Data collection, analysis and aggregation at VCC Library and Learning Centre

James Fournie
Coordinator, Library Systems and Technical Services
jfournie@vcc.ca
Some of the things we do at the library

• Lend books, media, devices, etc.
• Provide electronic resources
• Provide space
• Teach workshops/other instruction
• Help people find things
Questions we ask to improve service

• How many people come to the library?
• When is the library the busiest?
• What items are people borrowing the most?
• What electronic resources do people use?
• How many books did we buy this year?
Trying to answer questions: LLCMG

**Student Voices Subcommittee**
- Feedback from people
- Marketing/engagement
- Surveys

**Statistics Subcommittee**
- Mining data
- Deciding on new data to collect
- Regular reports
The many-headed hydra of data sources
Integrated Library System (Sierra)

Circulation statistics
(Reports – CSV/Excel Output)

Booking statistics
(MySQL Database, Reports)

Acquisitions – ordering, receiving, etc.
(PostgreSQL database)

Bibliographic data – information about the books in our collection
(MARC standard)
Vendor data

• Journal Articles/Indexes: ProQuest, EBSCO, JSTOR, etc
• Counting Online Usage of Networked Electronic Resources (COUNTER) Code of Practice
• Non-COUNTER
• Also MARC data
<table>
<thead>
<tr>
<th>Database</th>
<th>Publisher</th>
<th>Platform</th>
<th>User Activity</th>
<th>Reporting Period Total</th>
<th>Jan-2018</th>
<th>Feb-2018</th>
<th>Mar-2018</th>
<th>Apr-2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Search Complete</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td>2018-01-01 to 2018-04-30</td>
<td>10861</td>
<td>2739</td>
<td>3007</td>
<td>3558</td>
</tr>
<tr>
<td>Academic Search Complete</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>33493</td>
<td>9843</td>
<td>8584</td>
<td>9103</td>
</tr>
<tr>
<td>Academic Search Complete</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Result Clicks</td>
<td></td>
<td>19048</td>
<td>3646</td>
<td>4360</td>
<td>6200</td>
</tr>
<tr>
<td>Academic Search Complete</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Record Views</td>
<td></td>
<td>11857</td>
<td>2206</td>
<td>2680</td>
<td>3660</td>
</tr>
<tr>
<td>Academic Search Complete</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td></td>
<td>642</td>
<td>146</td>
<td>197</td>
<td>167</td>
</tr>
<tr>
<td>AHFS Consumer Medication Info</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>10665</td>
<td>0</td>
<td>0</td>
<td>4766</td>
</tr>
<tr>
<td>AHFS Consumer Medication Info</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Result Clicks</td>
<td></td>
<td>18</td>
<td>0</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>AHFS Consumer Medication Info</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Record Views</td>
<td></td>
<td>16</td>
<td>0</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Air University Library Index to Military Pe</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td></td>
<td>29170</td>
<td>8536</td>
<td>6838</td>
<td>7986</td>
</tr>
<tr>
<td>Air University Library Index to Military Pe</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>902</td>
<td>199</td>
<td>287</td>
<td>232</td>
</tr>
<tr>
<td>Air University Library Index to Military Pe</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Result Clicks</td>
<td></td>
<td>29625</td>
<td>8636</td>
<td>6938</td>
<td>8133</td>
</tr>
<tr>
<td>Alt HealthWatch</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td></td>
<td>137</td>
<td>39</td>
<td>40</td>
<td>26</td>
</tr>
<tr>
<td>Alt HealthWatch</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>60</td>
<td>17</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Alt HealthWatch</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Result Clicks</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alt HealthWatch</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Record Views</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aphasiology Archive</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td></td>
<td>29181</td>
<td>8539</td>
<td>6833</td>
<td>7992</td>
</tr>
<tr>
<td>Aphasiology Archive</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>846</td>
<td>176</td>
<td>290</td>
<td>220</td>
</tr>
<tr>
<td>Aphasiology Archive</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Result Clicks</td>
<td></td>
<td>29734</td>
<td>8654</td>
<td>6976</td>
<td>8179</td>
</tr>
<tr>
<td>Aphasiology Archive</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Record Views</td>
<td></td>
<td>594</td>
<td>73</td>
<td>118</td>
<td>205</td>
</tr>
<tr>
<td>Applied Science &amp; Technology Source</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Regular Searches</td>
<td></td>
<td>376</td>
<td>63</td>
<td>119</td>
<td>163</td>
</tr>
<tr>
<td>Applied Science &amp; Technology Source</td>
<td>EBSCO</td>
<td>EBSCOhost</td>
<td>Searches-federated and automatic</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Self-collected Data (Oracle)

