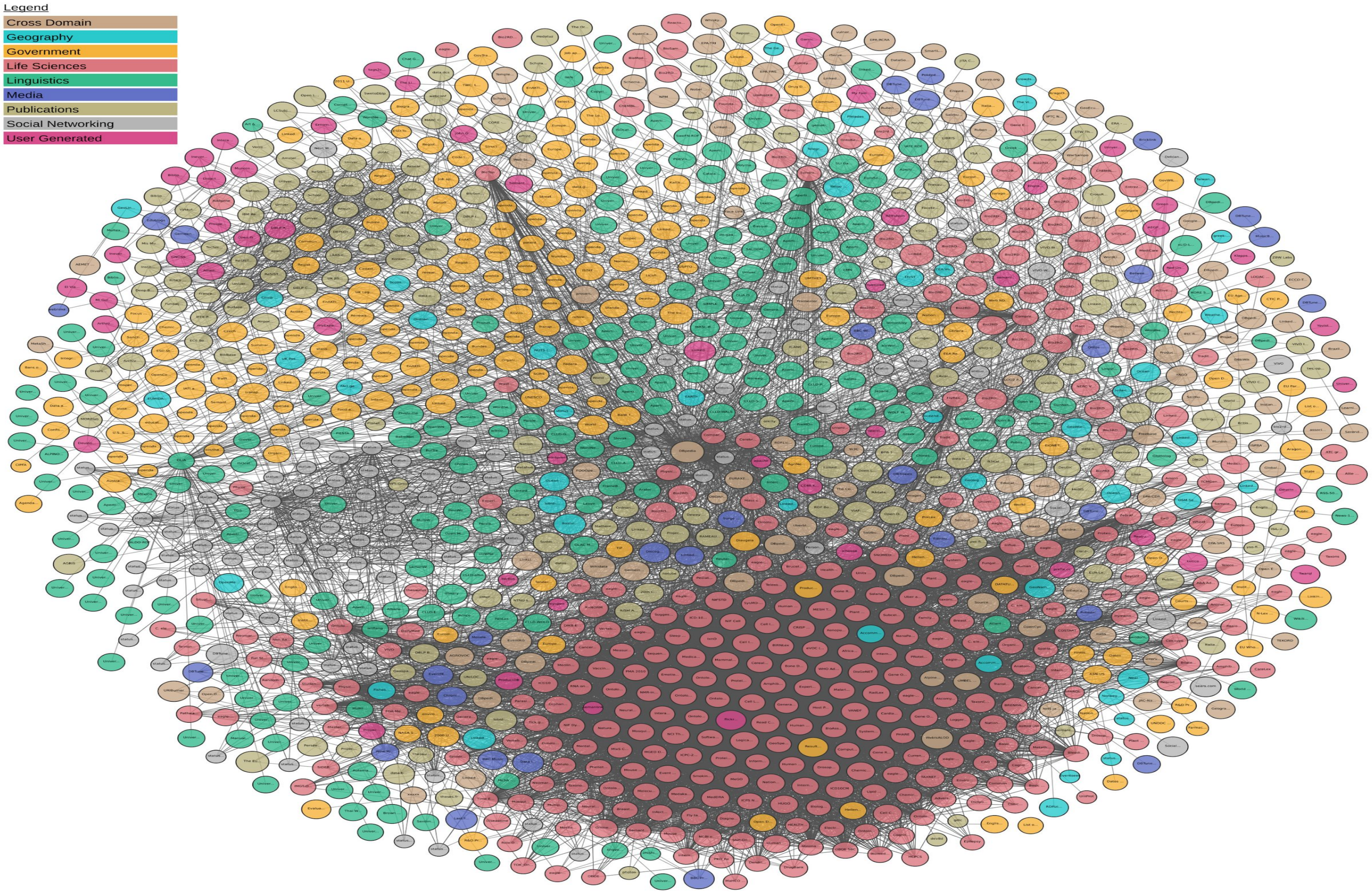
A full stack tool for preparing Library data for Linked Open Data application

A Demo

Danoosh Davoodi

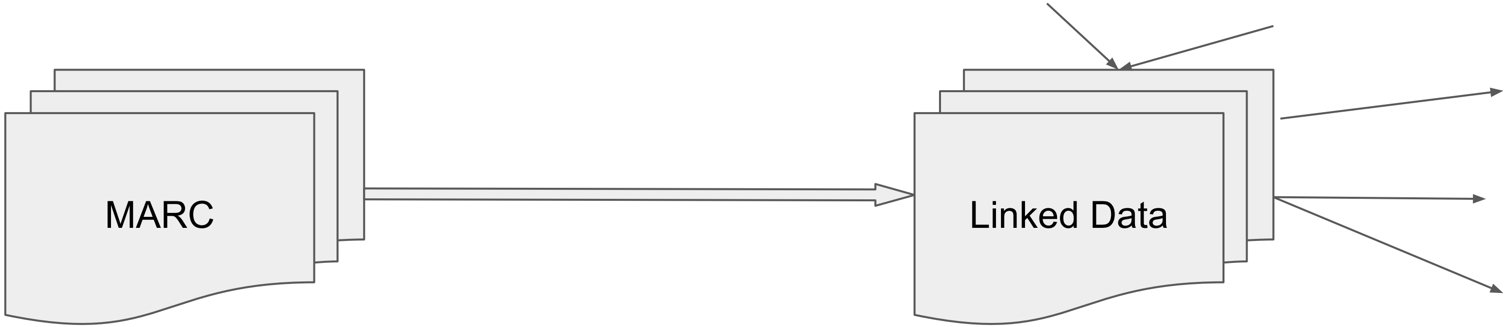
Linked Data - briefly

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.



