OpenRefine for metadata reconciliation

Produced by the Digital Collections Working Group of the Content Creation & Dissemination Team
   Janet Hauck, Whitworth University (chair)
   Anneliese Dehner, Metadata Applications Librarian, Alliance (author of this document)
   Laura Zeigen, Oregon Health & Science University
   Julia Simic, University of Oregon
   Theodore Gerontakos, University of Washington
   Jodi Allison-Bunnell, Alliance (ex officio)

Credit and thanks goes to Christina Harlow, who has created several OpenRefine reconciliation services and documentation for using these services. Much of the content in this document has been gleaned from her wonderful work.

Version 1.0, May 2017

Table of Contents

Overview

Systems-specific limitations

Reconciliation tutorials
   Creators & Contributors
   Genre
   Subject

Enrichment in the Harvester

Additional reconciliation services

Overview of OpenRefine reconciliation methods

Overview

This document is intended to guide Alliance digital content managers through the process of reconciling local metadata and enriching creators, contributors, and genres in the Alliance Harvester. Overviews of OpenRefine reconciliation methods and additional reconciliation services are provided as a gateway to explore local reconciliation beyond the requirements set by the Alliance Dublin Core Best Practices.
Reconciliation logically follows metadata cleanup, and this document is designed to follow OpenRefine for Metadata Cleanup, an introduction to OpenRefine.

**Reconciliation**

Reconciliation, in this case, consists of comparing values in your local metadata with values in an external dataset, and when they match, pulling in text strings from the external data. This is the process of normalizing local text strings to controlled vocabulary text strings. Reconciliation can be done manually--looking up values in LCNAF, LCSH, etc.--but manual reconciliation is time consuming. When using OpenRefine, reconciling local metadata to controlled vocabulary terms can be accomplished in a semi-automated process. This document covers several methods for reconciling metadata in OpenRefine.

**Enrichment**

Enrichment, in this case, consists of comparing values in your local metadata with values in an external dataset, and when they match, pulling in URIs from the external data. This is the process of normalizing local text strings to URIs. Because most DAMs and IRs are not structured to accommodate linked data, enrichment is done in and by the Alliance Harvester. When contributing a set in the Alliance Harvester, Alliance members can choose to enrich creators and contributors to Virtual International Authority File (VIAF) URIs. And they can choose to enrich genre values to Getty Art & Architecture Thesaurus (AAT) URIs. This enriched metadata will be an interoperable product, ready for DPLA and other aggregations.

**Benefits of reconciliation & enrichment**

- Harvested Alliance metadata will be more flexible and interoperable
- The metadata we add to Primo and DPLA will be more uniform, resulting in a better experience for our users
- As a consortium, we will be prepared when linked data becomes the norm

**Systems-specific limitations**

**Digital Commons**

The Digital Commons metadata schema structures Creator data in separate fields: One field for the creator's first name, one for the middle name or middle initial, and one for the last name. Due to this data structure, reconciling creators in Digital Commons to Library of Congress Name Authority File terms is not possible.

**Reconciliation tutorials**

In OpenRefine, reconciliation follows 5 key steps:

1. Picking a column to reconcile
2. Picking a vocabulary to reconcile with
3. Telling OpenRefine about the vocabulary
4. Starting the reconciliation process

5. Reviewing and choosing matches

Four reconciliation services are listed below, along with instructions for their configuration and use. These services are recommended because they are easy to use and they reconcile to vocabularies addressed in the Alliance Dublin Core Best Practices.

There are many other open source reconciliation services, built by developers in the OpenRefine community. For information about additional services, see the Additional reconciliation services and Overview of OpenRefine reconciliation methods sections below.

Creators & Contributors

Two services are recommended for creators and contributors because their results differ slightly due to differences in their matching algorithms. You may prefer to use one or to use them in combination.

LoC Reconciliation Service: Library of Congress Name Authority File (LCNAF)

Created by Christina Harlow. Use this reconciliation service to reconcile creators and contributors to LCNAF preferred labels in http://id.loc.gov/authorities/names. More information about this service can be found on github.

The instructions below will guide you through the process of adding and using this reconciliation service in OpenRefine. They assume you have already installed OpenRefine on your computer, created a project, and opened the project. For help with these steps, please refer to OpenRefine for Metadata Cleanup.

