Web of Science Data Normalization

Overview

The goal of this process is to refine Web of Science data about open access publications from UNC Charlotte affiliated authors so that it can be used for outreach to encourage deposits in Niner Commons. The process focuses on normalizing the open access, department, and author data to make it easier to sort and filter for targeted outreach: Niner Commons Outreach: WoS Citation Data. Following outreach, the data will be transformed using the Metadata crosswalk: Web of Science >>> Niner Commons to create the MODS records for the works that are approved for deposit into Niner Commons.

Pulling the data

The following steps detail how to pull the raw data set from Web of Science, which we will then transform in OpenRefine.

1. In Web of Science, perform a document search. Select "Affiliation" in the field dropdown and then enter "University of North Carolina Charlotte" in the search box.
2. In the Quick Filters section on the left, check "Open Access" to include only open access works.
3. Use the Publication Years section to filter down to only the desired years. (Web of Science only allows for exports of 1,000 records at a time, so some filtering will likely be necessary to allow for export.)
4. Click Export, then choose Excel. Choose the second record option to export the full list of results. Under Record Content, select "Full Record."
5. Open the exported file in Excel and do some preliminary cleaning.

a. These columns should be kept. Columns with a * will be used for creating the metadata later on.
    i. Publication Type*
    ii. Author Full Names*
    iii. Article Title*
    iv. Source Title
    v. Language*
    vi. Document Type*
    vii. Author Keywords*
    viii. Abstract*
    ix. Addresses
    x. Reprint Addresses*
    xi. Email Addresses
    xii. Funding Orgs
    xiii. Funding Text
    xiv. Publisher
    xv. Publication Year
    xvi. DOI*
    xvii. Early Access Date*
    xviii. Open Access Designations
b. These columns may also be helpful for later sorting of data depending on the type of outreach being done:
    i. Times Cited
    ii. Research Areas
c. Move the Open Access Designations column to the front of the spreadsheet.

Removing works already in Niner Commons

These steps will compare the Web of Science spreadsheet to a spreadsheet of works in Niner Commons in Excel to identify any works in the Web of Science data that are already in Niner Commons. The formula will look at the DOI column in the Web of Science spreadsheet and compare it to the DOI column in the Niner Commons data and produce TRUE or FALSE based on if a match is identified. This is done before the Web of Science spreadsheet is loaded into OpenRefine to begin the data cleaning process.

1. In the Niner Commons spreadsheet, find and replace “doi:” with nothing.
2. In the Web of Science spreadsheet, insert a new column called “Match” after the DOI column.
3. Enter the following formula in cell 2 of the Match column: **=NOT(ISERROR(MATCH(W2, 'Niner Commons Data.csv'!$D$2:$D$18, 0)))**
   Note that the exact values in the formula will vary based on the columns used in the spreadsheets and the name of the Niner Commons data file, so you may need to make adjustments to these variables:
    a. **W2** : this should be the first value in the DOI column in the Web of Science data

    b. **Niner Commons Data.csv** : this should be the name of the Niner Commons spreadsheet
    c. **$D$2:$D$18** : this should be the range of the DOI column in the Niner Commons spreadsheet
4. Carry down the formula to the rest of the cells in the Match column.
5. Filter the Match column to only show the “TRUE” values.
6. Delete those rows from the spreadsheet.

Cleaning the data in OpenRefine

“Open Access Designations”

Web of Science compiles detailed open access information for published works, found in the “Open Access Designations” column. While the information is helpful, it proves challenging to sort as many works have multiple designations listed. This process utilizes a hierarchy to prioritize open access designations and create a new column with a single designation for each publication to assist with targeted outreach based on open access type. Copyright issues are a big barrier to entry for including works in an institutional repository. Because of those challenges, this process focuses on gold works, which have the most straightforward open access copyright situation. For this same reason, bronze works were given the least priority due to the murky nature of their licensing.

Complete the following steps to transform the Open Access Designations to make them easier to sort.

1. Remove spaces in the column name, so it reads OpenAccessDesignations. *[Edit column > Rename this column]*
2. Add a column named “Simplified Open Access” based on the OpenAccessDesignations column using this expression: **if(value.contains("gold"), 'Gold', null)** *[Edit column > Add column based on this column]*
3. Facet the Simplified Open Access column by blank. *[Facet > Customized facets > Facet by blank]*
4. Click “true” in the facet to filter down to those with blanks in the Simplified Open Access column. This step is necessary to avoid overwriting data in the column from a previous transform.
5. Transform the Simplified Open Access column using the following expressions in this order. As each expression is performed, the affected rows should drop out as they are no longer blank. *[Edit cells > Transform]*
    a. **if(cells.OpenAccessDesignations.value.contains("hybrid"), 'Hybrid', null)**
    b. **if(cells.OpenAccessDesignations.value.contains("Green Published"), 'Green Published', null)**
    c. **if(cells.OpenAccessDesignations.value.contains("Green Accepted"), 'Green Accepted', null)**
    d. **if(cells.OpenAccessDesignations.value.contains("Green Submitted"), 'Green Submitted', null)**

    e. **if(cells.OpenAccessDesignations.value.contains("Bronze"), 'Bronze', null)**
    f. Remove the facet on the Simplified Open Access column.
6. Remove the facet on the Simplified Open Access column.

