequency</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are no one off schedules programmed

Recurrent

Search:

<table>
<thead>
<tr>
<th>Parser Script</th>
<th>Next Job</th>
<th>Last Job</th>
<th>First Run at</th>
<th>Environment</th>
<th>Mode</th>
<th>Frequency</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digitalnz Blog</td>
<td>08 Nov 09:30</td>
<td>08 Jun 09:32</td>
<td>08 Oct 2014</td>
<td>staging</td>
<td>Full And Flush</td>
<td>monthly at 09:30 (30 9 8 **)</td>
<td>09:30</td>
</tr>
<tr>
<td>Kete Christchurch</td>
<td>19 Oct 08:50</td>
<td>28 Sep 08:52</td>
<td>12 Jun 2013</td>
<td>staging</td>
<td>Incremental</td>
<td>daily at 08:50 (50 8 ** *)</td>
<td>08:50</td>
</tr>
<tr>
<td>Kete Hamilton</td>
<td>25 Oct 09:05</td>
<td>27 Sep 09:08</td>
<td>12 Jun 2013</td>
<td>staging</td>
<td></td>
<td>weekly at 09:05 (5 9 ** 3)</td>
<td>09:05</td>
</tr>
<tr>
<td>Kete Hawkes Bay</td>
<td>25 Oct 09:05</td>
<td>27 Sep 09:08</td>
<td>12 Jun 2013</td>
<td>staging</td>
<td></td>
<td>weekly at 09:05 (5 9 ** 3)</td>
<td>09:05</td>
</tr>
</tbody>
</table>
## Finished Jobs - Production

<table>
<thead>
<tr>
<th>Type</th>
<th>Parser</th>
<th>Parser Type</th>
<th>Enrichment</th>
<th>Operator</th>
<th>First Run at</th>
<th>Mode</th>
<th>Duration</th>
<th>Records harvested</th>
<th>Validation/Failure Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enrichment</td>
<td>Nga Manu Flickr</td>
<td>record</td>
<td>flickr_geo_enrich</td>
<td>Scheduled</td>
<td>October 18, 2017 07:16:35</td>
<td>2 secs</td>
<td>0</td>
<td>/</td>
<td>View details</td>
</tr>
<tr>
<td>Harvest</td>
<td>Nga Manu Flickr</td>
<td>record</td>
<td></td>
<td>Scheduled</td>
<td>October 18, 2017 07:16:03</td>
<td>Full And Flush</td>
<td>32 secs</td>
<td>851</td>
<td>0/0</td>
</tr>
<tr>
<td>Harvest</td>
<td>Forest and Bird Website</td>
<td>record</td>
<td></td>
<td>Scheduled</td>
<td>October 18, 2017 05:40:04</td>
<td>Full And Flush</td>
<td>3 mins 20 secs</td>
<td>109</td>
<td>0/0</td>
</tr>
<tr>
<td>Enrichment</td>
<td>Youtube the wireless</td>
<td>record</td>
<td>youtube_license_enrich</td>
<td>Scheduled</td>
<td>October 18, 2017 05:32:32</td>
<td>9 mins 49 secs</td>
<td>427</td>
<td>/</td>
<td>View details</td>
</tr>
<tr>
<td>Harvest</td>
<td>Youtube the wireless</td>
<td>record</td>
<td></td>
<td>Scheduled</td>
<td>October 18, 2017 05:32:03</td>
<td>Full And Flush</td>
<td>29 secs</td>
<td>427</td>
<td>0/0</td>
</tr>
<tr>
<td>Harvest</td>
<td>Anyquestions</td>
<td>record</td>
<td></td>
<td>Scheduled</td>
<td>October 18, 2017 05:18:04</td>
<td>Full And Flush</td>
<td>1 min 1 sec</td>
<td>56</td>
<td>0/0</td>
</tr>
</tbody>
</table>
### Harvest progress: NZJ Forestry Science

<table>
<thead>
<tr>
<th>Operator</th>
<th>Ting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Environment</td>
<td>production</td>
</tr>
<tr>
<td>Mode</td>
<td>Full And Flush</td>
</tr>
<tr>
<td>First Run at</td>
<td>July 27, 2017 10:12:17</td>
</tr>
<tr>
<td>End Time</td>
<td>July 27, 2017 10:12:17</td>
</tr>
<tr>
<td>Duration</td>
<td></td>
</tr>
<tr>
<td>Records/sec</td>
<td></td>
</tr>
<tr>
<td>Records harvested</td>
<td>0</td>
</tr>
<tr>
<td>Records posted to API</td>
<td>0</td>
</tr>
<tr>
<td>Records Retried</td>
<td>0</td>
</tr>
<tr>
<td>Status Message</td>
<td>404 Not Found</td>
</tr>
</tbody>
</table>

**Harvest failure**

- 404 Not Found

**Validation errors** (0)
New Staging Schedule

* Parser Script

* Begin
18/10/2017 10:50 +1300

Recurring

Simple Schedule

Frequency

At time
00 - 00 -

Offset
Other Features
Kris Ceismic Enrich

```ruby
# ============== start of CEISMIC enrichment block
enrichment :kris_ceismic_enrich, priority: -1, required_for_active

  requires :ceismiccheck do
    ceismictext = ''
    ceismictext += primary[:title].first if primary[:title].present
    ceismictext += ''
    ceismictext += primary[:description].first if primary[:description].present
    ceismictext += ''
    ceismictext += primary[:subject].to_a.join(' ') if primary[:subject].present
    coll = "CEISMIC" if ceismictext.downcase.include? "chr"
    coll
  end

  attribute :collection, default: requirements[:ceismiccheck]

end
# ============== end of CEISMIC enrichment block
```
### Link Check Rules - Production

**Search:**

<table>
<thead>
<tr>
<th>Partner</th>
<th>Source Name</th>
<th>Source ID</th>
<th>Active?</th>
<th>Updated At</th>
</tr>
</thead>
<tbody>
<tr>
<td>95bFM</td>
<td>Bfm-rss</td>
<td>bfm-rss</td>
<td>true</td>
<td>2017-03-14T12:04:05.032+13:00</td>
</tr>
<tr>
<td>Christchurch Art Gallery</td>
<td>chch-art-galleries</td>
<td>chch-art-galleries</td>
<td>true</td>
<td>2014-03-24T15:56:29.000+13:00</td>
</tr>
</tbody>
</table>
Thanks
boost_nz_ Opening up the worlds information with Supplejack. Check out our blog or DigitalNZ to find out more 🌍 #boost #boostteam #boostnerds #supplejack #dnz #digitalnz #metadat #opensource #tshirts #friday #rubyonrails #search #dataharvesting
Supplejack

METADATA AGGREGATION AND DISCOVERY