1. If the cells in your Creator column contain multiple values, split the multi-valued cells.

   Screenshots above show (from left to right): a multi-valued cell before splitting, where to access the 'split' function, and cells that have been split.

2. To open reconciliation options, click the drop-down menu icon at the top of the Creator column.
3. Hover over “Reconcile” in the drop-down menu.

4. Select “Start reconciling...” from the flyout menu, and a box will open that looks like the box below. If you have added any reconciliation services/extensions, they will be listed in the outlined box at left.

5. Click the “Add Standard Service...” button, and a box will open like the screenshot below:
6. Enter the URL for the LoC Reconciliation Service service and click the “Add Service” button. LoC Reconciliation Service URL: https://lc-reconcile.herokuapp.com/

There will be a slight delay while the reconciliation service loads. When the service has loaded, the reconciliation dialog box will update to look like the screenshot below:

You only need to add the service one time. When you click “Start Reconciling” in the future, you will see “LoC Reconciliation Service” listed with the other reconciliation services in the outlined box at left.

7. Select your reconciliation options:
   - Click the “LoC“ radio button. There are three types listed in the middle outlined box. “Names” searches LCNAF only, “Subjects” searches LCSH only, and “LoC” searches both. While it may seem counterintuitive, LoC is the most successful choice. But feel free to play around with “Subject” and “Names” to determine the best choice for your needs.
• If you would like to manually review every match, UN-check “Auto-match candidates with high confidence”
• Click the “Start Reconciling” button to initiate the reconciliation process

You will see your OpenRefine project again, and a yellow progress box will display at the top of the browser window. Reconciliation will take a while, depending on how much data you have.

8. Understanding your results. When the process is complete, you will see the closest matches the reconciliation service found for each cell in the Creator column. In the “Facet/Filter” panel at left, you will see “judgment” and “best candidate’s score” facet boxes. Show your data as “rows” to use these facets effectively.
**Matches:** A single dark blue link indicates the text string successfully matched a controlled vocabulary term, and the reconciliation service is confident that this is the correct match. You can click the blue link to review its authority record and verify the match yourself.

**Multi-matches:** Checkboxes and light blue links indicate multiple matches were found for your text string. The matches are ordered by the reconciliation score, highest score first.

**No match:** Checkboxes along with grey text reading “Create new topic” indicate that a no match was for your text string.

9. Manually review the cells with multiple matches. You can click one of these potential matches to review the match’s authority record.

When you find the match you want, click the box with one check (to the left of the match) and only this cell will be updated. When a match has been chosen the checkboxes will disappear and the link color will change to dark blue:
OR when you find the match you want, click the box with two checks (to the left of the match) to update this cell and all other Creator cells containing the same text string. When a match has been chosen the checkboxes will disappear and the link color will change to dark blue:

10. If you would like to keep your original text strings, you can add a column for the new reconciled data. To create a new column for your reconciled data, follow the steps below:

- Click the drop-down menu icon at the top of the column containing reconciled data.
- Hover over “Edit column”.
- Click “Add column based on this column...”, and a GREL box will open like the one below.
● Type a name for your new column in the “New column name” input at the top of the GREL box
● In the “Expression” box, replace value with one of the expressions below. When you do, you will see the the new column’s values in the right column of the “Preview” table:
  ○ To pull in the label only: cell.recon.match.name
  ○ To pull in the URI only: cell.recon.match.id
  ○ To pull in the label and the URI, separated by a pipe (|) (for easier column splitting later): cell.recon.match.name + " | " + cell.recon.match.id
● Click the “OK” button, and you will see your new column of reconciled data.

Now you need to clear the reconciled data and restore your original text strings in the Creator column. To do this, follow the steps below:
- Click the drop-down menu icon at the top of the Creator column.
- Hover over “Reconcile”.
- Hover over “Actions”.
- Click “Clear reconciliation data”, and you will see your original creator values in the Creator column.

11. If you split multi-valued cells before reconciling your data (in step 1), rejoin those cells now.

Conciliator: Library of Congress Name Authority File (LCNAF) / VIAF
Created by Jeff Chiu. Use this reconciliation service to reconcile creators and contributors to LCNAF labels in https://viaf.org/. More information about this service can be found here.

The instructions below will guide you through the process of adding and using this reconciliation service in OpenRefine. They assume you have already installed OpenRefine on your computer, created a project, and opened the project. For help with these steps, please refer to OpenRefine for Metadata Cleanup.

