Proj 1 - MySQL Assignment | Knowledge Discovery

Susan Burtner

THE CONCEPT (abridged version, taken from the MAT 259 course website**)**: In any database, there lies hidden knowledge. What does a database contain, and what can MySQL queries reveal? Your first assignment is to find something of interest based on your own cultural / knowledge interests. Here are some options:

1) Topics of Cultural Interest

2) **The Database Organizational Structure** <- *what I'll be looking at*

3) Data Analytics Query Methods

GETTING THE DATA:

- Use the MySQL Workbench to write a query by which to retrieve the data from the SPL database
- Use the **spl_2016 database** which gets updated daily.

DO THE ASSIGNMENT:

Do a report on your research, and propose what to visualize based on time, or volume of activity. Possibly make a comparison between book(s) and movie(s) and soundtracks (cd).

Once you have all the material - click on "POST REPLY" to this link and add your info to complete the assignment.

Map Use and Musings at the Seattle Public Library - 2016

I'm interested in knowing how maps are catalogued at the Seattle Public Library (SPL) in 2016. Maps create an interesting challenge for library and database organizers, becuase they exist beyond just the medium of text, but as images or collections of vector-based features. My research questions try to explore how the library organizes maps, and the structure by which they are made accessible.

Research Questions relating to Organizational Structure:

RQ1: How many maps are available at SPL?

Why am I asking this? It would be helpful to first know how many maps are available at SPL before knowing which ones are checked out, and before performing any analyses of cultural interest. This will be helpful for later visualizations that require normalization.

```
SELECT COUNT(barcode)
FROM spl_2016.inraw
WHERE itemtype LIKE '%map';
```

> *287*

Since the `barcode` value is distinct, we know that there are 287 entries in the database that are designated as "maps."

RQ2: How many map collections does SPL have and what are they?

* **Why am I asking this? * The map collections can be an indication of the *coverage* of the globe by subject area. Collections with more maps may imply increased interest in that particular area, or perhaps, a greater diversity of map representation. For example, areas with more "things" will have more maps related to each of those "things," such hiking, camping, animal activity, etc.

```
SELECT COUNT(DISTINCT collcode)
FROM spl_2016.inraw
WHERE itemtype LIKE '%map';
```

> *8*

There are 8 different types of map collections at SPL. What are they?

```
SELECT DISTINCT collcode
FROM spl_2016.inraw
WHERE itemtype LIKE '%map';
```

collcode
cadocpt
camap
camapr
camus

collcode
canf
caref
naatlr
nanf

Follow-up question... what do these collection designations mean?

We know from this link on the course website that **collcode**:

> *This is a string of characters that encodes several data for each item, including the physical home (aka branch), collection type, and collection name.*

But it's unclear what the names of the collections actually mean. We may have to do additional exploraiton on SPL's website. I'm assuming that "map" is just that, and "ref" is just that as well, but otherwise, I couldn't find much information on the collection metadata, and since the current SPL website is for 2018, I didn't much digging to see what these retroactive designations meant.

RQ3: Do maps have a Dewey class? And if so what are they?

* **Why am I asking this? * We know that maps are *generally* non-fiction, however, they aren't *read* like texts are. If the maps do have Dewey class values, what are they?

```
SELECT inraw.bibNumber, deweyClass.deweyClass, inraw.subj
FROM spl_2016.inraw, spl_2016.deweyClass
WHERE inraw.bibNumber = deweyClass.bibNumber AND
    inraw.itemtype LIKE '%map'
ORDER BY deweyClass DESC;
```

bibNumber	deweyClass	subj
1848209	979	NULL
1757400	912	NULL
1757400	912	NULL
1693625	911	NULL
1693625	911	NULL
1693625	911	NULL
1693625	911	NULL
1693625	911	NULL
1693625	911	NULL

bibNumber	deweyClass	subj
558499	910	NULL
447232	**784**	NULL
2659183	355	NULL
3031147	230	NULL
2010269	224	NULL
2694116	**188**	NULL
1603624	177	NULL
1663926	**173**	NULL

When we grab just the distinct Dewey classes, we get:

> *979, 912, 911, 910, 784, 355, 230, 224, 188, 177, 173*

If we cross reference these classes with the Dewey classification csv given on the course website, we see that the maps fall under these subjects:

Dewey class	Description
173	**Ethics of family relationships**
177	Ethics of social relations
188	**Stoic philosophy**
224	Prophetic books of Old Testament
230	Christianity
355	Military science
784	**Instruments & Instrumental ensembles & their music**
910	Geography & travel
911	Historical geography
912	Graphic representations of surface of earth and of extraterrestrial worlds
979	Great Basin & Pacific Slope region of United States

Follow-up question (and sanity check)... is this right? Let's pick a few `bibNumbers` (in bold) and check.

```
SELECT deweyClass.deweyClass, deweyClass.bibNumber, title.title, subject.subject
FROM spl_2016.deweyClass, spl_2016.title, spl_2016.subject, spl_2016.inraw
WHERE inraw.itemtype LIKE '%map' AND
    deweyClass.bibNumber = title.bibNumber AND
    title.bibNumber = subject.bibNumber AND
```

```
    subject.bibNumber = inraw.bibNumber AND
    deweyClass.bibNumber IN ('1663926', '2694116', '447232')
ORDER BY deweyClass.deweyClass ASC;
```

Dewey class	bibNumber	title	subject
173	1663926	Ouachita National Forest pocket guide	Ouachita National Forest Ark and Okla
188	2694116	Shasta Trinity National Forest California	Outdoor recreation California Shasta National Forest Maps
188	2694116	Shasta Trinity National Forest California	Outdoor recreation California Trinity National Forest Maps
188	2694116	Shasta Trinity National Forest California	Shasta National Forest Calif Maps
188	2694116	Shasta Trinity National Forest California	Trinity National Forest Calif Maps
784	447232	Sing a song of holidays and seasons home neighborhood and community	Childrens songs

Not sure why that last one has the subject "Children's songs," so maybe this is an error. Regardless, these entries do indeed seem to be maps!

