

Technology Selection for Texas - Beyond the Portal

Nicholas Woodward (njw@austin.utexas.edu)



Grant

Develop a statewide metadata aggregation service for Texas digital collections

Collaboration with TSLAC, HPL and existing hub at UNT

TDL is uniquely positioned to provide a home for statewide metadata aggregation

Inclusion in DPLA would benefit institutions all across the state whose collections are not part of the Portal to Texas History

Metadata hubs survey

Phone interviews with six metadata hubs

Technology stacks: REPOX, custom Rails applications, Heidrun, others

Diverse metadata sources and technology stacks, but large reliance on
ContentDM and DSpace

Most solutions depend on command line and lack GUI

Cross-functional teams to handle all steps of the process

Supplejack

Developed at Digital Library of New Zealand

Open source developer community

Scalable to millions of records and hundreds of collections

Technology stack

Ruby on Rails applications for the API, Harvest Manager and Worker

MongoDB to store metadata

Solr search index

Scheduled metadata harvesting Ruby libraries for cron jobs

Parallel metadata harvesting using multiple Ruby servers

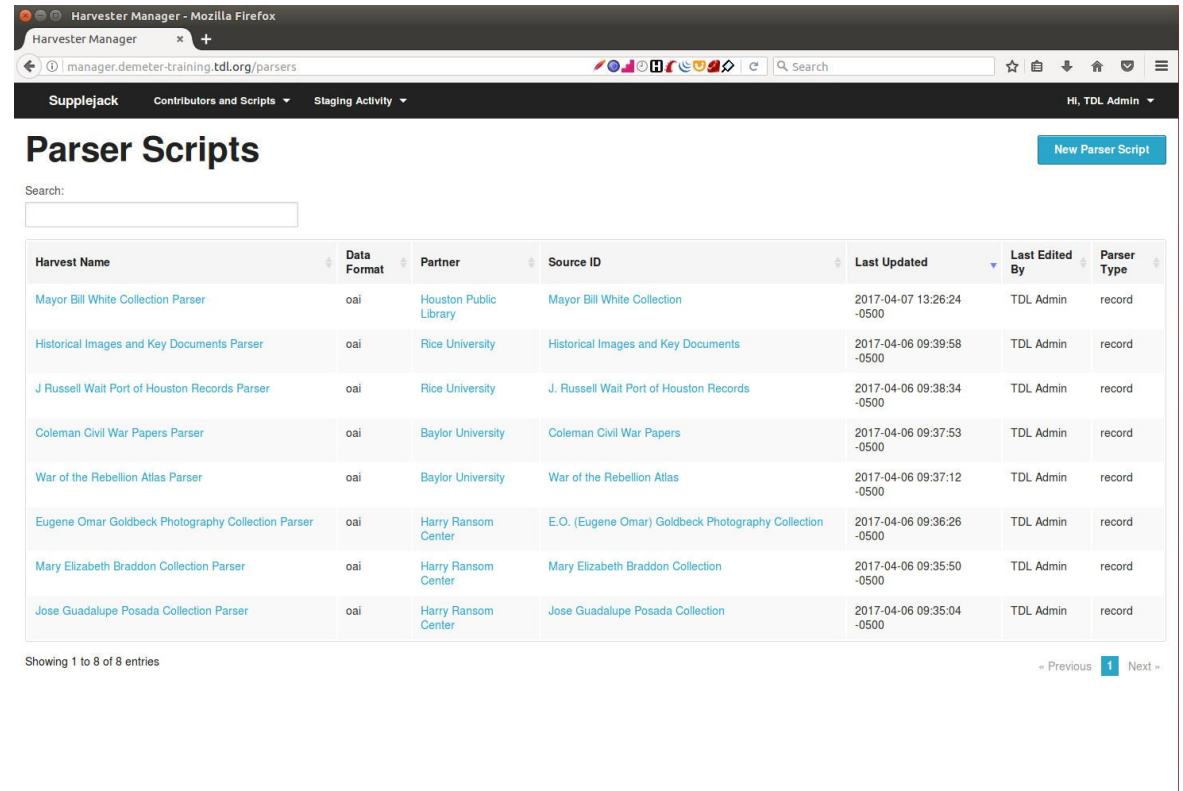
Collection parsers

Four institutions

Collections in ContentDM
and DSpace

OAI-PMH endpoints

Compound objects



The screenshot shows the Harvester Manager web application in a Mozilla Firefox browser. The page title is "Harvester Manager" and the URL is "manager.demeter-training.tdl.org/parsers". The navigation bar includes "Supplejack", "Contributors and Scripts", and "Staging Activity". The main heading is "Parser Scripts" with a "New Parser Script" button. A search bar is present. Below is a table with 8 entries, each representing a parser script. The table columns are: Harvest Name, Data Format, Partner, Source ID, Last Updated, Last Edited By, and Parser Type. The entries are for various collections including Mayor Bill White Collection, Historical Images and Key Documents, J. Russell Wait Port of Houston Records, Coleman Civil War Papers, War of the Rebellion Atlas, Eugene Omar Goldbeck Photography Collection, Mary Elizabeth Braddon Collection, and Jose Guadalupe Posada Collection.