1-5. Begin by following steps 1-5 in the “LoC Reconciliation Service: Library of Congress Name Authority File (LCNAF)” section above.

6. Enter the URL for the Conciliator Service service and click the “Add Service” button.
Conciliator Service URL: http://refine.codefork.com/reconcile/viaf/LC

There will be a slight delay while the reconciliation service loads. When the service has loaded, the reconciliation dialog box will update to look like the screenshot below:
You only need to add the service one time. When you click “Start Reconciling” in the future, you will see “VIAF - LC” listed with the other reconciliation services in the outlined box at left.

7. Select your reconciliation options:
   - Click the “Person” radio button, if it’s not automatically selected
   - If you would like to manually review every match, UN-check “Auto-match candidates with high confidence”
   - Click the “Start Reconciling” button to initiate the reconciliation process.

You will see your OpenRefine project again, and a yellow progress box will display at the top of the browser window. This process will take a while, depending on how much data you have.

8-11. Follow steps 8-11 in the “LoC Reconciliation Service: Library of Congress Name Authority File (LCNAF)” section above to understand and interpret your results, choose matches, and save your reconciled data to a new column (if desired).

Genre

RDF extension for Getty Art & Architecture Thesaurus (AAT)

Use this extension to reconcile genre values to AAT terms in a RDF file. This reconciliation method can be used for any RDF file of data (a local or regional vocabulary, languages, subjects, geospatial terms, etc.). If you can download or create the RDF file, you can use this method.
The instructions below will guide you through the process of adding and using this reconciliation service in OpenRefine. They assume you have already installed OpenRefine on your computer, created a project, and opened the project. For help with these steps, please refer to OpenRefine for Metadata Cleanup.

Download and Install DERI RDF Extension

The DERI RDF extension is required to reconcile values against an RDF file or a SPARQL endpoint. For more information about this extension, see “DERI RDF Extension” in the “Overview of OpenRefine reconciliation methods” section below.

1. Find your version of OpenRefine. The screenshot below shows you how to find it in the OpenRefine UI:

2. Find the location of your OpenRefine installation on your computer by clicking the “Browse workspace directory” link in the bottom left corner of the OpenRefine UI, as shown below:
3. The link should open the folder where your OpenRefine installation is located on your computer. Does your “OpenRefine” folder contain a folder named “extensions”? If not, create a new folder in the “OpenRefine” folder, and name it “extensions”.

4. Keep the “extensions” folder open on your desktop.

5. Quit OpenRefine.

6. Download the DERI RDF extension:
   - If you are using OpenRefine 2.5, [download RDF extension 0.8.0](#)
   - If you are using OpenRefine 2.6 - beta 1, [download RDF extension 0.9.0](#)

7. Move the RDF extension zip file into the “OpenRefine” > “extensions” folder on your computer.

8. Unzip the file.

9. If you downloaded RDF extension 0.9.0
   - Download this [java file](#), and add it to the folder at this location:
     OpenRefine > extensions > rdf-extension (the name of this folder varies with the version) > MOD-INF > lib

10. Restart OpenRefine, and you will now see the “RDF” button in the top right corner of an OpenRefine project:

---

Add the AAT reconciliation service

1. Download [this RDF file](#) to your computer. It contains all AAT preferred English labels and record URIs.

2. Open a project in OpenRefine.
3. Click the “RDF” button.
4. Hover over “Add reconciliation service”.
5. Click “Based on RDF file...” and the “Add file-based reconciliation service” dialog box will open.

![Add file-based reconciliation service dialog box]

6. Type a name for this service in the “Name:” input.
7. Click the radio button to the left of “Upload file:”
8. Click the “Choose File” button to upload the RDF file you downloaded in step 1.
9. Select “RDF/XML” from the “File format:” drop-down menu.
10. Check the “rdfs:label” option in the “Label properties” box.
11. Click “OK” to add the service. Loading the RDF file and adding this service will take as long as 15 minutes. When the service has been added, you will be taken back to your OpenRefine project.
12. Verify that the reconciliation service was added:
● Click the drop-down menu icon of any column in your project.
● Hover over “Reconcile”.
● Click “Start reconciling...” and you will see a reconciliation dialog box like the one below. If the AAT reconciliation service added successfully, you will see it listed in the outlined box at left:

Reconcile genre values to AAT terms
Follow the steps below to reconcile genre text strings in your metadata to AAT terms:

1. If the cells in your Type column contain multiple values, split the multi-valued cells.

   Screenshots above show (from left to right): a multi-valued cell before splitting, where to access the ‘split’ function, and cells that have been split.