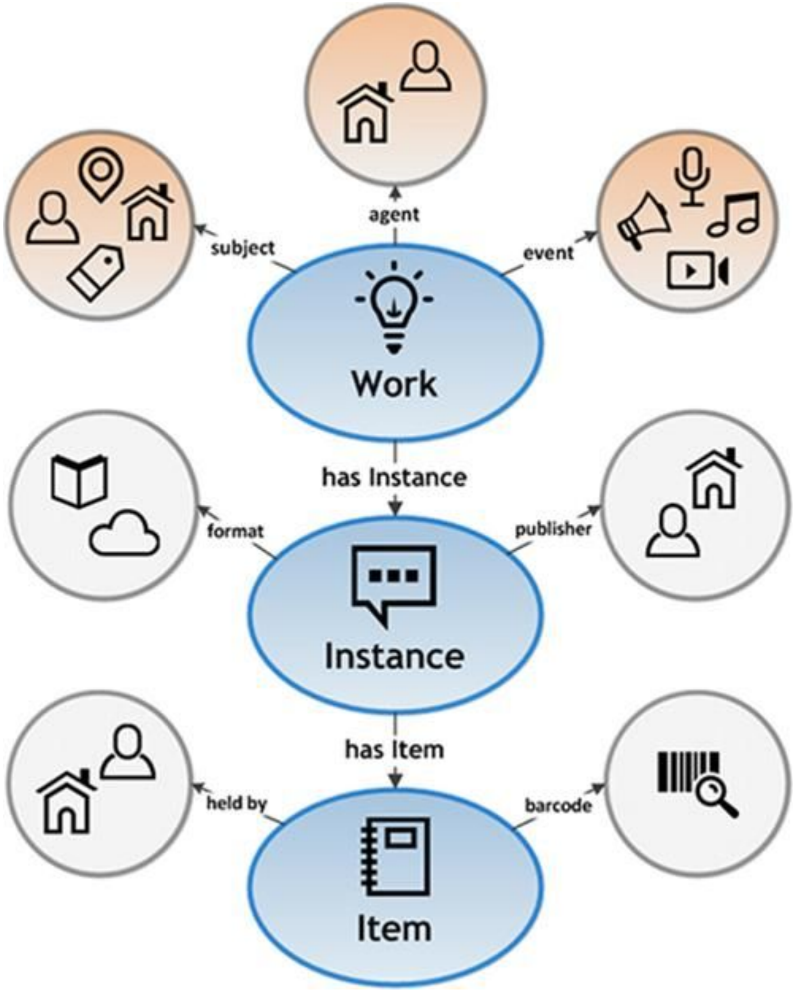
The Linked Open Data Cloud from lod-cloud.net

Libraries??!!



bf:

BIBFRAME


agent
subject
event
Work
has Instance
format
publisher
Instance
has Item
held by
barcode
Item

Move to Linked Data at UAL

ERA

education & research archive

SHARE

Discovery

Virtual

Environment

www.share-vde.org

LD4L

Linked Data for Libraries

CANLINK

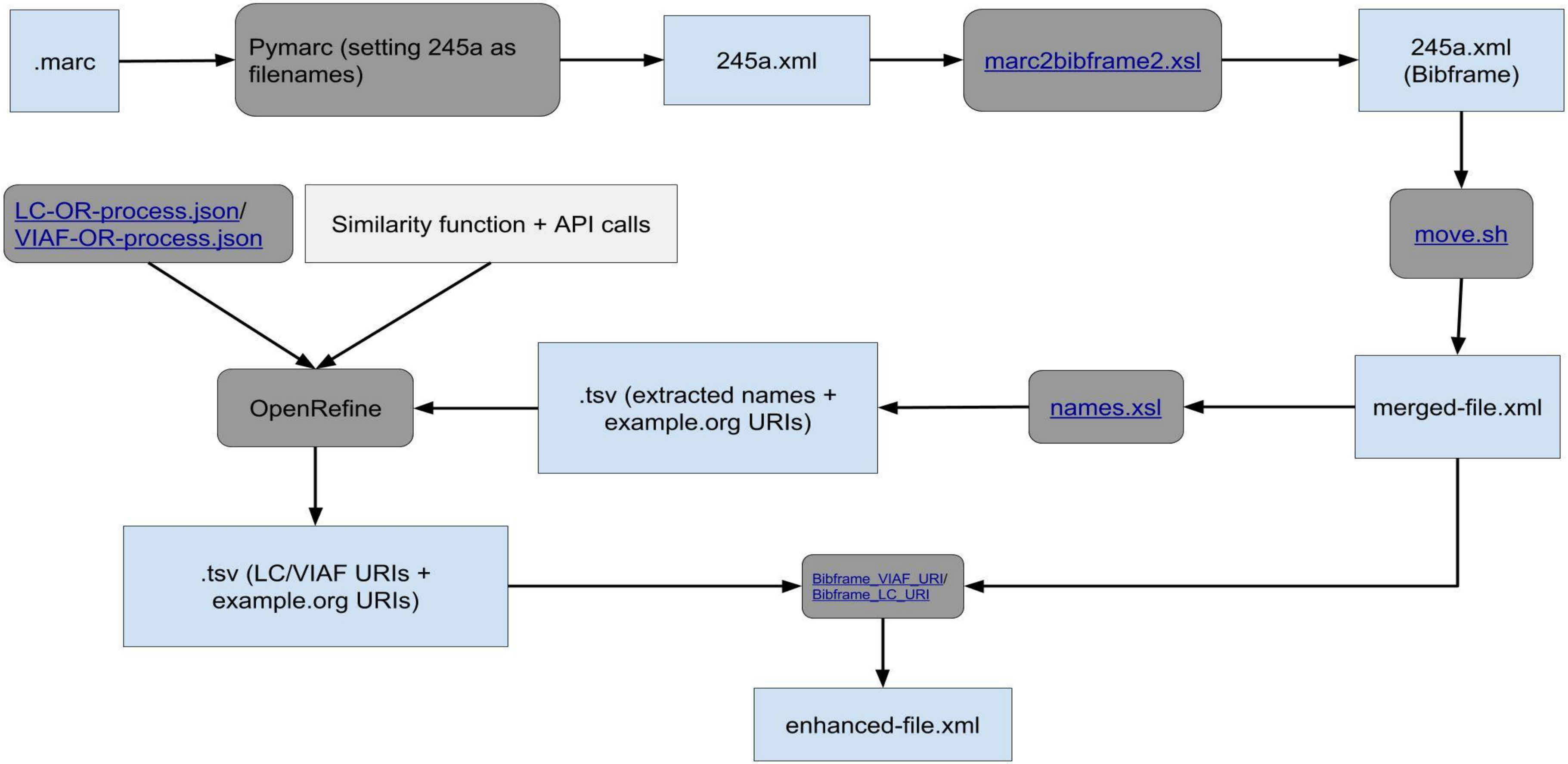
...

Data Enrichment for LD

Test are capability to transform our MARC or MARC/XML data to BIBFRAME and do some URI enrichment on authors, subject, etc

Knowledge that we had:

- Oxygen
- Python
- BASH scripting
- OpenRefine


.marc
Pymarc (setting 245a as filenames)
245a.xml
marc2bibframe2.xsl
245a.xml (Bibframe)
move.sh
merged-file.xml
names.xsl
.tsv (extracted names + example.org URIs)
OpenRefine
LC-OR-process.json/ VIAF-OR-process.json
Similarity function + API calls
.tsv (LC/VIAF URIs + example.org URIs)
Bibframe_VIAF_URI/ Bibframe_LC_URI
enhanced-file.xml

```
<bf:Agent rdf:about="http://example.org/6815285#Agent100-13">
  <rdf:type
rdf:resource="http://id.loc.gov/ontologies/bibframe/Person"/>
  <bflc:name00MatchKey>Veksner, Simon,</bflc:name00MatchKey>
  <bflc:primaryContributorName00MatchKey>Veksner,
       Simon,</bflc:primaryContributorName00MatchKey>
  <bflc:name00MarcKey>1001 $aVeksner,
Simon,$eauthor.</bflc:name00MarcKey>
  <rdfs:label>Veksner, Simon,</rdfs:label>
</bf:Agent>
```

