Outshine with the Superfine Frankenstein Pipeline at Timberline

Visualizing Cost Per Use in Power BI

05/2024
Introductions

• Lydia Harlan
  Budget Analyst for Collections, Discovery, and Digital Strategy

• Kristin Buxton
  Head of Science Liaisons

• Gabriele Hayden
  Head of Data Services
Roadmap

• Why we built a cost per use model in Power BI
• Our iterative process
• Views from our dashboard
• Ongoing work
• Discussion and Questions
Discussion

• How do you gather usage data?
• Do you do it in house or through a vendor?
• How do you share it?
• Are any of you using Power BI or another data visualization tool?
• If so, what information are you gathering and presenting?
What Did We Do and Why Did We Do It?

Why: No comprehensive cost per use data, manual usage stats collection, and little analysis

What: Could we use Power BI to create a cost per use model?

Who: “Change will not come if we wait for some other person or some other time. We are the ones we've been waiting for. We are the change that we seek.” (Thanks Obama.)

When: May 2022 – May 2023 but also some pre-work starting November 2021.
A Technical Adventure

We began this project not knowing what technology we needed. We each brought existing skills in Excel, plus

- Lydia: SQL queries in our ILS reporting module (Alma Analytics), access to previous Cost Per Use spreadsheets
- Kristin: former software engineer (mostly C++), a "customer" for the finished project
- Gabriele: basic Power BI and Python

We each contributed based on our skill set.
Along the way we learned

- How APIs work, the SUSHI API protocol
  - reading articles and documentation

- Power BI languages Power Query (M) and DAX
  - 16 hours of formal training, reading documentation, troubleshooting

Plus individual skill development in Python (Gabriele) and Alma Analytics (Lydia)
How

• Get cost
• Get use
• Join, transform data into Cost Per Use
• Turn static data loads into API queries
• Make it look pretty
• Share it
• Use it
Wait, how?

It’s all connected…somehow.

Disparate documentation.

Documentation is for developers, not us.

Multiple dependencies.
Approaches to the Pipeline
Technical Definitions

- COUNTER 5
- SUSHI
- API
Technical Definitions

• **COUNTER 5** – Standard for reporting usage statistics for electronic resources

• **SUSHI** – An API for retrieving COUNTER 5 usage reports

• **API** – Application programming interface. A way to exchange data between computers/servers programmatically
Proof of concept in Excel
Proof of concept in Power BI
Query SUSHI from Power BI?

1. Publisher 1
2. Publisher 2
3. Publisher 3
4. Publisher 4

- Alma
  - Cost report

- SUSHI API
- SUSHI API
- SUSHI API
- SUSHI API

- Power BI

- Report
No! Query SUSHI from Alma
Solution! Use Python for Power BI query.
Our Dashboard
Power Query Interface showing M Code

cost_data

```m
let Source = Python.Execute("# packages#(1f)#(1f)#(1f)#(1f)import requests # for API calls#(1f)#(1f)#(1f)#(1f)import xml.etree.ElementTree as ET # to par
df1 = Source[{Name="df"]}{Value},
#"Changed Type" = Table.TransformColumnTypes(df1,{{"0", Int64.Type}, {"ISSN", type text}, {"Title (Normalized)", type text},
#"Removed Columns" = Table.RemoveColumns(#"Changed Type"),{"0"}, "REPORT_SUM(Transaction Expenditure Amount BY Title (Normalized)
#"Renamed Columns" = Table.RenameColumns(#"Removed Columns"),{{"Title (Normalized)", "Title Normalized"}}),
#"Replace Beginning the with nothing" = Table.ReplaceValue(#"Renamed Columns", each [Title Normalized], each if Text.StartsWith(#"Replace Beginning the with nothing", each [Title Normalized], each if Text.EndsWith(#"Replace Beginning the with nothing", each [Title Normalized], each if Text.EndsWith("",""), Replacer.ReplaceText,{"ISSN"})),
#"Split Column by Delimiter" = Table.SplitColumn(#"Replaced Value", "ISSN", Splitter.SplitTextByDelimiter("; ", QuoteStyle.C
#"Changed Type1" = Table.TransformColumnTypes(#"Split Column by Delimiter"),{{"ISSN.1", type text}, {"ISSN.2", type text}}),
#"Renamed Columns1" = Table.RenameColumns(#"Changed Type1"),{{"ISSN.1", "ISSN1"}, {"ISSN.2", "ISSN2"}}),
#"Added Conditional Column" = Table.AddColumn(#"Renamed Columns1", "Fund Code Display", each if Text.Contains([Fund Code], "
#"Filtered Rows" = Table.SelectRows(#"Added Conditional Column", each true)
```

Model view shows relationships between tables
# Extract the column names from the data

# Get the schema namespace (would it be better to pull this from the xml itself?)
ns = {'xsd': 'http://www.w3.org/2001/XMLSchema',
      'saw-sql': 'urn:saw-sql'}

# Get the column headers (human-readable names) and
# names (the XML tag names) from the schema as a list
column_names = [elem.attrib['{urn:saw-sql}columnHeading']
                for elem in root.findall('./xsd:element', ns)]

column_xml = [elem.attrib['name'] # elem.attrib doesn't have 'saw-sql' b/c not in xml
              for elem in root.findall('./xsd:element', ns)]
What’s in here, how does it work

Power BI data pipeline matches cost and use data on ISSN and Title using M (Power Query) and DAX languages.

• Lots of details to get clean data.

Calculate cost per use in a fuzzy way.

