Visualization of Historical Prices for Journal Subscriptions

Mark Lerret, Anne Weir, Jordan Price, Chang-Han Chang, Anantharaman Janakiraman

Introduction

The Scholarly Publishing and Academic Resources Coalition (SPARC) works to enable the open sharing of sale data of research results and educational materials. Publishers who sell these materials have access to all of the sale data, but the institutions that buy the materials do not. This market asymmetry leaves institutions at a disadvantage when determining the market price of the materials that the institution wants to purchase. Increased access to this data allows institutions to gain access to these materials at a more competitive and affordable price, which will democratize access to knowledge, accelerate discovery, and increase the return on our investment in research and education. As part of this mission, SPARC has developed the Big Deal Knowledge Bases, which details what thousands of peer institutions have paid for journal subscription packages.

Currently SPARC has made this data available to the public for free [1]. The data is in a table format and users can filter out certain criteria and view the data that is relevant to them. SPARC wants this data to be transformed into a more visually appealing output that is user friendly, that will allow buyers to make more informed purchases.

1 Insight Needs

The SPARC team provided several insight needs that are of interest. Other insight needs were gleaned by this project's team when studying and exploring the data.

- What is the average sale price of a given journal collection?
- What does the distribution of sales prices for a given journal collection look like?
- How do different parameters such as FTE (institutions of certain sizes) and Carnegie Basic Classification affect pricing?
- What publishers charge the most/least for collections?
- How are prices changing over the years?
- Are prices different for different geographic regions?
- What is a fair price for a given collection for a given institution?
- How have economic downturns (e.g., Covid, 2008 recession) affected pricing?

1.1 Stakeholder Analysis

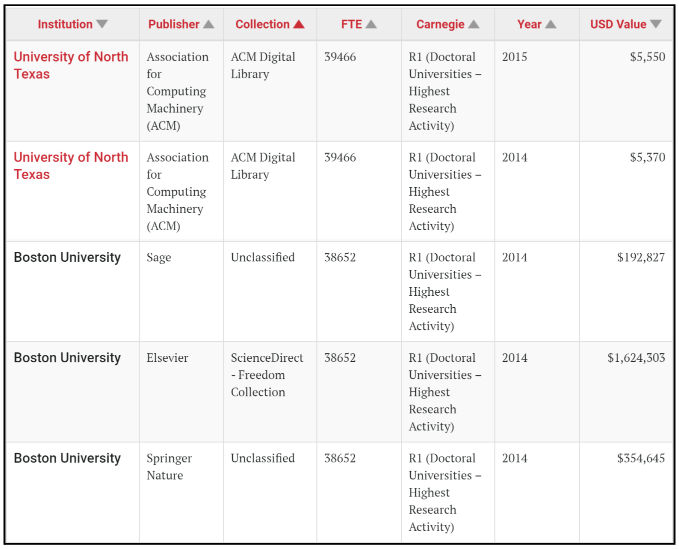
Our primary stakeholder is the SPARC organization because it is our client, but the primary focus of our project was to enable price discovery of research and other academic resources for buyers and sellers of these resources. This market is constantly evolving, and insights are wide ranging and constantly changing. Therefore, our project did not focus on conducting analysis for users of SPARC resources, but instead, it focused on providing interactive and dynamic visualizations that allow end users to analyze the information themselves. We are confident that interacting with our dashboards and analyzing our visualizations will help buyers and sellers of academic resources answer the insight needs listed above. Answering these insight needs will ultimately allow buyers and sellers to negotiate more fair prices in the same way Kelley Blue Book does so in the market for used cars.

2 Data Acquisition

SPARC collects and compiles the data set manually. Our project team acquired the data from the SPARC website [1], because it is available to the public for free.