2. To open reconciliation options, click the drop-down menu icon at the top of the Type column (or whichever column contains your genre values).
3. Hover over “Reconcile”.

4. Click “Start reconciling...” and you will see a reconciliation dialog box like the one below:

5. Click “AAT” in the list of reconciliation services in the outlined box at left, and when the service loads, you will see reconciliation options. It may take several minutes for the service to load because the RDF file is quite large.

6. When the AAT service loads, select your reconciliation options:
   - If you would like to manually review every match, UN-check “Auto-match candidates with high confidence”
   - Click the “Start Reconciling” button to initiate the reconciliation process.
You will see your OpenRefine project again, and a yellow progress box will display at the top of the browser window. This process will take a while depending on how much data you have.

7. Follow steps 8-11 in the “LoC Reconciliation Service: Library of Congress Name Authority File (LCNAF)” section above to understand and interpret your results, choose matches, and save your reconciled data to a new column (if desired).

**Subject**

RDF extension for Library of Congress Subject Headings (LCSH)

Created by the Free Your Metadata group. Use this reconciliation service to reconcile subjects to LCSH terms via the SPARQL endpoint created by the Free Your Metadata group. More information about this service can be found here.

The instructions below will guide you through the process of adding and using this reconciliation service in OpenRefine. They assume you have already installed OpenRefine on your computer, created a project, and opened the project. For help with these steps, please refer to OpenRefine for Metadata Cleanup.

The DERI RDF Extension is required to reconcile subjects to LCSH terms with this service. Refer to “Download and Install DERI RDF Extension” in the “RDF extension: Getty Art & Architecture Thesaurus (AAT)” section above for instructions on downloading and installing the extension.
Add LCSH reconciliation service

1. Open a project in OpenRefine.

2. Click the “RDF” button.

3. Hover over “Add reconciliation service”.

4. Click “Based on SPARQL endpoint...”, and the “Add SPARQL-based reconciliation service” dialog box will open:

5. Type a name for this service in the “Name:” input.

6. In the “Endpoint URL” input, type: http://sparql.freeyourmetadata.org/

7. In the “Graph URI” input, type: http://id.loc.gov/authorities/subjects
8. Select “Virtuoso” from the “Type:” drop-down menu.

9. Check the “skos:prefLabel” option in the “Label properties” box.

10. Click “OK” to add the service. When the service has been added, you will be taken back to your OpenRefine project.

11. Verify that the reconciliation service was added:

   - Click the drop-down menu icon of any column in your project.
   - Hover over “Reconcile”
   - Click “Start reconciling…” and you will see a reconciliation dialog box like the one below. If the LCSH reconciliation service added successfully, you will see it listed in the outlined box at left:
Reconcile subject values to LCSH terms

Follow the steps below to reconcile subject text strings in your metadata to LCSH terms:

1. If the cells in your Subject column contain multiple values, split the multi-valued cells.

   Screenshots above show (from left to right): a multi-valued cell before splitting, where to access the ‘split’ function, and cells that have been split.

2. Click the drop-down menu icon at the top of your Subject column (or whichever column contains your subject values)

3. Hover over “Reconcile”

4. Click “Start reconciling...” and you will see a reconciliation dialog box like the one below
5. Select “LCSH” from the list of reconciliation services in the outlined box at left. It will take a several minutes to load the service, more if the SPARQL endpoint is busy at the moment.

6. Select your reconciliation options:
   - Select the radio button for skos:Concept
   - If you would like to manually review every match, UN-check “Auto-match candidates with high confidence”
   - Click the “Start Reconciling” button.

You will see your OpenRefine project again, and a yellow progress box will appear at the top of the browser window. This process will take a while depending on how much data you have.
7. Follow steps 8-10 in the “LoC Reconciliation Service: Library of Congress Name Authority File (LCNAF)” section above to understand and interpret your results, choose matches, and save your reconciled data to a new column (if desired).

**Enrichment in the Harvester**

Please refer to Alliance documentation: Contributing a set to the harvester for details about enriching creators, contributors, and genre values in the Alliance Harvester.