RQ4: How many of these maps were left 'Uncategorized'?

* ***Why am I asking this?*** * It's apparent from the previous research questions that maps seem like they would be hard to organize. I am curious if most of the maps are physical or digital copies for example, and if so, how they are organized both in storage and from publically accessible locations in the library. RQ1 in particular raises an interesting thought. I used `barcode` for that query, which implies these are physical maps. But are there digital maps as well? It's not clear we would see this, as this database has tables like `inraw` and `outraw`, which implies that this database contains transactions for physical objects, but perhaps not digital entities (like a digitized map.) Regardless, I would like to see how many maps have no `callNumber`, so I modify my first query. The following query helps to see what proportion of maps are easily or (not so easily) categorized.

```
SELECT COUNT(barcode)
FROM spl_2016.inraw
WHERE itemtype LIKE '%map' AND
    callNumber = 'UNCAT';
```

> *32*

Thirty two out of 287 maps have no `callNumber`, which is essentially the "address" of the item. But how would one find these in the library?

RQ5: What subjects do the maps have?

* *Why am I asking this?* ** RQ3 makes me wonder what the available maps are even about. We can first figure out how many distinct subjects there are, but it would also be cool to know what are the top 20 subjects.

```
SELECT COUNT(DISTINCT subject.subject)
FROM spl_2016.subject, spl_2016.inraw
WHERE inraw.itemtype LIKE '%map' AND
    subject.bibNumber = inraw.bibNumber;
```

> *277*

There 277 different subjects for which the maps are about. What are the top 20 of these?

```
SELECT subject.subject, COUNT(subject.subject) AS total
FROM spl_2016.subject, spl_2016.inraw
WHERE inraw.itemtype LIKE '%map' AND
    subject.bibNumber = inraw.bibNumber
GROUP BY subject.subject
ORDER BY total DESC
LIMIT 20;
```

rank	subject	total
1	| 75	
2	Mount Baker Snoqualmie National Forest Wash Maps	21
3	Maps	18
4	Topographic maps	15
5	Outdoor recreation Washington State Mount Baker Snoqualmie National Forest Maps	10
6	Endangered species California	9
7	Endangered species Northwest Pacific	9
8	Spotted owl	9
9	Wildlife management California	9
10	Mount Baker National Forest Wash Maps	9
11	Wildlife management Northwest Pacific	9
12	Outdoor recreation Washington State Mount Baker National Forest Maps	9
13	Outdoor recreation Washington State Snoqualmie National Forest Maps	9
14	Snoqualmie National Forest Wash Maps	9

rank	subject	total
15	Mount Baker Wilderness Wash Maps	7
16	Noisy Diobsud Wilderness Wash Maps	7
17	Willamette National Forest Or Maps	7
18	United States Historical geography Maps	6
19	Pasayten Wilderness Wash Maps	5
20	Gifford Pinchot National Forest Wash Maps	5

It's interesting that the top "subject" is no subject at all! And most of the rest of the maps are on Washington state or national forests, though the "Topographic maps" could be about anything.

RQ6: Which regions of the world are covered by the given maps?

* ***Why am I asking this?*** What is absent from the database is just as important as what's in it. The creation of maps also reflect ongoing changes in the political and administrative boundaries of the world. While boundaries may seem stable, they obviously weren't always so, and there are still many regions of the world in which cartographic boundaries are unclear or disputed, and this may either spur more map creation or revision.

We can see what's **not** being included in the collection of SPL's maps by looking at RQ4 and revisiting the results from RQ3. When I modify the query in RQ4 to include the Dewey class, I am reminded that MOST map listings do not have a Dewey class (and the top 20 certainly don't.) However, by revisiting the results from RQ3 and comparing them to our list of Dewey classifications, we can get a cursory view of which subjects are not being represented by the library.

The below table shows just some of the Dewey classes that were **not** included when RQ3 was conducted.

Dewey class	Description
914	Geography of & travel in Europe
915	Geography of & travel in Asia
916	Geography of & travel in Africa
917	Geography of & travel in North America
918	Geography of & travel in South America
919	Geography of & travel in *Australasia, Pacific Ocean islands, Atlantic Ocean islands, Arctic islands, Antarctica, & on extraterrestrial worlds

And basically the rest of the 900's could be encompassed within a map classification.

Also, there are no maps about extraterrestrial worlds? That's too bad...

Discussion & Conclusion

Finding geographic information is tricky. A lot of geographic information systems (GIS) research is being conducted in geographic information retrieval and search. While it's not as easy as it seems, visualization is increasingly being used to overcome some of the challenges of finding, exploring and analyzing geographic data. Most physical and natural processes transcend political boundaries, and yet, this is where most data collection takes place. Even so, most maps that are available exist at a sub-administrative scale with a certain subject in mind, such as recreation (hiking in Washington) or tracking animal populations (mapping endangered species). While the 2016 SPL cataloguing of maps seems decent, there is room for improvement (an opportunity for GIS practitioners!)

In future work I would like to start looking more at the **cultural interest** of the available maps. I could compare what is in the database to what is actually being checked out. I would also love to explore which physical maps are being looked at versus digital copies of maps online. Furthermore, I think **word embeddings** offer a great opporunity to group the textual aspect of maps in a meaningful way, and see if this can somehow help improve the process of geographic information retrieval.