DASH! Subject Headings index (for experimental prototype)

To support the subject view in the prototype, which included information from LCSH, Wikidata, and PeriodO, we built a Solr index to enable faster access of this information. During LD4P3, we had used the Fuseki server provided by U of Iowa (thanks Dave Eichmann!) which contained the LCSH dataset to retrieve LCSH information. The server has 488,367 LCSH headings. Querying against the server, we obtained a TSV file with the URI and label of these headings and then created an index with this information. Possibly due to some errors with the script, our final index has a total of 421,220 entries where each Solr record represents a different LCSH URI.

Below, we provide an overview of the number of fields corresponding to different data sources in our index, and then provide examples of each of these types of data below.

Description	***Number***
Total number of records	421,220
Records with components (i.e. subheadings of any kind)	157,690
Records with LCSH geographic components	7,710
Records with LCSH temporal components (i.e. either a starting date or stopping date retrieved from LCSH)	6,034
Records with either a PeriodO start or stop date	1,474
Records with PeriodO spatial information	1,476
Records with Wikidata URIs for the subject heading	42,060
Records with Wikidata URIs for LCSH geographic components in the subject heading	6,277

Minimal entry: LCSH URI and Label

Here is the JSON for a minimal LCSH entry in this index:


```
{
        "id":"http:__id.loc.gov_authorities_subjects_sh99014295",
        "uri_s":"http://id.loc.gov/authorities/subjects/sh99014295",
        "label_s":"Rugrats (Fictitious characters)",
        "label_t":["Rugrats (Fictitious characters)"],
        "_version_":1692415613796876288
}
```


Our script copied the label into two fields, “label_s” which allows for exact string matching and “label_t” which allows for partial text matching against the label. The URI is stored in the “uri_s” field.

LCSH Geographic Components

To this index, we also added component information for subject headings, identifying geographic and temporal components separately. We retrieved this information by executing additional queries against the Fuseki server to identify URIs for components with subject headings. Below is an example of a subject heading “Cathedrals–Europe” that has two subdivisions or components: “Cathedrals” and “Europe”. URIs and labels for both are stored in the “components_json_s” field. Because “Europe” corresponds to a geographic component (i.e. identified of Geographic type), we have also extracted the URI into the geo_uri_ss field and the label into the geo_label_ss field.


```
{
        "id":"http:__id.loc.gov_authorities_subjects_sh2009118527",
        "uri_s":"http://id.loc.gov/authorities/subjects/sh2009118527",
        "label_s":"Cathedrals--Europe",
        "label_t":["Cathedrals--Europe"],
        "components_json_s":"[{\"uri\":\"http://id.loc.gov/authorities/subjects/sh85045631\",\"label\":\"
Europe\"},{\"uri\":\"http:   //id.loc.gov/authorities/subjects/sh85021018\",\"label\":\"Cathedrals\"}]",
        "geo_uri_ss":["http://id.loc.gov/authorities/subjects/sh85045631"],
        "geo_wd_ss":["http://www.wikidata.org/entity/Q46"],
        "geo_label_ss":["Europe"],
        "_version_":1692415662490648576
}
```

LCSH Temporal Components

Similar to our process for retrieving geographic information, we also extracted temporal components from LCSH headings to include in the index. We identified the labels representing centuries and hardcoded a list mapping those labels to particular start and end dates. For example, we mapped the LCSH URI <http://id.loc.gov/authorities/subjects/sh2002012476> with the label “20th century” to the start date 1900 and end date of 1999. The example below shows the temporal component included in the “components_json_s” field. The start and end years are represented in the “temp_start_i” and “temp_stop_i” fields respectively.


```
{
        "id":"http:__id.loc.gov_authorities_subjects_sh2009115915",
        "uri_s":"http://id.loc.gov/authorities/subjects/sh2009115915",
        "label_s":"Art, Romanian--20th century",
        "label_t":["Art, Romanian--20th century"],
        "components_json_s":"[{\"uri\":\"http://id.loc.gov/authorities/subjects/sh85007875\",\"label\":\"Art,
Romanian\"},{\"uri\":\"http://id.loc.gov/authorities/subjects/sh2002012476\",\"label\":\"20th century\"}]",
        "temp_start_i":1900,
        "temp_stop_i":1999,
        "_version_":1692415662483308549
}
```


PeriodO information

As we noted, we took the LCSH info mapped in PeriodO to retrieve temporal and geographic information for subject headings and corresponding Wikidata URIs for geographic places. We downloaded this dataset and then ran scripts to parse the data and add information to the index.