• Interlibrary Loan statistics
• Desk Question Tally Form
• Gate count tally form
### Interlibrary Loan Statistics

- **Status**: Filled
- **Delivery Method**: Pickup
- **Patron Type**: Student
- **Num. of Approaches**: 3

### Public Services Statistics Form

<table>
<thead>
<tr>
<th>Category</th>
<th>Mode</th>
<th>FAQ (optional)</th>
<th>Duration:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directional</td>
<td>In Person</td>
<td>myVCC problems</td>
<td>optional</td>
</tr>
<tr>
<td>Reference</td>
<td>Telephone</td>
<td>Can I print here?</td>
<td></td>
</tr>
<tr>
<td>Technical</td>
<td>E-mail</td>
<td>Copyright</td>
<td></td>
</tr>
<tr>
<td>A/V Equipment</td>
<td>Virtual Ref.</td>
<td>Wireless problems</td>
<td></td>
</tr>
<tr>
<td>Ready Reference</td>
<td>1-1 Appointment In Person</td>
<td>Citation</td>
<td></td>
</tr>
<tr>
<td>Referred</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Records Management</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Note**: Desk - BC
- **Key**: to return
# GateCount Update Form

Please use this form with either Firefox or Chrome

## Downtown

<table>
<thead>
<tr>
<th>Time</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
<th>Date</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>6am</td>
<td>2,204</td>
<td>2,204</td>
<td>18</td>
<td>2,204</td>
<td>2,204</td>
<td>2,204</td>
<td>25</td>
<td>2,204</td>
<td>41</td>
<td>2,204</td>
<td>46</td>
<td>2,204</td>
<td>50</td>
</tr>
<tr>
<td>7am</td>
<td>2,204</td>
<td>2,204</td>
<td>18</td>
<td>2,204</td>
<td>2,204</td>
<td>2,204</td>
<td>25</td>
<td>2,204</td>
<td>41</td>
<td>2,204</td>
<td>46</td>
<td>2,204</td>
<td>50</td>
</tr>
<tr>
<td>8am</td>
<td>2,204</td>
<td>2,204</td>
<td>18</td>
<td>2,204</td>
<td>2,204</td>
<td>2,204</td>
<td>25</td>
<td>2,204</td>
<td>41</td>
<td>2,204</td>
<td>46</td>
<td>2,204</td>
<td>50</td>
</tr>
<tr>
<td>9am</td>
<td>2,204</td>
<td>2,204</td>
<td>18</td>
<td>2,204</td>
<td>2,204</td>
<td>2,204</td>
<td>25</td>
<td>2,204</td>
<td>41</td>
<td>2,204</td>
<td>46</td>
<td>2,204</td>
<td>50</td>
</tr>
<tr>
<td>10am</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11am</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12pm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Update Counters**

## Totals

- **Total for May 23, 2018** at DTN: 493
- **Total for May 2018** at DTN: 6,213
  - *NOT including after hours traffic*
- **Total for 2018** at DTN: 65,937
  - *NOT including after hours traffic*
- **Total for fiscal year 2018 / 2019** at DTN: 16,909
  - *NOT including after hours traffic*
Other data

• Banner data for Learning Centre (CSV files)
• Workstation logs (flat log file)
• EZProxy (off-campus access)
• Website analytics
• And more!
VCC Library Data

Data Inputs

• Oracle Database
• MySQL Database
• PostgreSQL Database
• CSV / Excel
• MARC bibliographic data
• COUNTER reports
• Banner
• Other
How do we harness this tidal wave of data?
SELECT T1.bib_level, T1.bib_mttype, count(T2.record_num) FROM (SELECT bib_view.id, user_defined_bcode1_muser.name as bib_level, user_defined_bcode2_muser.name as bib_mttype FROM sierra_view.bib_view, sierra_view.user_defined_bcode1_muser, sierra_view.user_defined_bcode2_muser WHERE bib_view.bcode1 = user_defined_bcode1_muser.code AND bib_view.bcode2 = user_defined_bcode2_muser.code ) T1, (SELECT item_view.id, item_view.record_num FROM sierra_view.item_view WHERE (item_view.location_code LIKE '%sta' OR item_view.location_code LIKE '%vid' OR ...) ) T2
<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>485</td>
<td>403</td>
<td>564</td>
</tr>
<tr>
<td>Feb</td>
<td>319</td>
<td>518</td>
<td>266</td>
</tr>
<tr>
<td>Mar</td>
<td>189</td>
<td>422</td>
<td>356</td>
</tr>
<tr>
<td>Apr</td>
<td>429</td>
<td>631</td>
<td>434</td>
</tr>
<tr>
<td>May</td>
<td>201</td>
<td>734</td>
<td>106</td>
</tr>
<tr>
<td>Jun</td>
<td>246</td>
<td>370</td>
<td>127</td>
</tr>
<tr>
<td>Jul</td>
<td>172</td>
<td>22</td>
<td>586</td>
</tr>
<tr>
<td>Aug</td>
<td>258</td>
<td>97</td>
<td>75</td>
</tr>
<tr>
<td>Sep</td>
<td>273</td>
<td>453</td>
<td>397</td>
</tr>
<tr>
<td>Oct</td>
<td>417</td>
<td>305</td>
<td>313</td>
</tr>
<tr>
<td>Nov</td>
<td>771</td>
<td>181</td>
<td>234</td>
</tr>
<tr>
<td>Dec</td>
<td>446</td>
<td>110</td>
<td>276</td>
</tr>
<tr>
<td>YTD</td>
<td>4206</td>
<td>4246</td>
<td>3734</td>
</tr>
<tr>
<td>% Change</td>
<td>0</td>
<td>1%</td>
<td>-12%</td>
</tr>
</tbody>
</table>
Some problems

