DASH! Data analysis

At the beginning of the DASH! phase, we reviewed subject heading and author heading information. An important step in exploring linked data integration is assessing where there may be points of intersection between the catalog and other linked data sources and how extensive these intersections may be. Here is a link to a very informal jamboard trying to combine pieces of information. This document provides a more targeted analysis.

Subject browse index

Description	Number	Solr query or filter
Authorized headings	1,720,035	authority=true
Authorized headings which are used directly for items	773,829	authority=true, mainEntry=true
Types and related counts for authorized headings	"Personal Name": 665,346 "Topical Term": 447,011, "Corporate Name":327,014 "Geographic Name":179,784 "Work": 88,640 "Event": 6,868 "Genre/Form Term": 5,372 "Chronological Term":0	authority=true, facet.field=headingTypeDesc
Types and related counts for authorized headings where headings are the main entry	"Topical Term":271,105 "Personal Name":246,717 "Corporate Name":124,621 "Geographic Name":98,274 "Work":29,505 "Genre/Form Term":1,976 "Event":1,631 "Chronological Term":0	authority=true, mainEntiry=true, facet.field=headingTypeDesc,

Author browse index

Description	Number	Solr query or filter
Authorized headings	4,999,524	authority=true
Authorized headings which are used directly for items	2,597,883	authority=true, mainEntry=true
Types and related counts for authorized headings	"Personal Name": 4,015,649 "Corporate Name": 851,940 "Event": 131,422 "Work": 247 "Topical Term": 175 "Geographic Name": 91	authority=true, facet.field=headingTypeDesc
Types and related counts for authorized headings where headings are the main entry	"Personal Name": 2,183,578 "Corporate Name":353,640 "Event":60,380 "Topical Term":160 "Work":73 "Geographic Name":52	authority=true, mainEntiry=true, facet.field=headingTypeDesc,

In addition, we wanted to review how many catalog records may relate to a temporal or geographic subject heading.

Description	Number and filters
Catalog records with geographic subject heading	3,220,876 (using field subject_geo_filing)
Catalog records with temporal subject heading	959,900 (using field subject_era_filing)

(* Note the numbers above were retrieved 02/03/2022)

For our data sources, we wished to employ LCSH, LCNAF, Wikidata and PeriodO data that matched LCSH headings. The following results were obtained by running queries against U of Iowa's Fuseki server which uses data downloads from LCSH.

- LCSH temporal components
  - 17,561 subject headings with temporal component
  - 11,553 distinct temporal components (i.e. of type <http://www.loc.gov/mads/rdf/v1#Temporal>)
    - 34 distinct temporal components which are URIs and not blank nodes
- LCSH geographic components
  - 87,542 subject headings with geographic component
  - 114,410 distinct geographic components (i.e. of type <http://www.loc.gov/mads/rdf/v1#Geographic>)
    - 48,677 distinct geographic components which are URIs and not blank nodes

Querying wikidata, we find the following matches to LC identifiers:

- 1,340,603 distinct Wikidata entities which use an LOC identifier. (We used the property P244).
  - 51,864 Wikidata entities pointing to an LOC identifier starting with "sh" (i.e. subject heading)
  - 1,289,158 Wikidata entities pointing to an LOC identifier starting with "n" (i.e. name headings)
- 1,342,872 distinct LOC identifiers designated for Wikidata entities (i.e. this many unique identifiers are found in the object position for P244
  - 51,824 LOC identifiers which start with "sh" i.e. subject headings
  - 1,290,993 LOC identifiers which start with "n" i.e. name headings

(*Note wikidata results were retrieved 2/3/2022)

We also used LCSH mappings from PeriodO which contain 1478 total mappings (as of 2/3/2022).