In the example below, “Burr conspiracy, 1805-1807” maps to a period_o entity with identifier "p06c6g399wq". This identifier could be used to generate the associated PeriodO link http://n2t.net/ark:/99152/p06c6g399wq .The PeriodO data gives us a start and stop date which we store in the “periodo_start_i” and “periodo_stop_i” fields respectively. PeriodO also gives us a Wikidata URI and label for spatial coverage for this subject heading, which we save in the “spatial_coverage_ss” and “spatial_coverage_label_ss” fields respectively.


```
{
        "id":"http:__id.loc.gov_authorities_subjects_sh85018171",
        "uri_s":"http://id.loc.gov/authorities/subjects/sh85018171",
        "label_s":"Burr Conspiracy, 1805-1807",
        "label_t":["Burr Conspiracy, 1805-1807"],
        "periodo_s":"p06c6g399wq",
        "periodo_start_i":1805,
        "periodo_stop_i":1807,
        "spatial_coverage_ss":["http://www.wikidata.org/entity/Q30"],
        "spatial_coverage_label_ss":["United States"],
        "wikidata_uri_s":"http://www.wikidata.org/entity/Q2994776",
        "_version_":1692415625426632710
}
```


Wikidata URIs

Our scripts for populating this index also included queries for retrieving Wikidata URIs for the main subject heading as well as for LCSH geographic components or subdivisions. In addition, we added Wikidata URIs related to spatial information extracted from PeriodO.

Our PeriodO example we included above shows two examples of the inclusion of Wikidata URIs.We have highlighted the fields below:

```
{
    "id":"http:__id.loc.gov_authorities_subjects_sh85018171",
    "uri_s":"http://id.loc.gov/authorities/subjects/sh85018171",
    "label_s":"Burr Conspiracy, 1805-1807",
    "label_t":["Burr Conspiracy, 1805-1807"],
    "periodo_s":"p06c6g399wq",
    "periodo_start_i":1805,
    "periodo_stop_i":1807,
    "spatial_coverage_ss":["http://www.wikidata.org/entity/Q30"],
    "spatial_coverage_label_ss":["United States"],
    "wikidata_uri_s":"http://www.wikidata.org/entity/Q2994776",
    "_version_":1692415625426632710
}
```


The subject heading "Burr Conspiracy, 1805-1807" has the URI <http://id.loc.gov/authorities/subjects/sh85018171> which relates to the Wikidata URI <http://www.wikidata.org/entity/Q2994776>. The Solr record above uses the field “wikidata_uri_s” to store the Wikidata URIs related to the main subject heading represented in that record. PeriodO information provides the label, “United States” and Wikidata URI, <http://www.wikidata.org/entity/Q30> , for the related location. The Solr record stores this PeriodO Wikidata URI in the “spatial_coverage_ss” field.


```
{
    "id":"http:__id.loc.gov_authorities_subjects_sh2009118527",
    "uri_s":"http://id.loc.gov/authorities/subjects/sh2009118527",
    "label_s":"Cathedrals--Europe",
    "label_t":["Cathedrals--Europe"],    "components_json_s":"[{\"uri\":\"http://id.loc.gov/authorities/subjects/sh85045631\",\"label\":\"Europe\"},{\"uri\":\"http://id.loc.gov/authorities/subjects/sh85021018\",\"label\":\"Cathedrals\"}]",
    "geo_uri_ss":["http://id.loc.gov/authorities/subjects/sh85045631"],
    "geo_wd_ss":["http://www.wikidata.org/entity/Q46"],
    "geo_label_ss":["Europe"],
    "_version_":1692415662490648576
}
```


This example, which we had included to show geographic components retrieved from LCSH, also includes Wikidata URIs for those geographic components. For the LCSH heading “Cathedrals–Europe”, “Europe” is identified by the LCSH URI <http://id.loc.gov/authorities/subjects/sh85045631>. Querying Wikidata for a related URI for this heading gives us the URI <http://www.wikidata.org/entity/Q46> which the Solr record above stores in the “geo_wd_ss” field.