Gap Analysis Script

Cherie Turner

6/26/2017

SUMMARY: This script uses the data gathered for a larger gap analysis project on monographs holdings and interlibrary loan requests for monographs to provide some guidance for smaller subsets of the collection. It makes a recommendation for changes to the purchasing level for small sections of the collection based on circulation and ILL requests for items in that section.

The script requires four major inputs:

- Item level monograph data, including circulation
- ILL request data for monographs
- List of all relevant call numbers
- Category descriptions for each relevant call number

The steps used to export and clean the circulation and interlibrary loan data are described in a paper by Bronicki et al [i](1). The circulation data used was exported from Innovative's Sierra, and the ILL data from Illiad. The fields for each of these data sets are described in the Source Data section below. A separate data set containing the list of relevant call numbers and category descriptions must be created for this analysis. Details on that process are also available in the Source Data section of this document.

SOURCE DATA:

Monographs Data: This analysis requires that the monograph circulation data contain only three fields:

- full_call_num – LC call number pulled from the bib and item call number fields
- checkout_total – count of total circulations
- YTD_checkout – count of circulations in the year prior to the data export

The monographs data was saved as a csv file to be read into R, named books.csv.

Some additional fields are likely to be helpful for deeper analysis, or for interpretation:

- title
- publisher
- pub_year – publication year
- cat_date – catalog date; used as a proxy for date added to the collection
- location – location within the library; useful for identifying items located in the stacks vs. course reserves, etc.

Some additional fields were retained from data processing and cleanup in earlier studies. These fields are not necessary for this analysis.

- Bib_recordNum – internal bibliographic record number
- Bib_call_num – LC call number pulled from the bibliographic record (missing from some records)
- Item_call_num – LC call number pulled from the item record (missing from some records)
- LC_Category – LC Class separated out from the full call number (artifact of past analysis)
- LC_subcategory – LC Subclass (e.g. HD) separated out from the full call number (artifact of past analysis)

Interlibrary Loan Data: Similar to the monograph circulation data, only one field from the ILL data was absolutely required for the analysis:

- Call Number Congress / Other – LC Call number items with all other types of call numbers were removed during cleanup

The interlibrary loan data was saved as a csv file, named ILL.csv, to be read into R.

One additional field was used to filter out data prior to this analysis:

- Format – type of item (only books were of interest for this analysis and the past analysis from which the data was obtained)

Several additional fields were useful for deeper analysis and interpretation:

- Title
- Author
- Date 1
- Imprint
- Language

And several fields were retained through data processing which were not relevant to this analysis:

- Lender Full Name – name of the library loaning the item
- Lender State – state of the library loan the item
- OCLC Number
- LC_Classes – LC Class separated out from the Call Number Congress / Other field (artifact of past analysis)
- LC_Subclasses - LC Subclass (e.g. HD) separated out from the Call Number Congress / Other field (artifact of past analysis)
- ISBN
- Request Initiated Date
- Borrower Filled Date
- Library Type – type of library loaning the item

SCRIPT OVERVIEW:

This script works in several major steps:

- Designating the relevant call numbers for your collection of interest
- Adding description files that can attach whatever descriptions are relevant to you to the data
- Adding monographs data
- Creating a call number from the monographs data that can be easily matched with call numbers from the list of relevant call numbers created in the first step
- Sub-setting the monographs data to include only the data for your areas
- Creating a new csv file with only the monographs data for your areas
- Calculating the total number of books, checkouts, and year-to-date checkouts as well as mean, median, and maximum total checkouts for each section of your data
- Adding ILL data
- Creating an easily matched call number from the ILL data
- Sub-setting the ill data to include only data for your areas
- Creating a new csv file with only the ILL data for your areas
- Calculating the total number of ILL requests for books for each section of your data
- Merging the information on the number of items in our collection and checkout information from the monographic data with the information about requests from the ILL data by section
- Making ratio calculations for each section of the data (details on how each calculation was made on pg. 4)
- Creates a new csv file including the summary data for each category and the calculations listed in 14 (see pg. 4 for a description of each variable)
- Creates a plot in pdf form that shows for each category
    - The average use (as a red bar)
    - The standard deviation (as whiskers)
    - The maximum usage (as a point)

SUMMARY DESCRIPTION:

The following fields are included in the csv file exported by the script.

- row.names – factor of the script, not meaningful for analysis and can be deleted
- Description – the descriptions that you designated in your descriptor files
- LC_Subcategory – the LC subcategory for each description category
- Start_LC_Number – the first LC number in the range that is described
- End_LC_Number – the last LC number in the range that is described (NOTE: if the same descriptor is used in multiple locations the LC number ranges may appear to overlap, but this does not affect the calculations)
- Items – total number of items in each description category