↓

Veksner, Simon,	**http://id.loc.gov/ontologies/bibframe/Person**	http://example.org/6815285#Agent100-13

Name	Ingest_key	LC	VIAF
Veksner, Simon,	http://example.org/6815285#Agent100-13	no2011039513	169080997

```
<bf:Agent rdf:about="http://id.loc.gov/authorities/names/no2011039513">
   <rdf:type rdf:resource="http://id.loc.gov/ontologies/bibframe/Organization"/>
   <bflc:name10MatchKey>Canadian Medical Association.</bflc:name10MatchKey>
   <bflc:name10MarcKey>7102 $aCanadian Medical Association.</bflc:name10MarcKey>
   <rdfs:label>Canadian Medical Association.</rdfs:label>
   <bf:identifiedBy>
      <bf:Identifier>
         <rdf:value rdf:about="http://viaf.org/viaf/169080997"/>
      </bf:Identifier>
   </bf:identifiedBy>
</bf:Agent>
```

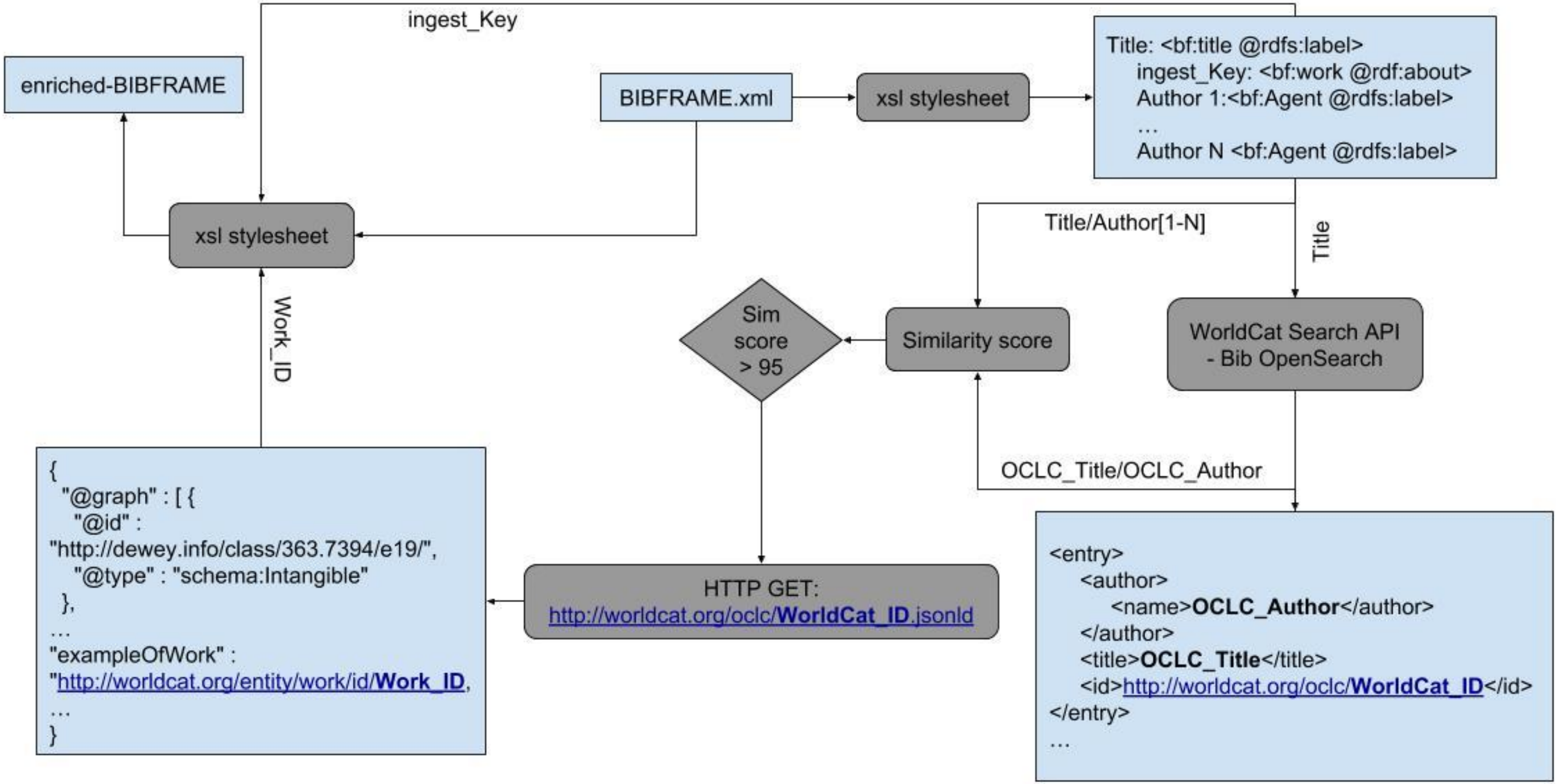
Process	**Time**	**Tool**
Converting .marc to MARC/XML	7 - 8 mins	pymarc
Converting MARC/XML to BIBFRAME (and merging)	40 - 50 mins	Oxygen / bash
Extracting names (or subjects) from the bibframe file	Less than 2 mins	Oxygen
OpenRefine process	Few seconds	OpenRefine - GREL
Enriching names with URIs (from VIAF)	30 - 35 mins	OpenRefine + VIAF recon java client
Enriching names with URIs (from LC)	90 - 120 mins	OpenRefine + LC recon client
Enriching subjects with URis (from LC)	70 - 90 mins	OpenRefine + LC recon client
OpenRefine process	Few seconds	OpenRefine - GREL
Ingesting (replacing example.org URIs)	60 - 70 mins (using Saxon EE on Oxygen) 100 - 120 mins (using Saxon HE on command-line)	Oxygen / Saxon command-line

UAL Linked Data Enrichment Tool

- Developed fully in python environment
- Has a GUI
- Flexibility to search as many as 6 APIs for URI enrichments
- Create statistical data (and visual graphs)
- Allow multi processing
- Works with BIBFRAME and MARC (as input)
- Merge BIBFRAME files for better processing