Harvest Name	Data Format	Partner	Source ID	Last Updated	Last Edited By	Parser Type
Mayor Bill White Collection Parser	oai	Houston Public Library	Mayor Bill White Collection	2017-04-07 13:26:24 -0500	TDL Admin	record
Historical Images and Key Documents Parser	oai	Rice University	Historical Images and Key Documents	2017-04-06 09:39:58 -0500	TDL Admin	record
J Russell Wait Port of Houston Records Parser	oai	Rice University	J. Russell Wait Port of Houston Records	2017-04-06 09:38:34 -0500	TDL Admin	record
Coleman Civil War Papers Parser	oai	Baylor University	Coleman Civil War Papers	2017-04-06 09:37:53 -0500	TDL Admin	record
War of the Rebellion Atlas Parser	oai	Baylor University	War of the Rebellion Atlas	2017-04-06 09:37:12 -0500	TDL Admin	record
Eugene Omar Goldbeck Photography Collection Parser	oai	Harry Ransom Center	E.O. (Eugene Omar) Goldbeck Photography Collection	2017-04-06 09:36:26 -0500	TDL Admin	record
Mary Elizabeth Braddon Collection Parser	oai	Harry Ransom Center	Mary Elizabeth Braddon Collection	2017-04-06 09:35:50 -0500	TDL Admin	record
Jose Guadalupe Posada Collection Parser	oai	Harry Ransom Center	Jose Guadalupe Posada Collection	2017-04-06 09:35:04 -0500	TDL Admin	record

Showing 1 to 8 of 8 entries

« Previous 1 Next »

Parser script editor

Version control

Metadata mapping

XSLT paths

Namespaces

Conditional logic and
functions in Ruby

The screenshot displays the Harvester Manager web interface in a Mozilla Firefox browser. The page title is "Mary Elizabeth Braddon Collection Parser" with a "record" tag. The main content area shows a Ruby script for parsing XML records. The script defines a class `MaryElizabethBraddonCollectionParser` that inherits from `SupplejackCommon::Oai::Base`. It sets the base URL to `http://hrc.contentdm.oclc.org/oai/oai.php` and the metadata prefix to `oai_dc`. The script then defines various attributes and their corresponding XPath expressions for metadata elements like title, description, identifier, creator, publisher, subject, date, type, format, language, relation, rights, coverage, and internal identifier. The script ends with a `fetch` call to retrieve the last record.

```
1 class MaryElizabethBraddonCollectionParser < SupplejackCommon::Oai::Base
2   base_url 'http://hrc.contentdm.oclc.org/oai/oai.php'
3
4   metadata_prefix 'oai_dc'
5   set 'p15878coll53'
6
7   namespaces dc: "http://purl.org/dc/elements/1.1/",
8             dcterms: "http://purl.org/dc/terms/",
9             oai_dc: "http://www.openarchives.org/OAI/2.0/oai_dc/"
10
11   attributes :content_partner, default: "Harry Ransom Center"
12   attributes :primary_collection, default: "Mary Elizabeth Braddon Collection"
13
14   attributes :title, xpath: "/record/metadata/oai_dc:dc:title"
15
16   attributes :description, xpath: "/record/metadata/oai_dc:dc:description", default: ""
17
18   attributes :identifier, xpath: "/record/metadata/oai_dc:dc:identifier"
19
20   attributes :creator, xpath: "/record/metadata/oai_dc:dc:creator"
21
22   attributes :publisher, xpath: "/record/metadata/oai_dc:dc:publisher"
23
24   attributes :subject, xpath: "/record/metadata/oai_dc:dc:subject"
25
26   attributes :date, xpath: "/record/metadata/oai_dc:dc:date", date: true
27
28   attributes :type, xpath: "/record/metadata/oai_dc:dc:type"
29
30   attributes :format, xpath: "/record/metadata/oai_dc:dc:format"
31
32   attributes :language, xpath: "/record/metadata/oai_dc:dc:language", default: "en"
33
34   attributes :relation, xpath: "/record/metadata/oai_dc:dc:relation"
35
36   attributes :rights, xpath: "/record/metadata/oai_dc:dc:rights"
37
38   attributes :coverage, xpath: "/record/metadata/oai_dc:dc:coverage"
39
40   attributes :internal_identifier do
41     fetch('/record/metadata/oai_dc:dc:identifier').select(:last)
42   end
43 end
```

On the right side of the interface, there are buttons for "Preview", "Edit current version", and a "History" section showing the "initial commit" by TDL. At the bottom, there are buttons for "Rename Parser", "Change Data Source", "Delete Parser Script", and "Disable Full & Flush harvest mode". A "Message" input field and an "Update Parser Script" button are located at the bottom left.