- YTD_Checkouts – total number of checkouts in the 1 year period before the data was accessed for each description category
- Checkouts – total number of all-time checkouts for each description category
- Mean_Checkouts – average number of checkouts for each description category
- Median_Checkouts – median number of checkouts for each description category
- Max_Checkouts – maximum number of checkouts for each description category
- Requests – total number of ILL requests in the 1 year period for which the data was collected
- P_Items - (Items in the relevant section)/(Total items in this area of the collection)
- P_YTD_Checkouts – (Number of YTD Checkouts for this section)/(Total Number of YTD Checkouts)
- YTD_Use – (P_YTD_Checkouts)/(P_Items); ratio of actual 1 year use to expected use
- YTD_Rating - described use level based on YTD_Use
    - If YTD_Use <1 YTD_Rating is Underused
    - If YTD_Use >1 YTD_Rating is Overused
- P_Checkouts – (Number of all-time checkouts for this section)/(total number of all-time checkouts)
- Use – (P_Checkouts)/(P_Items); ratio of actual all-time use to expected use
- Rating – described use level based on Use
    - If Use <1 Rating is Underused
    - If Use >1 Rating is Overused
- Usage_Change – YTD_Use – Use; means of tracking how usage may be changing
- Trend – described trend based on Usage_Change
    - If Usage Change > 5% Trend is Increasing Use
    - If Usage Change <5% Trend is Decreasing Use
    - All other cases Trend is Long Term Trend
- P_Requests – (ILL Requests in this section)/(Total Number of ILL requests)
- Borrowing – (P_Requests)/(P_Items); ratio of requests to collection size in that area
- ILL_Rating – described ILL demand based on Borrowing
    - If Borrowing < mean ILL_Rating is Low Demand
    - If Borrowing > mean ILL_Rating is High Demand
    - If Borrowing > mean + SD ILL_Rating is Very High Demand
    - If Borrowing > mean + 2(SD) ILL_Rating is Extremely High Demand
- YTD_Recommendation – recommendation for further action for each description category
    - If YTD_Rating = Overused and ILL_Rating = High Demand, Very High Demand, or Extremely High Demand YTD_Recommendation is Growth Opportunity
    - If YTD_Rating = Underused and ILL_Rating = High Demand, Very High Demand, or Extremely High Demand YTD_Recommendation is Change in Purchasing
    - If YTD_Rating = Overused and ILL_Rating = Low Demand YTD_Recommendation is No Changes
    - If YTD_Rating = Underused and ILL_Rating = Low Demand YTD_Recommendation is Ease Off

- Recommendation
    - If Rating = Overused and ILL_Rating = High Demand, Very High Demand, or Extremely High Demand Recommendation is Growth Opportunity
    - If Rating = Underused and ILL_Rating = High Demand, Very High Demand, or Extremely High Demand Recommendation is Change in Purchasing
    - If Rating = Overused and ILL_Rating = Low Demand Recommendation is No Changes
    - If Rating = Underused and ILL_Rating = Low Demand Recommendation is Ease Off

NOTES:

- For all calculations (including the mean and standard deviation used to calculate the ILL_Rating) each category is compared only to the other categories in your analysis. Analysis comparing to the entire collection is possible but only if descriptors are developed for the entire collection
- There will probably be areas of your collection where the YTD_Recommendation and the Recommendation do not match. It is recommended that you consider both carefully, and if in doubt err on the side of further analysis or on collecting rather than stopping collecting in an area.
- The mean, median, and maximum values for each section are intended to help you identify where an item or several items with high numbers of checkouts that could skew the script's recommendations for that category. When you see either very high maximum checkouts, or a mean and median value that are not close together consider those sections more carefully.

REUSING THE SCRIPT:

If you choose to adapt this script, edit the following areas of the script:

- lib – This variable changes the naming of the resulting csv and pdf files. Please replace "name" with your name.
- sn – These variables designate the LC subcategories that are relevant to this portion of the collection. Copy this row for each LC subcategory of interest, and in each row replace "LC" with a single LC subcategory (ex. QD). Re-label each variable name so that you have s1, s2 . . . sn
- snrange – These variables designate the LC numbers that are relevant to the portion of the collection that you are analyzing. Copy this row for each LC subcategory of interest, and in each row replace "num:num" with the LC call numbers that you care about for the matching sn (ex. if s1 is QD, and I care about QD 1-999, then "s1range<-c(1:999)")
    - If you have multiple distinct ranges within the same LC subcategory, use the same format, separating each range by a comma (ex. "s1range<-c(1:367, 480:499)")
- match – This variable merges the sn and snrange into a vector where each individual call number of interest is listed. For each pair of sn and snrange you should have "paste(sn, snrange)", with each set separated by a comma (ex. "match<-c(paste(s1, s1range), paste(s2, s2range))")
- Sn – These lines of code add the files containing your descriptors to the script.
    - If you choose to use one combined descriptor file then delete this row of code.
    - If you choose to use multiple descriptor files then copy this row for each descriptor file, and change "LCDescriptor.csv" on each row to the appropriate name for your csv descriptor files.
- descriptors – This line of code combines the descriptor files for each section into one data set.
    - If you chose to use one combined descriptor file then change the code to "descriptors<-read.csv("descriptor.csv")" and replace "descriptor.csv" with whatever name you choose for your csv descriptor file.
    - If you chose to use multiple descriptor files ensure that each number is listed in the descriptor script (S1 through Sn) with each entry separated by a comma (ex. "descriptors<-rbind(S1, S2, S3)")

You will likely also need to modify the script to accommodate for differences in the monograph data or interlibrary loan data available, or in the data cleanup conducted.

Prior to running this script you will need to:

- Create descriptions for your call numbers and save them in a csv file. Your file should include 4 columns:
    - LC_Subcategory – the LC subcategory for each individual call number

- LC_Number – each individual LC number (must specifically account for 1, 2, 3, 4 . . ., not provide ranges)
- Match – Use the excel function CONCATENATE(A1, " ", B1) to create a column containing a call number with a space between the subcategory and the number.
- Description – The description that you would like to be displayed for all items of that LC subcategory and number (ex. all QD 1 are General – Chemistry). These descriptions can be repeated across a range, or if helpful across multiple ranges. If you choose to use the same description across multiple ranges do be aware that your final summary csv file will not differentiate between those ranges.

---

[i] Bronicki, Jackie, Irene Ke, Cherie Turner, and Shawn Vaillancourt. 2015. "Gap Analysis by Subject Area of the University of Houston Main Campus Library Collection." *The Serials Librarian* 68 (1-4):230-242. doi: 10.1080/0361526X.2015.1017717.