Kitty Currier
MAT 259
20 January 2014
Revised 27 February 2014

Project 1: Item number vs. bar code

Project Description:

How do the item number and bar code relate to physical resources at the Seattle Public Library?

The *barcode* field only appears in two tables of the *spl2* database: *inraw* and *outraw*. Tables derived from these (e.g. *activity*, *callnum*, *collection*) include the item number but not the bar code as columns, suggesting that the bar code is redundant information. In this project I posed queries to investigate whether, in fact, the bar code and item number are both unique to individual items.

Queries/explanations:

To investigate this I asked several questions (all queried on *inraw*):

1. How many unique entries are in the *itemNumber* column?

   SQL:

   ```
   SELECT COUNT(DISTINCT itemNumber) FROM inraw
   ```

   Result: 3,480,694

   Duration: 366.602 sec.

2. How many unique entries are in the *barcode* column?

   SQL:

   ```
   SELECT COUNT(DISTINCT barcode) FROM inraw
   ```

   Result: 3,430,056

   Duration: (forgot to record the first time, and subsequently the duration was 0.000 sec.)

The difference between these results indicates that there are 50,638 more unique item numbers than bar codes. (I checked that for every item number there is a bar code, and vice versa.) This suggests that a single bar code may be associated with more than one item number. To check:

3. What is the greatest number of item numbers associated with a single bar code?

   SQL:

   ```
   SELECT barcode, COUNT(itemNumber) as numItems
     FROM
       (SELECT DISTINCT itemNumber, barcode FROM inraw LIMIT 1000000) AS
   q1
     GROUP BY barcode
     ORDER BY numItems DESC
   ```

   Result: (To avoid a timeout error I had to limit this query to searching only the first distinct 1,000,000 records, so the result is a sample. Only the first 5 rows are shown below.)

barcode	numItems
0010050278208	10
0010050277176	10
0010050068351	10
0010050275808	10
0010050064491	10
…	…

Duration: 59.421 sec.

This query uses a subquery to create a table (q1) with only the columns *itemNumber* and *barcode* from *inraw*, a step that in similar queries appeared to reduce processing time. Then it counts the number of distinct item numbers associated with each barcode and presents the results in descending order, listing the barcodes that are associated with the greatest number of item numbers at the top.

This limited subset of *inraw* data shows that some bar codes are associated with at least 10 unique item numbers.

I checked the reverse using the same logic—are there instances where a single item number is associated with more than one bar code?

4. What is the greatest number of bar codes associated with a single item number?

   SQL:

```
SELECT itemNumber, COUNT(DISTINCT barcode) as numBarcodes
  FROM
    (SELECT itemnumber, barcode FROM inraw) AS a1
  GROUP BY itemNumber
  ORDER BY numBarcodes DESC
```

   Result (first 5 rows of 3,480,694→this agrees with query #1):

itemNumber	numBarcodes
2422818	5
2441067	4
2452665	4
2194598	4
2310583	4
…	…

   All item numbers are associated with 5 or fewer bar codes.

   (The results of this query are saved in the accompanying file, **currier_proj1_query4.csv**.)

   Duration: 369.832 sec.

To summarize, an item number may be associated with up to 5 bar codes, and a bar code may be associated with 10 or more item numbers in some cases.

This is strange—how frequently are item numbers and bar codes found in a 1:2 or 1:many relationship?

5. How many item numbers are associated with more than one barcode?

   SQL:

```
SELECT
  SUM(CASE WHEN numBarcodes = 5 THEN 1 ELSE 0 END) AS 5barcodes,
  SUM(CASE WHEN numBarcodes = 4 THEN 1 ELSE 0 END) AS 4barcodes,
  SUM(CASE WHEN numBarcodes = 3 THEN 1 ELSE 0 END) AS 3barcodes,
  SUM(CASE WHEN numBarcodes = 2 THEN 1 ELSE 0 END) AS 2barcodes,
  SUM(CASE WHEN numBarcodes = 1 THEN 1 ELSE 0 END) AS 1barcode
  FROM
   (SELECT itemNumber, COUNT(DISTINCT barcode) as numBarcodes
     FROM
         (SELECT itemNumber, barcode FROM inraw) AS q1
     GROUP BY itemNumber) AS q2
```

   Result:

5barcodes	4barcodes	3barcodes	2barcodes	1barcode
1	25	719	26156	3453793

   Summing the first four columns, 26,900 item numbers are associated with >1 barcode, or < 1% of all unique item numbers.

   Duration: 338.491 sec.

This query begins with a subquery to create a table (q1) with only the columns *itemNumber* and *barcode*, a step that appears to reduce processing time. Next, a subquery acts on q1 to count the number of distinct bar codes associated with each item number. Finally, the query counts and returns the number of item numbers associated with 5, 4, 3, 2, and 1 bar code.

And the reverse, using the same logic:

6. How many bar codes are associated with more than one item number?

   SQL:

```
SELECT
  SUM(CASE WHEN numItems = 10 THEN 1 ELSE 0 END) AS 10items,
  SUM(CASE WHEN numItems = 9 THEN 1 ELSE 0 END) AS 9items,
  SUM(CASE WHEN numItems = 8 THEN 1 ELSE 0 END) AS 8items,
  SUM(CASE WHEN numItems = 7 THEN 1 ELSE 0 END) AS 7items,
  SUM(CASE WHEN numItems = 6 THEN 1 ELSE 0 END) AS 6items,
  SUM(CASE WHEN numItems = 5 THEN 1 ELSE 0 END) AS 5items,
  SUM(CASE WHEN numItems = 4 THEN 1 ELSE 0 END) AS 4items,
  SUM(CASE WHEN numItems = 3 THEN 1 ELSE 0 END) AS 3items,
  SUM(CASE WHEN numItems = 2 THEN 1 ELSE 0 END) AS 2items
  FROM
   (SELECT barcode, COUNT(DISTINCT itemNumber) as numItems
     FROM
         (SELECT itemNumber, barcode FROM inraw LIMIT 10000000) AS q1
     GROUP BY barcode) AS q2
```

   Result: (Again, I limited this search to a sample to avoid timeout.)

10items	9items	8items	7items	6items	5items	4items	3items	2items
228	165	98	54	32	24	99	207	304

   A small number of bar codes are associated with >1 item numbers.

Duration: 507.659 sec.

Analysis:

The item number associated with the highest number of barcodes corresponds to a Juv media CD of *Harry Potter and the Order of the Phoenix* that was checked out and in 135 times between 2006 and 2013. The barcode associated with this item number changes sequentially over time, suggesting that perhaps the physical bar code sticker had to be replaced several times. This pattern does not hold true, however, for other item numbers.

One of the bar codes associated with many item numbers appears to be used for ILL materials. As Karl explained, this is probably a common bar code that is scanned every time an item from a foreign library is checked out/in.