- Lots of manual steps
- Lots of room for error
- "Excel Magic" --
  - You must know where to paste the data
  - Must know how to handle outliers
  - Etc.
Reports/output

• Monthly Reports
• Quarterly/Term data
• Annual Reports
• BC Council of Post-Secondary Library Directors (CPSLD)
• Decision-making data for service desk
• Decision-making data for purchasing resources
• Decision-making data for discarding/cancelling resources
• Paper/email/spreadsheets
VCC Library Data

Data Inputs
• Oracle Database
• MySQL Database
• PostgreSQL Database
• CSV / Excel
• MARC bibliographic data
• COUNTER reports
• Banner
• Other

Data Outputs
• Spreadsheets
• Data warehouse
• Emails
• Graphs
• PDF of graphs
• Interactive graphs
• When: Regularly and ad hoc
1. Explore data with code
2. Evaluate data explorations
3. Convert explorations into recurring reports

Automation!
Automation: Python

- Writing code is repeatable, vs manual operations
- Python is easy to learn
- Python is widely used in scientific community
- Python has modules for wacky library standards (MARC, COUNTER)
- Python has modules for data analytics (pandas, NumPy)
- Python is free
Exploration: Jupyter Notebooks

• Use Python, R or other programming
• Shows thought process with notes
• Can be shared, built-upon
• Increasingly used in academic research
• Used by Callysto, CanCode initiative (Grade 5-12, BC, AB)
Bar Plot Example

Bokeh’s core display model relies on composing graphical primitives which are bound to data series. This is similar in spirit to Protovis and D3, and different than most other Python plotting libraries.

A slightly more sophisticated example demonstrates this idea.

Bokeh ships with a small set of interesting “sample data” in the bokeh.sampledata package. We’ll load up some historical automobile mileage data, which is returned as a Pandas DataFrame:

```python
In [3]: from bokeh.sampledata.autocmp import autocmp

grouped = autocmp.groupby("yr")
mpg = grouped["mpg"]
evg = mpg.mean()
std = mpg.std()
years = list(grouped.groups.keys())

american = autocmp[autocmp["origin"]==1]

japanese = autocmp[autocmp["origin"]==3]
```

For each year, we want to plot the distribution of MPG within that year:

```python
In [4]: p = figure(title="MPG by year (Japan and US)"

p.vbar(x=years, bottom=mpg.std, top=mpg+std, width=0.6,
      fill_alpha=0.2, line_color=None, legend="MPG ± stddev")

p.circle(xs=japanese["yr"], ys=japanese["mpg"], size=10, alpha=0.5,
        color="red", legend="Japanese")

p.triangle(xs=american["yr"], ys=american["mpg"], size=10, alpha=0.5,
            color="blue", legend="American")

p.legend.location = "top_left"

show(p)
```

MPG by Year (Japan and US)
Chunks of Python code Can be repurposed into pipeline
Luigi

• Developed at Spotify
• Open-source
• Allows you to define dependencies between tasks
• Will run all the dependencies for a job
Data Pipelines
Data pipelines
Data pipelines
18 Graphs
Automated!
Next Steps

- Clean up rough edges
- Documentation
- Interactive graphs
- Excel/Email output
- Improve manual forms
The End

- [https://math.stackexchange.com/questions/1024992/fight-against-the-hydra-graph-theory](https://math.stackexchange.com/questions/1024992/fight-against-the-hydra-graph-theory)
- [https://www.flickr.com/photos/intersectionconsulting/7537238368](https://www.flickr.com/photos/intersectionconsulting/7537238368)