2.1 Description of Data

Dataset:



Institution ▼	Publisher ▲	Collection ▲	FTE ▲	Carnegie ▲	Year ▲	USD Value ▼
University of North Texas	Association for Computing Machinery (ACM)	ACM Digital Library	39466	R1 (Doctoral Universities – Highest Research Activity)	2015	$5,550
University of North Texas	Association for Computing Machinery (ACM)	ACM Digital Library	39466	R1 (Doctoral Universities – Highest Research Activity)	2014	$5,370
Boston University	Sage	Unclassified	38652	R1 (Doctoral Universities – Highest Research Activity)	2014	$192,827
Boston University	Elsevier	ScienceDirect - Freedom Collection	38652	R1 (Doctoral Universities – Highest Research Activity)	2014	$1,624,303
Boston University	Springer Nature	Unclassified	38652	R1 (Doctoral Universities – Highest Research Activity)	2014	$354,645

Figure 1: Dataset

Figure 1 shows a sample of the data set. This data set has 7 attributes and 11,949 records,as of April 2021. Each record is an individual transaction where there is a buyer and a seller of some academic resource. This can be summarized by saying each record indicates a buyer, a seller, a specific product, a price, and two characteristics of the buying institution. The following is a description of each attribute:

1. *Institution:* This attribute identifies the buying institution. These institutions are comprised of universities, research organizations, and other users of research pricing data.
2. *Publisher*: This attribute identifies the selling institution that created the academic resource being transacted upon.
3. *Collection*: This attribute identifies the academic resource being transacted upon.
4. *FTE*: This attribute is a proxy for how large the buying institution is. FTE stands for full time equivalency, which counts the number of full-time personnel at the institution where each part-time staff member or student counts as some fraction between 0 and 1.
5. *Carnegie*: This attribute identifies the Carnegie Classification of the buying institution. Carnegie Classification, is a classification of colleges and universities in the United States. Hence the classification is not applicable for other countries. The categories include Doctoral Universities, Master's Colleges, Research Universities etc.

6. *Year*: This attribute identifies the year in which the transaction occurred.
7. *USD Value*: This attribute identifies the amount for which the academic resource was transacted for in US dollars.

3 Visualizations

In order to meet our client's insight needs, three dashboards were created using Tableau. The dashboards can be found and used by following this link: https://public.tableau.com/profile/danny.chang#!/vizhome/Sparc_withData/Home?publish=yes

Figure 2 is an image of the home page of the dashboards:

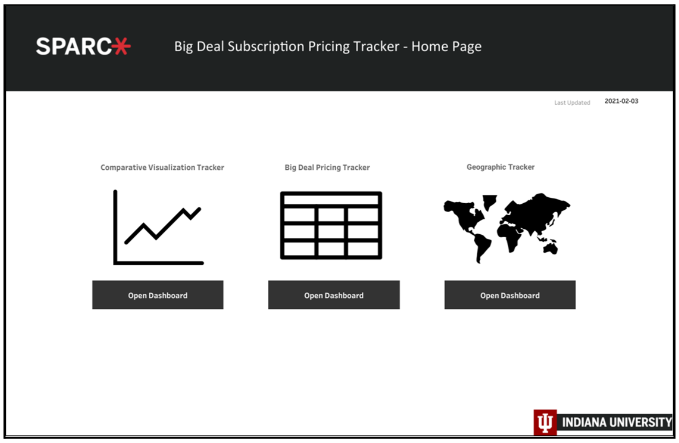


Figure 2: Visualization Home Page

The home page displays links to the Comparative Visualization Tracker, the Big Deal Pricing Tracker, and the Geographic Tracker dashboards. Each of the dashboards visualizes the academic resource pricing data differently, but has similar filtering capabilities. Below, each individual dashboard is described in greater detail.

3.1 Comparative Visualization Tracker

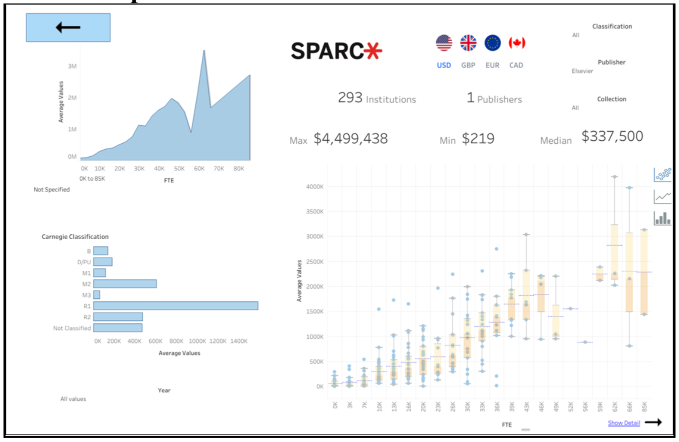


Figure 3: Comparative Visualization Tracker

Figure 3 shows the Comparative Visualization Tracker. This dashboard is intended to provide insights into the changes of the distribution of prices for a particular academic resource or collection of resources over time, for different ranges of buyer FTE, and for different buyer Carnegie Classifications.