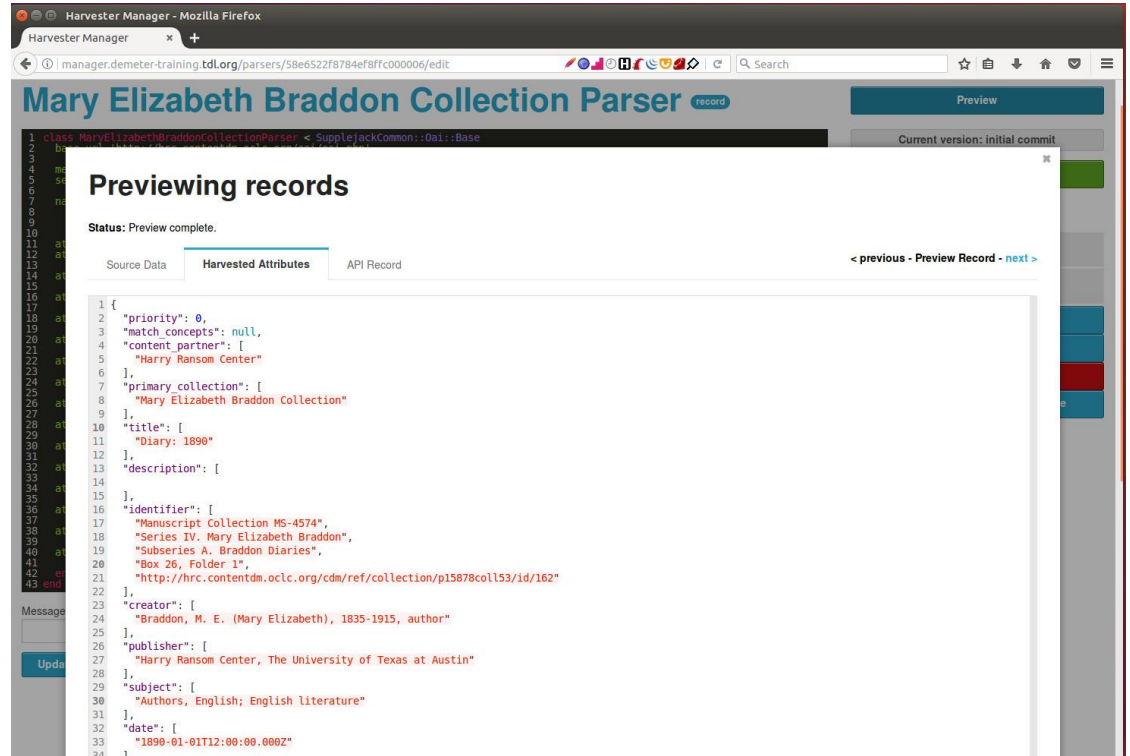
Harvest results preview

Results in real time

Raw data and metadata
mapping

API Record preview

Usefulness depends on
structure of feed results



The screenshot shows the Harvester Manager interface in Mozilla Firefox. The browser address bar displays the URL: `manager.demeter-training.tdl.org/parsers/58e6522f8784ef8ffc000006/edit`. The page title is "Mary Elizabeth Braddon Collection Parser". A "Preview" button is visible in the top right corner. The main content area is titled "Previewing records" and shows a status message: "Status: Preview complete." Below this, there are three tabs: "Source Data", "Harvested Attributes" (which is selected), and "API Record". The "Harvested Attributes" tab displays a JSON record structure. The "API Record" tab is also visible, showing a similar structure. The JSON record is as follows:

```
1 {
2   "priority": 0,
3   "match concepts": null,
4   "content_partner": [
5     "Harry Ransom Center"
6   ],
7   "primary collection": [
8     "Mary Elizabeth Braddon Collection"
9   ],
10  "title": [
11    "Diary: 1890"
12  ],
13  "description": [
14    "Manuscript Collection MS-4574",
15    "Series IV. Mary Elizabeth Braddon",
16    "Subseries A. Braddon Diaries",
17    "Box 26, Folder 1",
18    "http://hrc.contentdm.oclc.org/cdm/ref/collection/p15878coll53/id/162"
19  ],
20  "identifier": [
21    "Braddon, M. E. (Mary Elizabeth), 1835-1915, author"
22  ],
23  "publisher": [
24    "Harry Ransom Center, The University of Texas at Austin"
25  ],
26  "subject": [
27    "Authors, English; English literature"
28  ],
29  "date": [
30    "1890-01-01T12:00:00.000Z"
31  ]
32 }
```

The interface also includes a "Current version: initial commit" label and a "Update" button at the bottom left.

Successes...

Development and production Supplejack environments in AWS

Custom metadata API

Metadata harvesting of eight collections in ContentDM and DSpace

and Challenges

Environment setup and maintenance issues

Handling metadata enrichments asynchronously from harvests

Export and review of harvest results

Roadmap ahead

Manage the scheduling process for metadata harvests

Separate server for dynamic image thumbnail generation

OAI-PMH endpoint for re-harvesting of metadata

Expand on collections in the pilot to include others that meet minimum criteria (e.g. OAI-PMH endpoint, metadata standards, etc.)

Develop workflow and add staff to support regular metadata harvesting

Thank you



DIGITAL PUBLIC LIBRARY
OF AMERICA



TEXAS STATE LIBRARY
AND
ARCHIVES COMMISSION



UNT[®]

UNIVERSITY
OF NORTH TEXAS