APIs for URI enrichment

Source	Query
LoC	http://id.loc.gov/authorities/label/QUERY
	http://id.loc.gov/authorities/suggest/?q=QUERY
VIAF	https://viaf.org/viaf/search?query=local.personalNames+all+%22" + QUERY + "%22&sortKeys=holdingscount&recordSchema=BriefVIAF&httpAccept=application/json
	https://viaf.org/viaf/search?query=local.corporateNames+all+%22" + QUERY + "%22&sortKeys=holdingscount&recordSchema=BriefVIAF&httpAccept=application/json
	http://viaf.org/viaf/AutoSuggest?query="QUERY



```
Title: <bf:title @rdfs:label>
    ingest_Key: <bf:work @rdf:about>
    Author 1:<bf:Agent @rdfs:label>
    …
    Author N <bf:Agent @rdfs:label>
```

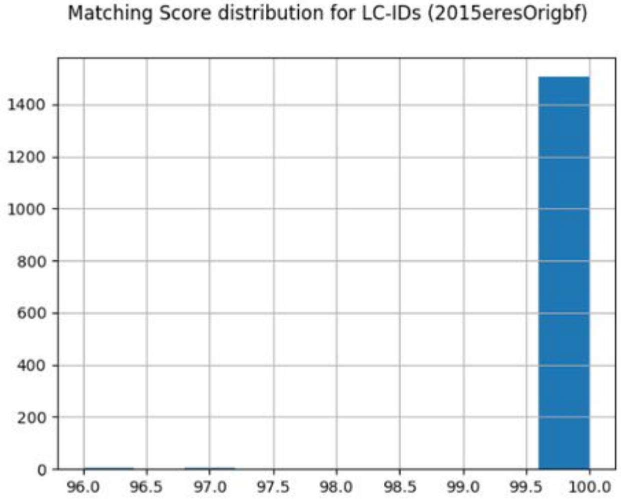
```
{
  "@graph" : [ {
    "@id" :
"http://dewey.info/class/363.7394/e19/",
    "@type" : "schema:Intangible"
  },
…
"exampleOfWork" :
"http://worldcat.org/entity/work/id/Work_ID,
…
}
```

```
<entry>
    <author>
        <name>OCLC_Author</author>
    </author>
    <title>OCLC_Title</title>
    <id>http://worldcat.org/oclc/WorldCat_ID</id>
</entry>
…
```

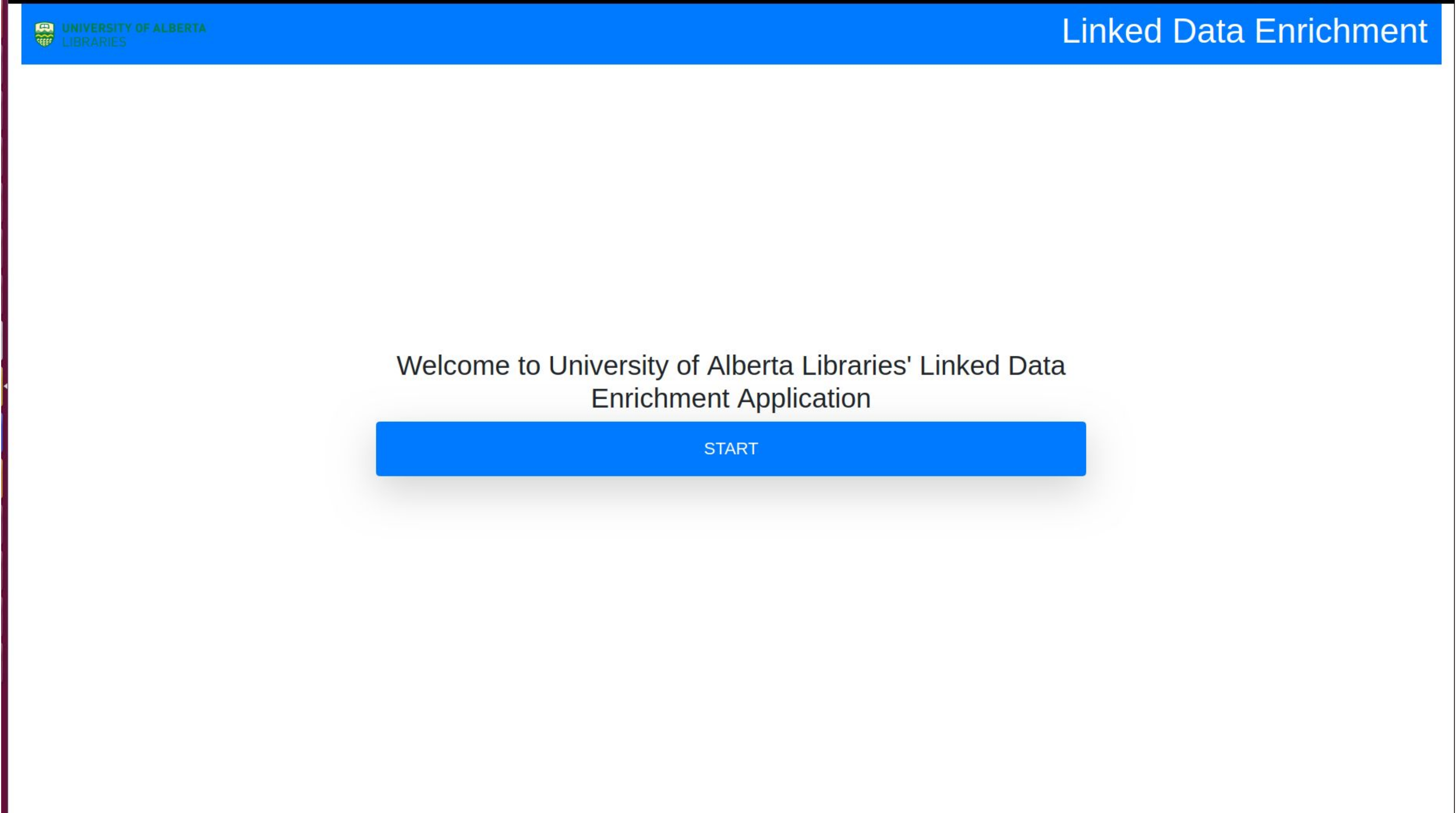
2015eresOrigbf was processed in 0:36:46

1816 names were extracted - 1723 unique names --- 1661 Personal names and 62 Corporate names

API searched	hits	hit_rate
VIAF personal	1463	84.91004063
VIAF general	11	0.6384213581
LC (suggest)	270	15.67034243
LC (didyoumean)	216	12.53627394
VIAF corporate	38	2.205455601