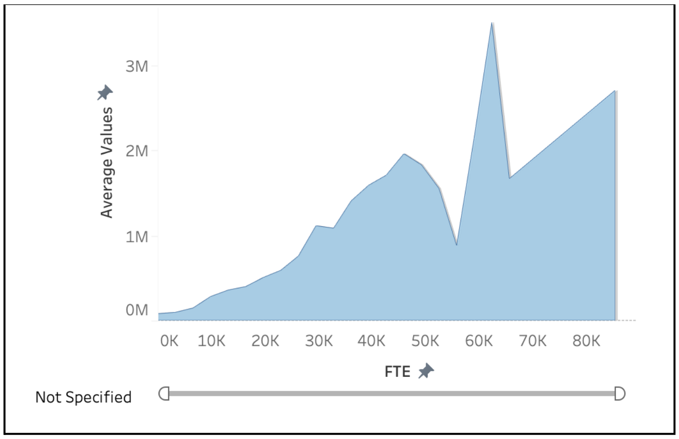


Figure 4: FTE Distribution

Figure 4 is an image of the interactive filter and line graph for the FTE attribute. The user can change the FTE range for which to analyze deals using the slider below the line graph, and toggle the inclusion of institutions with unspecified FTE with the "Not Specified" button. When changed, the filter is applied to the entire Comparative Visualization Tracker dashboard. The line graph shows how the average deal size in US Dollars changes with FTE, and allows the user to see the range of FTE and range of average deal size that are selected with the filter. The graph is also subject to other constraints set in other filters on this dashboard. Much like the filter for FTE, the filter for Carnegie Classification is interactive and shows average deal size for values of that attribute. A sample image of this visualization is provided below (Figure 5).

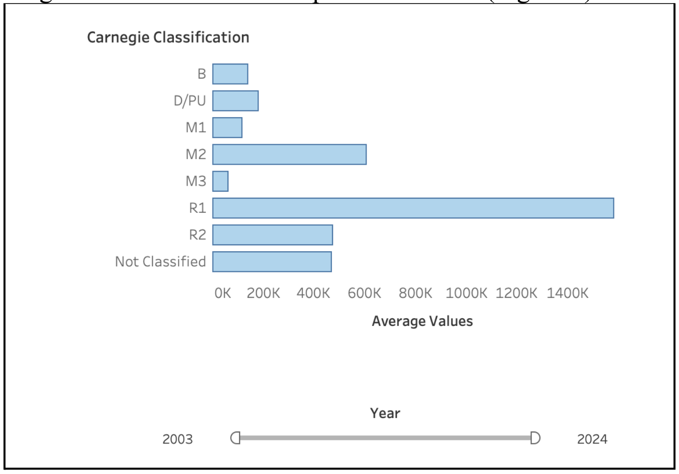
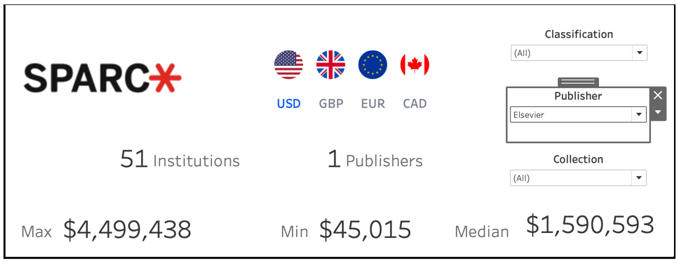


Figure 5: Carnegie Classification Distribution

A horizontal bar chart is used to show the distribution of average deal sizes across each Carnegie Classification. If a user wants to analyze data for a particular Carnegie Classification, that user can click the horizontal bar that corresponds to the desired Carnegie Classification. Additionally, there's an interactive slider below this graph that the user can use to control the period of time that person wants to analyze data for. Both of these filters constrain the data queried by all visualizations on the Comparative Visualization Tracker dashboard.



