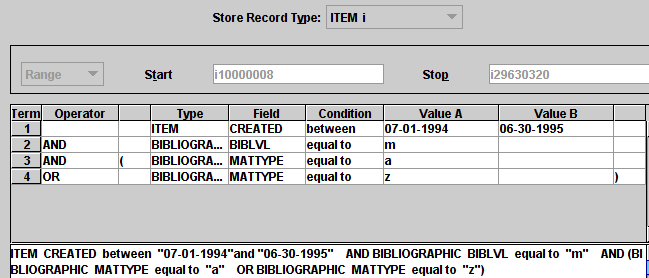
DH Project 2016  
Data Clean Up Steps

Notes:

Millennium search directions file name and locaiton > DH\_PUBLIC > Millennium\_search\_strategy-use-this-one

+++++++++++++++++++++



Searched for books with an order paid date for each fiscal year

Limited to print books

Our e-book records were not reliable.

Downloaded bibliographic information

Cleaned up and normalized the data

Did not use non-English books due to diacritics issues

Nor gov pubs.

Put the data into an Access database

Queried the database using variations of terror

Subject headings generated more results than titles

Subject headings reduce the number of terms and normalized the terms

We will continue gathering the data

We are making the data available to the library

Useful for outreach, collection evaluation and analysis and ?





1. Freeze top row
2. Insert a new column A for “Fiscal Year”
   1. Ex: 1995-1996 = FY 96
3. Remove text from OCLC # leaving only the number
   1. Steps: Copy and paste column to new sheet
   2. Click “text to columns” delimited by “
4. Convert MATTYPE
   1. a = print book
   2. z = ebook
5. Convert BIBLVL
   1. m=books
   2. s=serials
6. Convert ORD TYPE
   1. a=autoship
   2. f=firm
   3. s=standing order
   4. p=periodical
   5. m=membership
   6. r=fill in
   7. i=internet resource
7. Change file name from “to clean” to “clean”

Separate the title from the author.

Separate the unusable records (multiple OCLC numbers.)

Separate the gov pubs.

Separate the Non-English language titles.

In the end there will be 4 files for each of our master files: English, non-English, gov pubs, and not used.

Cleaning up the files:

Look at the records and take those that have multiple OCLC numbers and paste them into the not-used worksheet.

copy the title column to a new worksheet

name that worksheet "title-author-separation"

highlight the title column and delimit on the "/"

use =TRIM(a2) to remove leading spaces in the author and title fields.

Copy and paste special the title column back into the final worksheet.

Copy and paste special the author column back into the final worksheet.

to remove gov docs

filter the call number column to text that contains ":"

Color those cells a color

Sort the call number by color and cut and paste those titles to a worksheet titled “gov docs.”

To remove Non-English:

Filter the Lang column to everything but “Eng” and “blank”

Color the cells

Sort the lang column by color and cut and paste those titles to a worksheet titled “Non-English.”