• most expensive year and highest use year
• rounding
• holes in the data
Consider ILL

<table>
<thead>
<tr>
<th>Normalized Title</th>
<th>ConsiderILL</th>
<th>Max Yearly Requests</th>
<th>Max Cost</th>
<th>Cost Beyond ILL Threshold</th>
<th>Fund</th>
</tr>
</thead>
<tbody>
<tr>
<td>20th century mild</td>
<td>True</td>
<td>11</td>
<td>$313.50</td>
<td>$0</td>
<td>Other</td>
</tr>
<tr>
<td>administration society</td>
<td>True</td>
<td>59</td>
<td>$2,420.79</td>
<td>$1,365</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>advance functional materials</td>
<td>True</td>
<td>7</td>
<td>$10,843.37</td>
<td>$0</td>
<td>Science</td>
</tr>
<tr>
<td>advance materials</td>
<td>True</td>
<td>4</td>
<td>$8,505.30</td>
<td>$0</td>
<td>Science</td>
</tr>
<tr>
<td>advance in geometry</td>
<td>True</td>
<td>4</td>
<td>$799.40</td>
<td>$0</td>
<td>Science</td>
</tr>
<tr>
<td>african affairs</td>
<td>True</td>
<td>32</td>
<td>$861.05</td>
<td>$420</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>african technological studies</td>
<td>True</td>
<td>2</td>
<td>$863.77</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>agriculture</td>
<td>True</td>
<td>14</td>
<td>$452.10</td>
<td>$0</td>
<td>Design</td>
</tr>
<tr>
<td>agriculture and human values</td>
<td>True</td>
<td>10</td>
<td>$265.43</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>american journal of health behavior</td>
<td>True</td>
<td>2</td>
<td>$1,277.89</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>american journal of physical medicine rehabilitation</td>
<td>True</td>
<td>15</td>
<td>$305.31</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>american journal of nursing</td>
<td>True</td>
<td>152</td>
<td>$10,572.72</td>
<td>$1,960</td>
<td>Science</td>
</tr>
<tr>
<td>american journal of public administration</td>
<td>True</td>
<td>7</td>
<td>$313.50</td>
<td>$0</td>
<td>Other</td>
</tr>
<tr>
<td>american rivet</td>
<td>True</td>
<td>67</td>
<td>$1,645.03</td>
<td>$1,545</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>animal rights</td>
<td>True</td>
<td>11</td>
<td>$141.60</td>
<td>$0</td>
<td>Other</td>
</tr>
<tr>
<td>animals of mathematics</td>
<td>True</td>
<td>8</td>
<td>$689.70</td>
<td>$0</td>
<td>Science</td>
</tr>
<tr>
<td>animal studies of applied linguistics</td>
<td>True</td>
<td>4</td>
<td>$4,314.47</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>anthropological studies</td>
<td>True</td>
<td>40</td>
<td>$359.33</td>
<td>$0</td>
<td>Other</td>
</tr>
<tr>
<td>anthropology</td>
<td>True</td>
<td>15</td>
<td>$104.59</td>
<td>$0</td>
<td>Social Sciences</td>
</tr>
<tr>
<td>archeology and anthropology</td>
<td>True</td>
<td>7</td>
<td>$668.90</td>
<td>$0</td>
<td>Other</td>
</tr>
<tr>
<td>archeology</td>
<td>True</td>
<td>5</td>
<td>$120.48</td>
<td>$0</td>
<td>Other</td>
</tr>
</tbody>
</table>

Total: 7763 requests for $725,293.77 with a total cost of $141,365.
Consider ILL Formula

This is our formula for Cost Beyond ILL Threshold, assuming documents through Reprints Desk cost $35 and we can request an item 20 times before having to pay for copyright:

Cost Beyond ILL Threshold = IF('cost_data'[Electronic Use MAX]-20<0,0,
('cost_data'[Electronic Use MAX]-20)*35)

The Consider ILL column is a simple True / False statement:

ConsiderILL = IF(AND('cost_data'[Cost Beyond ILL Threshold] < 'cost_data'[Transaction Expenditure Amount],cost_data[Electronic Use MAX] > 0 ), TRUE, FALSE)
Maintenance, Replication, and Future Work
Ongoing maintenance

• Every month, after Alma performs the SUSHI harvest, load the Power BI model and refresh the data.

• Update vendor usernames and passwords, SUSHI credentials, API keys, and base URLs as needed.

• Respond to changes in platform functionality that can break report.

• Add new vendors as they enable COUNTER 5 reporting.
Skills & tech you need to replicate this project

- Alma, Alma Analytics, COUNTER 5, APIs, SUSHI, Python, and Power BI.
- Power BI Desktop requires a Windows PC.
- Basic familiarity with programming with SQL in Alma Analytics, with M and DAX in Power BI, and with Python.
- Infinite patience.
- Friends.

See our article coming soon!
Our code: github.com/uodataservices/cost_per_use
Future work

• Add more report types beyond tr_j1.
• Database evaluation is particularly tricky. Open to ideas.
• How to manage non-COUNTER compliant data?
Discussion

What insights/data points/criteria do you use to make data driven decisions?

What are we missing in our current dashboard?

Where could you see a data dashboard helping conversations at your organization?
Questions?
(and cat tax)

Lydia Harlan, lharlan@uoregon.edu
Kristin Buxton, kbuxton@uoregon.edu
Gabriele Hayden, ghayden@uoregon.edu
References


Thank you!
(and more cats)