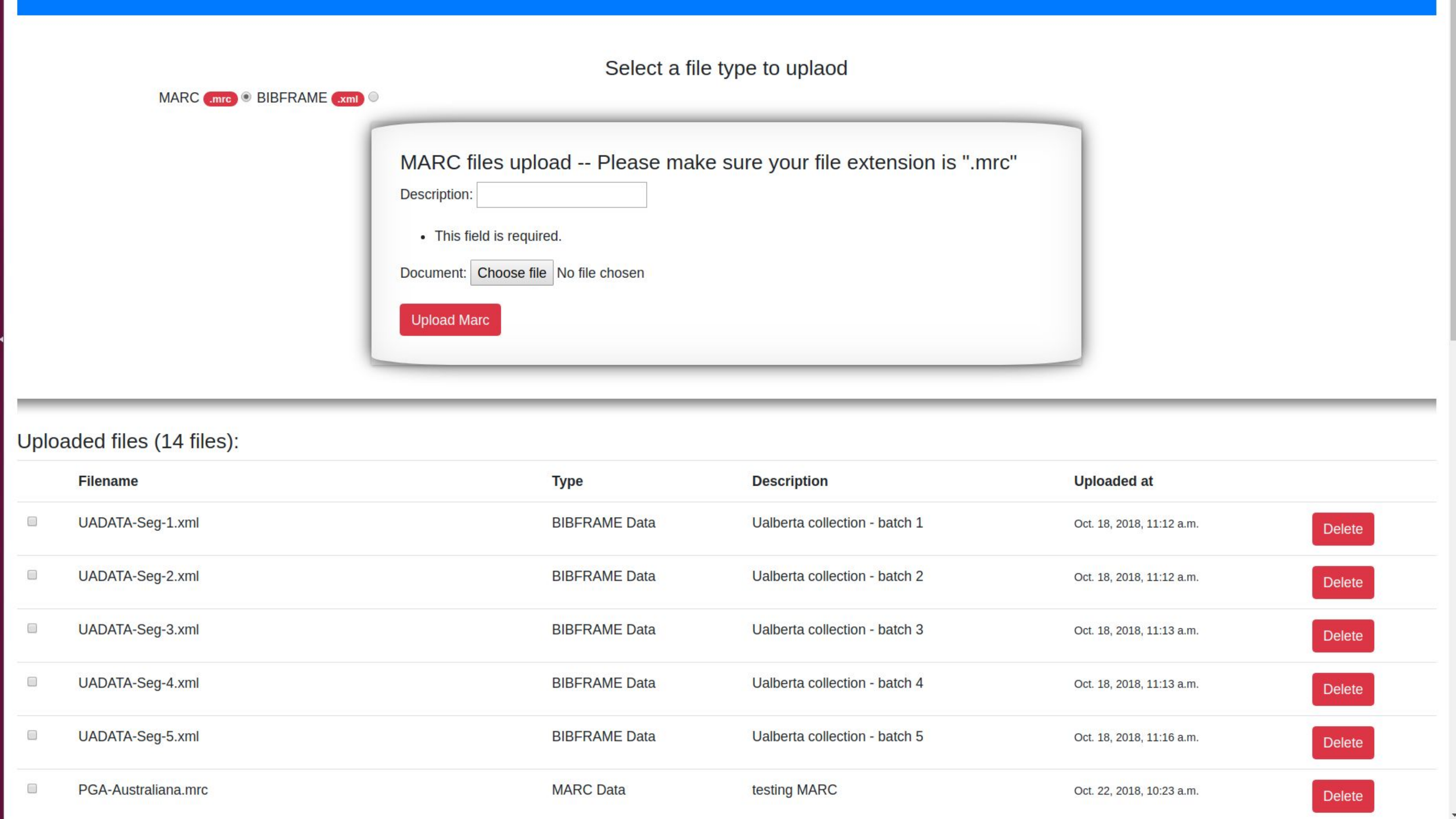

Matching Score distribution for LC-IDs (2015eresOrigbf)

names enriched	average matching score	median matching score	variance of matching score	standard-div of matching score	hit rate



Linked Data Enrichment

Welcome to University of Alberta Libraries' Linked Data Enrichment Application



Select a file type to uplaod

MARC .mrc ● BIBFRAME .xml ○

MARC files upload -- Please make sure your file extension is ".mrc"

Description:

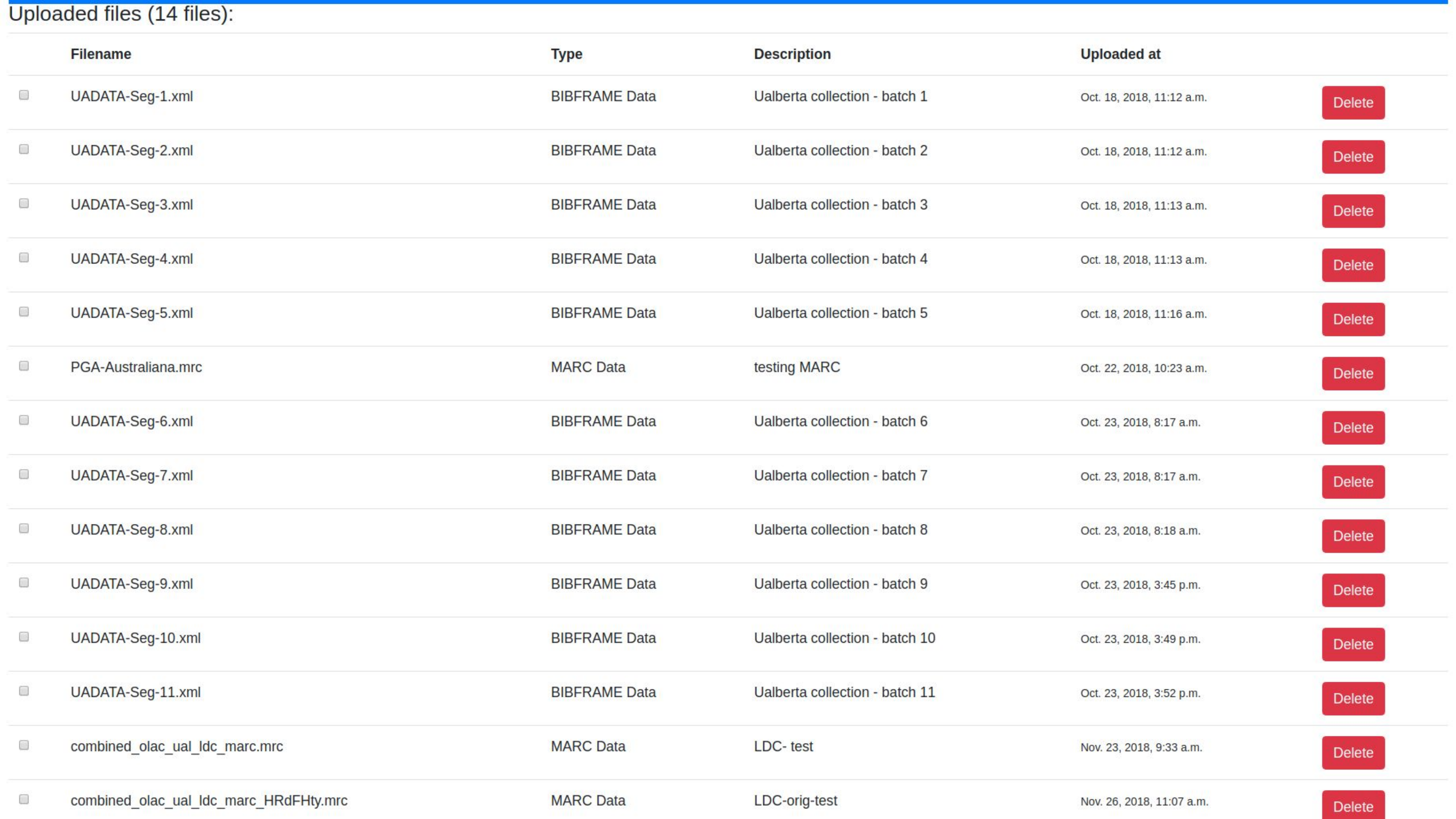
- This field is required.