**Additional reconciliation services**

There are many open source reconciliation services available, created by developers in the OpenRefine community. Here are two additional services that may be of interest to you. Both of these must be hosted on your computer to work.

*Faceted Application of Subject Terminology (FAST)*

Created by Christina Harlow. This reconciliation service must be locally hosted (on your computer) to work with OpenRefine. Installation and configuration requires comfort with working in a command line interface and basic knowledge of python.

*GeoNames Reconciliation Service*

Created by Christina Harlow. This reconciliation service must be locally hosted (on your computer) to work with OpenRefine. Installation and configuration requires comfort with working in a command line interface and knowledge of python.

**Overview of OpenRefine reconciliation methods**

Three reconciliation methods are presented here because they all have valid use cases and nuances to consider. Each method requires a different level of expertise and each interacts differently with external data sources.

*Add column by fetching URL...*

OpenRefine can use API queries, in the form of a URL, to retrieve data. The URL uses a value in one of your OpenRefine columns to retrieve information from an external dataset by making an HTTP request to an external API and then posting the entire data response in the cell of a new OpenRefine column.
Pros

- Lots of available APIs out there that can be used with this method (see “Example APIs” section below)

Cons

- Data retrieval is very slow.
- This is a two step process, and the second step requires some scripting knowledge. In the first step, data is pulled in from an external dataset. And in the second step, relevant information is extracted from the retrieved data by parsing it with a GREL expression.

Example APIs that work with this method

- [OCLC’s Virtual International Authority File (VIAF) API](#)
- [Library of Congress Subject Headings (LCSH)](#)
- [Faceted Application of Subject Terminology (FAST) Linked Data API](#)
- [Geonames API & MWDL’s experiences using it with OpenRefine](#)
- [Geocoding names and addresses with the Google Maps API](#)

**Standard reconciliation service API**

Reconciliation services allow you to look up your OpenRefine values in an external dataset, match your OpenRefine values to terms in the external dataset, and use the external values in your own data. A reconciliation service is a program independent of OpenRefine. Some services are hosted remotely (on the web) and some must be hosted locally (on your computer). Locally hosted services are faster. A reconciliation service uses a REST API to get data from an external dataset, parse the data, and return the data to OpenRefine

Pros

- Data retrieval is faster than the ‘Add column by fetching URL...’ method described above
- More sophisticated than the ‘Add column by fetching URL...’ method described above. The reconciliation service does the retrieval, parsing, matching, and ranking of results for you.
- Data is retrieved and presented in a user-friendly format.

Cons

- Several reconciliation services must be hosted locally, on your computer. The process of configuring a reconciliation service locally requires working in a command line interface and may require basic knowledge of python.
Example reconciliation services

- **LC Reconcile** created by Christina Harlow
- **Conciliator** created by Jeff Chiu
- **GeoNames Reconciliation Service** created by Christina Harlow
- **Faceted Application of Subject Terminology (FAST)** created by Christina Harlow

**DERI RDF extension**

OpenRefine functionality can be enhanced with extensions, which can be downloaded and installed on your computer. DERI's RDF Extension makes it possible to use a reconciliation service against linked data sources with SPARQL endpoints (SPARQL is a language used to query linked data and a SPARQL endpoint is the URL to which such queries can be sent) or RDF documents.

**Pros**

- Data retrieval is faster than the “Add column by fetching URL...” method described above
- More sophisticated than the “Add column by fetching URL...” method described above. The retrieval, parsing, and matching are done for you.
- Data is retrieved and presented in a user-friendly format.
- Can be used to reconcile against a RDF document (eg. downloaded from the Library of Congress Linked Data Service, downloaded from Getty, or created from your own local vocabulary)
- Exports to RDF serializations (RDF/XML and RDF N-triples at the moment) after you set up an RDF skeleton for mapping your records to nodes. This is not a format that Alliance DAMs and IRs can currently import. But the functionality can be used to create an RDF file of local vocabulary terms.

**Cons**

- No longer actively supported by DERI. This means that some issues you might run into will not necessarily have anyone there who can answer or update the codebase.
- Use depends on the SPARQL server of the external dataset. For example, the Getty SPARQL endpoint does not work with this extension.

**Examples of RDF extensions**

- **Library of Congress Subject Headings (LCSH)** created by the Free Your Metadata group
- Create your own with an RDF file of controlled vocabulary terms.