"Departments"

Web of Science does not have a column with clearly defined department data. Instead, we can pull this information, when available, from the address column in the spreadsheet. In general, the following data cleaning measures will assign approximately 30% of the records with departmental metadata. This was also the case when the Web of Science data was reconciled against the UNC System salary database, which lists employees and their departments. Remaining Web of Science records will need manual intervention to add departmental metadata. Web of Science has a "Research Areas" field that could be referred to when sorting the data for outreach, but reconciling that to UNCC departments for metadata purposes is not recommended, as some research areas are interdisciplinary and do not map readily to a department. Looking ahead, if a more comprehensive faculty directory becomes available, this process could be updated to reconcile the two data sources and hopefully automatically assign departmental metadata to more records.

Complete the following steps to pull out the academic department data and begin to normalize it.

1. Add a column named "Departments" based on the Reprint Addresses column using this Python/Jython for the expression:

```
import re
pattern = re.compile(r"((Dept|Sch|School|Department).+?),", re.I)
list = []
for i in pattern.findall(value):
    list.append(i[0])
return ";".join(list)
```

   [Edit column > Add column based on this column]

2. Find and replace "Dept" with "Department of". [Edit cells > Replace]
3. Find and replace "Dept" with "Department of". [Edit cells > Replace]
4. Find and replace "Sch" with "School of". *[Edit cells > Replace]*
5. Cluster and edit on the Departments column using the nearest "neighbor method" method with "ppm". *[Edit cells > Cluster and edit]*
6. Reference this document for controlled terms. Note that not all departments in the spreadsheet will be from UNCC.
7. Create a text facet on the Departments column to further clean the departments, referencing the above document. *[Facet > Text facet]*

“Author Names”

Author data from Web of Science also requires attention to normalize minor variations in how the same author is listed in different works and to trim down the number of authors as some publications have 10+ authors listed. Normalizing this data will assist in outreach by making it easier to identify authors with multiple publications.

Complete the following steps to separate out the author data into rows, trim down the number of authors to a maximum of 5 for each record, and normalize data to clean up name variations.

1. Split the `Author Full Names` column into several columns. *[Edit column > Split into several columns]*
    a. Split by separator “; ” (semicolon-space).
    b. Split into 6 columns at most.
    c. Uncheck “Remove this column.” This will preserve the original column in the spreadsheet so the full list of authors is not lost for publications with more than 5 authors.
2. Transpose the new columns into rows. *[Transpose > Transpose cells across columns into rows]*
    a. Select `Author Full Names 1` for the “From Column.”
    b. Select `Author Full Names 5` for the “To Column.” (If you aren’t limiting the number of authors, then select the last `Author Full Names` column.)
    c. Select transpose into one column, named “Clean Author Names.”
3. Trim leading and trailing whitespace in the `Clean Author Names` column. *[Edit cells > Common transforms > Trim leading and trailing whitespace]*
4. Cluster and edit the `Clean Author Names` column to clean up potential variations within how the same author is represented in the spreadsheet. There are various clustering methods that look at different criteria, so running both of these is necessary to catch the majority of variations. *[Edit cells > Cluster and edit]*
    a. First use the “key collision” method with the “fingerprint” keying function.
    b. Then switch to the “nearest neighbor” method with “ppm.” Increase the radius to pull in more potential matches.
5. Delete the `Author Full Names 6` column. *[All > Edit columns > Re-order/remove columns]*

Copying down data

Complete the following steps to copy down the information in a record to the newly created rows for the additional authors. (There is a “fill down” function under “Edit Cells” but it doesn’t recognize records in OpenRefine, so it will continue to carry down the value until it hits a row with a value. That could be problematic if there is a record that has an intentional blank column as it would carry down the value from the previous record, which is why this transform is needed instead.)

1. Transform the `All` column with this expression: **row.record.cells[columnName].value[0]** Change the expression language to GREL. *[All > Transform]*

2. After clicking OK, a screen will appear allowing you to select the columns to transform and the order in which the transform happens.
3. Uncheck the Clean Author Names column.
4. Click and drag the first column down to the bottom so that the transform is carried out on that column last. (The first column in the spreadsheet maintains the record relationships between rows. Once this transform is carried out on that column the records will disappear and the entire spreadsheet will be treated as separate rows, so it is important that this column is changed last.)