Document: Choose file No file chosen

Upload Marc

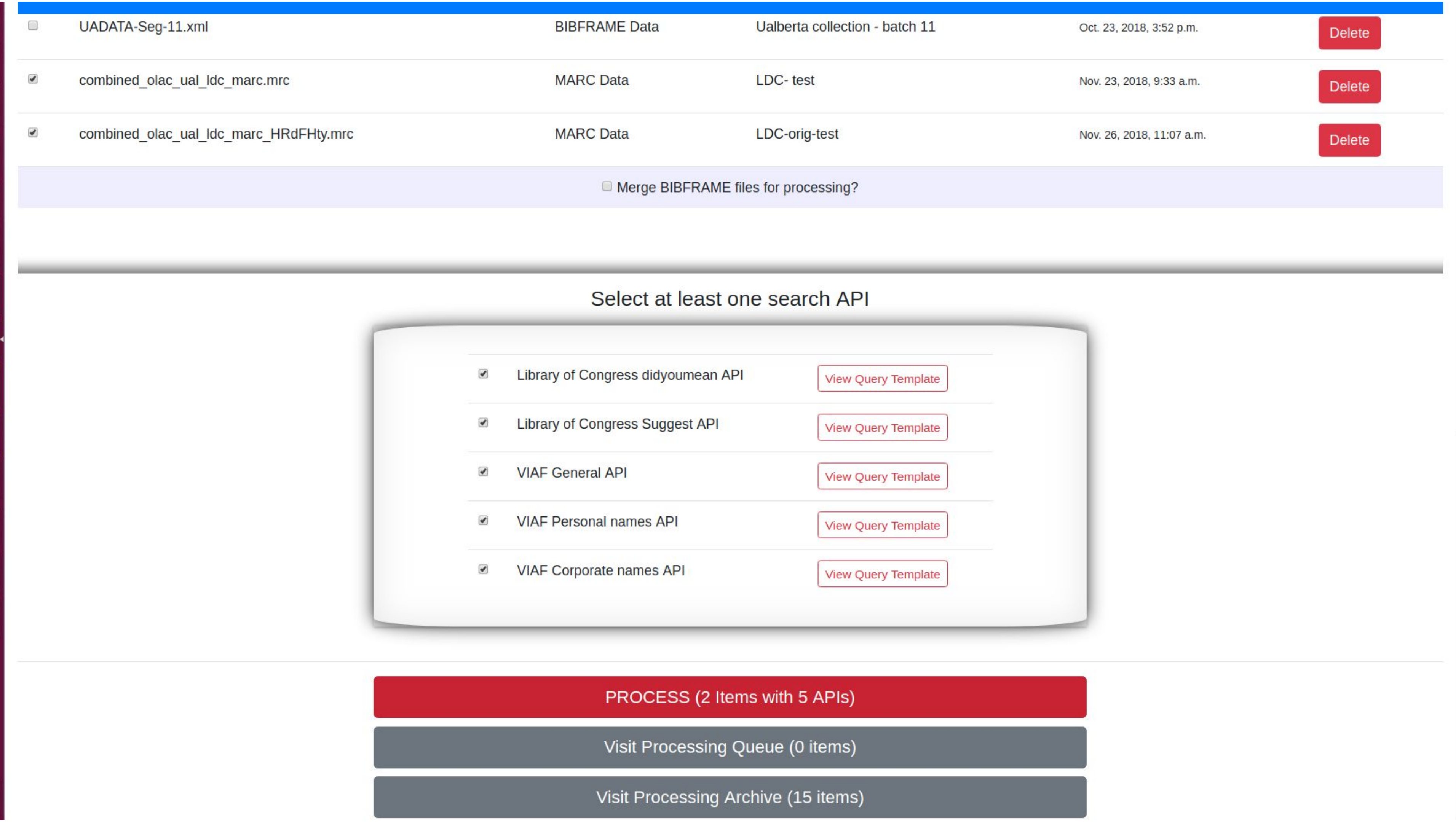
Uploaded files (14 files):

	Filename	Type	Description	Uploaded at	
□	UADATA-Seg-1.xml	BIBFRAME Data	Ualberta collection - batch 1	Oct. 18, 2018, 11:12 a.m.	Delete
□	UADATA-Seg-2.xml	BIBFRAME Data	Ualberta collection - batch 2	Oct. 18, 2018, 11:12 a.m.	Delete
□	UADATA-Seg-3.xml	BIBFRAME Data	Ualberta collection - batch 3	Oct. 18, 2018, 11:13 a.m.	Delete
□	UADATA-Seg-4.xml	BIBFRAME Data	Ualberta collection - batch 4	Oct. 18, 2018, 11:13 a.m.	Delete
□	UADATA-Seg-5.xml	BIBFRAME Data	Ualberta collection - batch 5	Oct. 18, 2018, 11:16 a.m.	Delete
□	PGA-Australiana.mrc	MARC Data	testing MARC	Oct. 22, 2018, 10:23 a.m.	Delete



Uploaded files (14 files):

	Filename	Type	Description	Uploaded at	
☐	UADATA-Seg-1.xml	BIBFRAME Data	Ualberta collection - batch 1	Oct. 18, 2018, 11:12 a.m.	Delete
☐	UADATA-Seg-2.xml	BIBFRAME Data	Ualberta collection - batch 2	Oct. 18, 2018, 11:12 a.m.	Delete
☐	UADATA-Seg-3.xml	BIBFRAME Data	Ualberta collection - batch 3	Oct. 18, 2018, 11:13 a.m.	Delete
☐	UADATA-Seg-4.xml	BIBFRAME Data	Ualberta collection - batch 4	Oct. 18, 2018, 11:13 a.m.	Delete
☐	UADATA-Seg-5.xml	BIBFRAME Data	Ualberta collection - batch 5	Oct. 18, 2018, 11:16 a.m.	Delete
☐	PGA-Australiana.mrc	MARC Data	testing MARC	Oct. 22, 2018, 10:23 a.m.	Delete
☐	UADATA-Seg-6.xml	BIBFRAME Data	Ualberta collection - batch 6	Oct. 23, 2018, 8:17 a.m.	Delete
☐	UADATA-Seg-7.xml	BIBFRAME Data	Ualberta collection - batch 7	Oct. 23, 2018, 8:17 a.m.	Delete
☐	UADATA-Seg-8.xml	BIBFRAME Data	Ualberta collection - batch 8	Oct. 23, 2018, 8:18 a.m.	Delete
☐	UADATA-Seg-9.xml	BIBFRAME Data	Ualberta collection - batch 9	Oct. 23, 2018, 3:45 p.m.	Delete
☐	UADATA-Seg-10.xml	BIBFRAME Data	Ualberta collection - batch 10	Oct. 23, 2018, 3:49 p.m.	Delete
☐	UADATA-Seg-11.xml	BIBFRAME Data	Ualberta collection - batch 11	Oct. 23, 2018, 3:52 p.m.	Delete
☐	combined_olac_ual_ldc_marc.mrc	MARC Data	LDC- test	Nov. 23, 2018, 9:33 a.m.	Delete
☐	combined_olac_ual_ldc_marc_HRdFHty.mrc	MARC Data	LDC-orig-test	Nov. 26, 2018, 11:07 a.m.	Delete



☐	UADATA-Seg-11.xml	BIBFRAME Data	Ualberta collection - batch 11	Oct. 23, 2018, 3:52 p.m.	Delete
☑	combined_olac_ual_ldc_marc.mrc	MARC Data	LDC- test	Nov. 23, 2018, 9:33 a.m.	Delete
☑	combined_olac_ual_ldc_marc_HRdFHty.mrc	MARC Data	LDC-orig-test	Nov. 26, 2018, 11:07 a.m.	Delete

☐ Merge BIBFRAME files for processing?

Select at least one search API

☑ Library of Congress didyoumean API

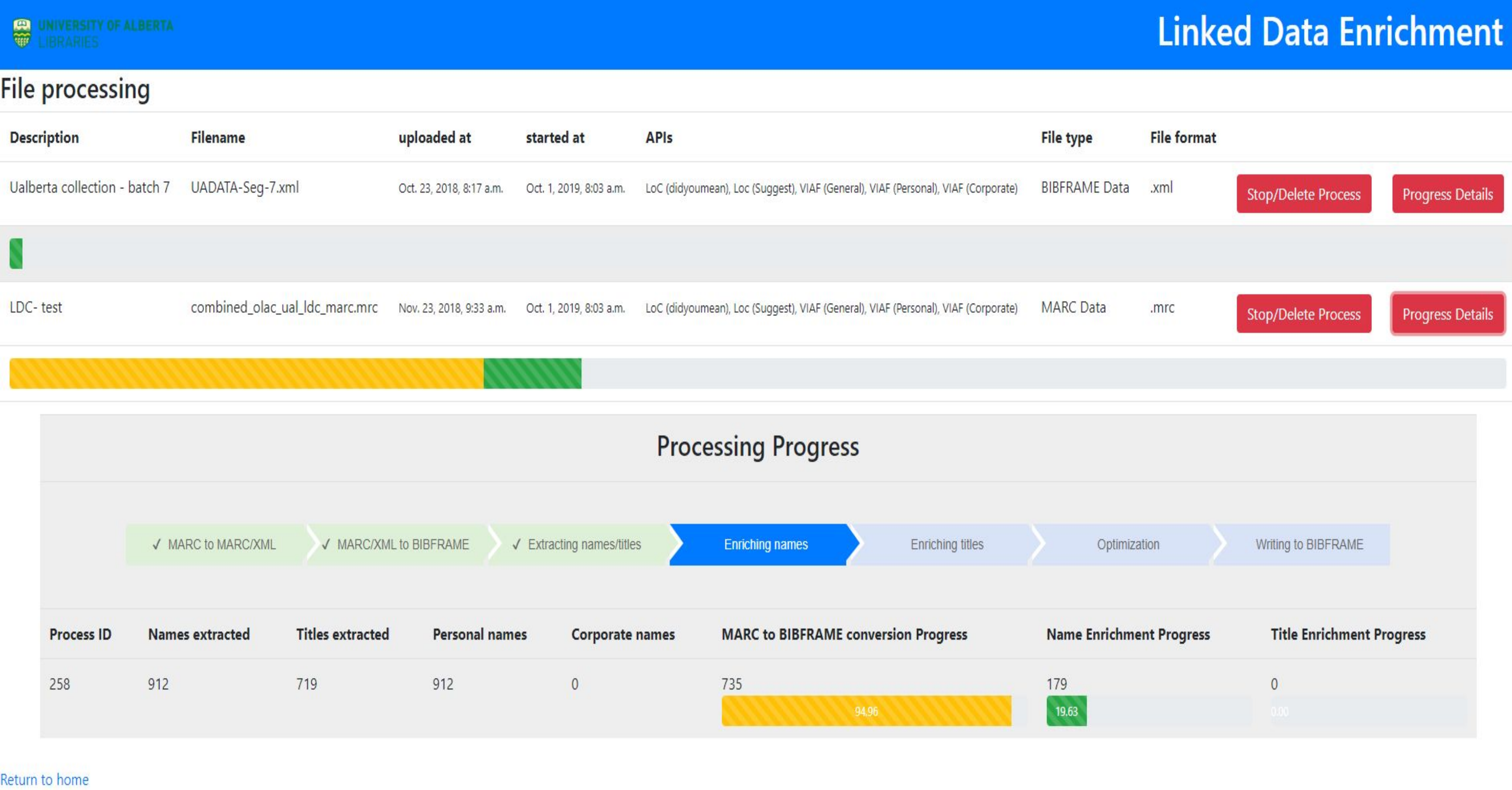
☑ Library of Congress Suggest API

☑ VIAF General API

☑ VIAF Personal names API

☑ VIAF Corporate names API

View Query Template
View Query Template
View Query Template
View Query Template
View Query Template
PROCESS (2 Items with 5 APIs)
Visit Processing Queue (0 items)
Visit Processing Archive (15 items)



Linked Data Enrichment

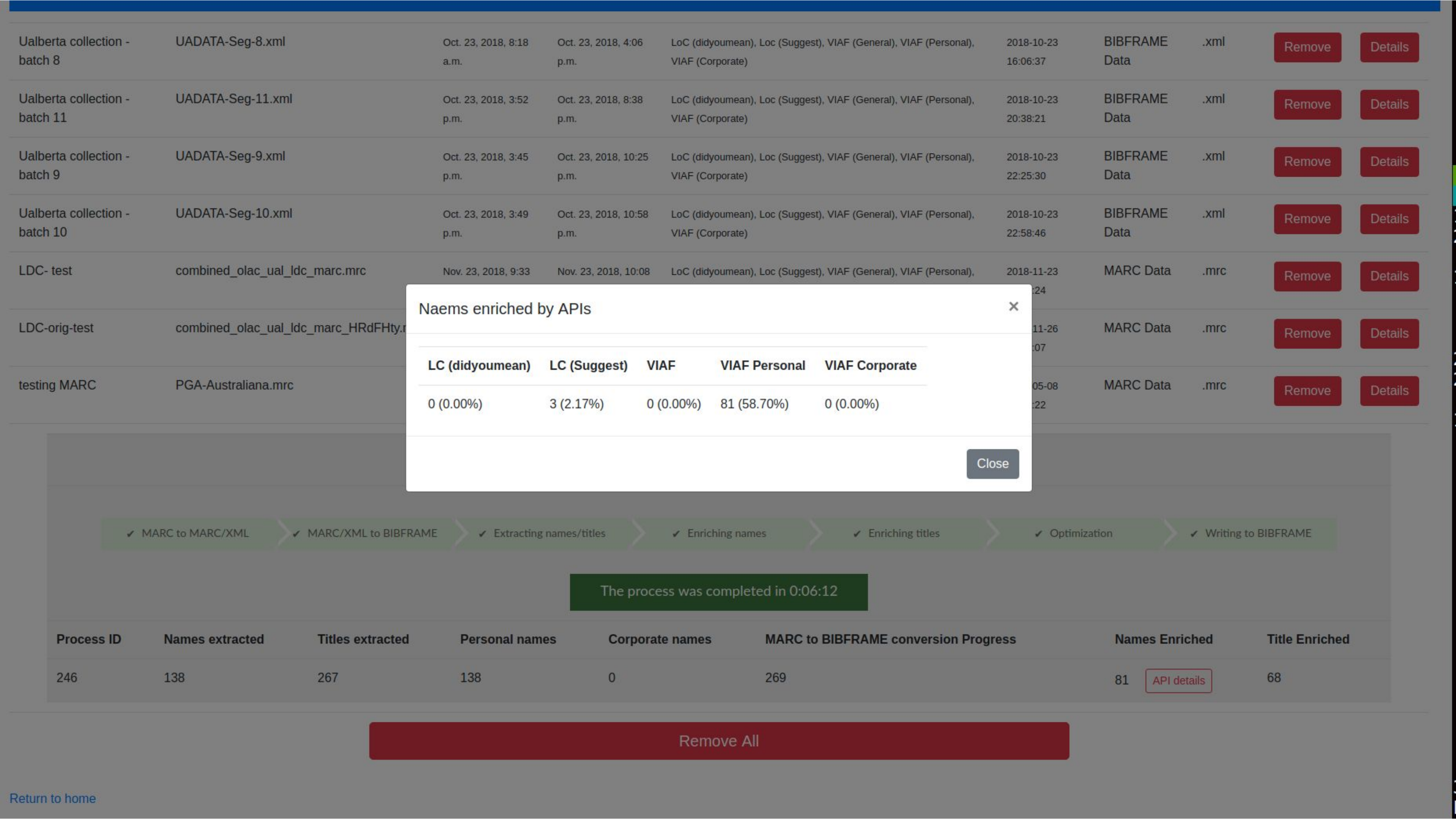
File processing

Description	Filename	uploaded at	started at	APIs	File type	File format		
Ualberta collection - batch 7	UADATA-Seg-7.xml	Oct. 23, 2018, 8:17 a.m.	Oct. 1, 2019, 8:03 a.m.	LoC (didyoumean), Loc (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	BIBFRAME Data	.xml	Stop/Delete Process	Progress Details
LDC- test	combined_olac_ual_ldc_marc.mrc	Nov. 23, 2018, 9:33 a.m.	Oct. 1, 2019, 8:03 a.m.	LoC (didyoumean), Loc (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	MARC Data	.mrc	Stop/Delete Process	Progress Details

Processing Progress

Process ID	Names extracted	Titles extracted	Personal names	Corporate names	MARC to BIBFRAME conversion Progress	Name Enrichment Progress	Title Enrichment Progress
258	912	719	912	0	735 94.96	179 19.63	0 0.00

Return to home



Ualberta collection - batch 8	UADATA-Seg-8.xml	Oct. 23, 2018, 8:18 a.m.	Oct. 23, 2018, 4:06 p.m.	LoC (didyoumean), LoC (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	2018-10-23 16:06:37	BIBFRAME Data	.xml	Remove	Details
Ualberta collection - batch 11	UADATA-Seg-11.xml	Oct. 23, 2018, 3:52 p.m.	Oct. 23, 2018, 8:38 p.m.	LoC (didyoumean), LoC (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	2018-10-23 20:38:21	BIBFRAME Data	.xml	Remove	Details
Ualberta collection - batch 9	UADATA-Seg-9.xml	Oct. 23, 2018, 3:45 p.m.	Oct. 23, 2018, 10:25 p.m.	LoC (didyoumean), LoC (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	2018-10-23 22:25:30	BIBFRAME Data	.xml	Remove	Details
Ualberta collection - batch 10	UADATA-Seg-10.xml	Oct. 23, 2018, 3:49 p.m.	Oct. 23, 2018, 10:58 p.m.	LoC (didyoumean), LoC (Suggest), VIAF (General), VIAF (Personal), VIAF (Corporate)	2018-10-23 22:58:46	BIBFRAME Data	.xml	Remove	Details
LDC- test	combined_olac_ual_ldc_marc.mrc	Nov. 23, 2018, 9:33	Nov. 23, 2018, 10:08	LoC (didyoumean), LoC (Suggest), VIAF (General), VIAF (Personal),	2018-11-23 :24	MARC Data	.mrc	Remove	Details
LDC-orig-test	combined_olac_ual_ldc_marc_HRdFHty.r				11-26 :07	MARC Data	.mrc	Remove	Details
testing MARC	PGA-Australiana.mrc				05-08 :22	MARC Data	.mrc	Remove	Details

Naems enriched by APIs

×

LC (didyoumean)	LC (Suggest)	VIAF	VIAF Personal	VIAF Corporate
0 (0.00%)	3 (2.17%)	0 (0.00%)	81 (58.70%)	0 (0.00%)

Close

✓ MARC to MARC/XML ✓ MARC/XML to BIBFRAME ✓ Extracting names/titles ✓ Enriching names ✓ Enriching titles ✓ Optimization ✓ Writing to BIBFRAME

The process was completed in 0:06:12

Process ID	Names extracted	Titles extracted	Personal names	Corporate names	MARC to BIBFRAME conversion Progress	Names Enriched	Title Enriched
246	138	267	138	0	269	81 API details	68

Remove All

Return to home

Coming Up

- User registration and authentication.
- User profiles (private, public uploads)
- Download results (zip/tar packages)
- ...

Thank You