SPARC*

USD GBP EUR CAD

Classification: (All)

Publisher: Elsevier

Collection: (All)

51 Institutions

1 Publishers

Max $4,499,438

Min $45,015

Median $1,590,593

Figure 6: Interactive Filters

Figure 6 shows the upper right hand corner of the dashboard which has some more simple interactive filters. Aside from including an additional filter for Carnegie Classification, other filters for the Publisher and Collection attributes are also included as drop-down lists. The currency icons can be clicked so that a viewer can view deal price statistics in different currencies. This filter makes use of an open source python API which converts the transaction records between currencies using the real-time exchange rate between the two currencies. All of these filters constrain all visualizations on the Comparative Visualization Tracker Dashboard.

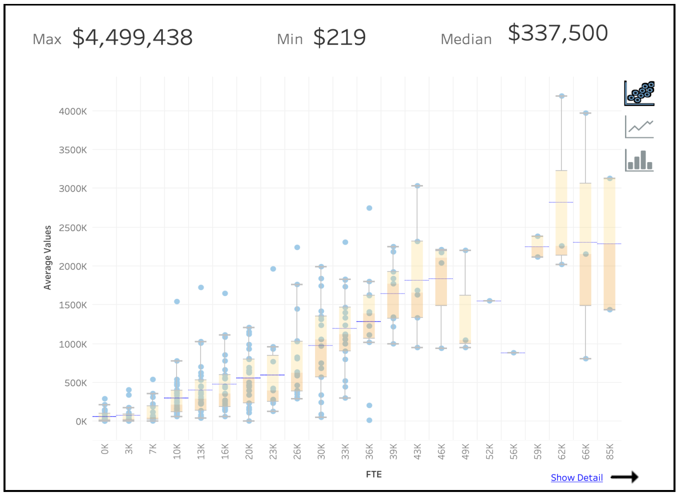


Figure 7: Deal Sizes Box Plot Visualization

Figure 7 above is a sample image of the primary visualization of this dashboard. The deal sizes averaged over time for each institution are plotted against ranges of FTE. Additionally, boxplots of these data points are shown for the same ranges of FTE. Users can also use the panel in the upper right hand corner of the graph to change the visualization that is displayed. The other two visualizations, instead of showing deal sizes averaged over all years within filter constraints for different ranges of FTE, show deal sizes averaged over all ranges of FTE within filter constraints over time.

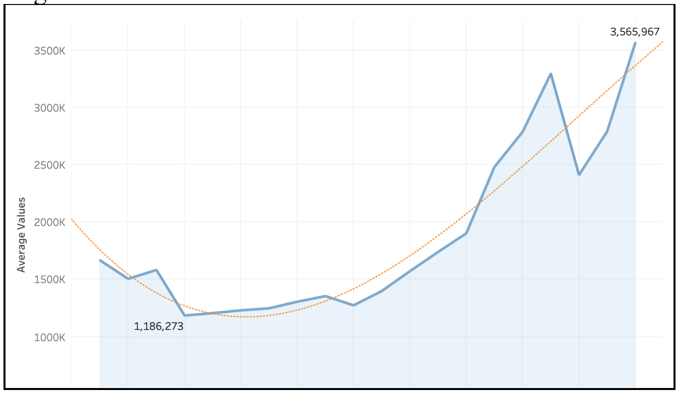
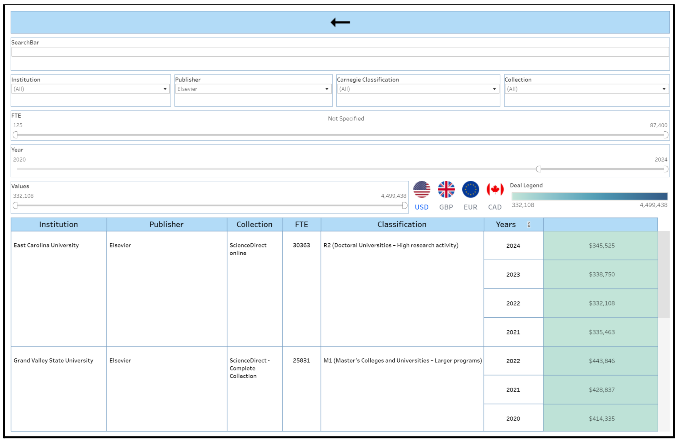


Figure 8: Deal Size Distribution

As seen in figure 8 above, one of the ways changes in deal sizes over time can be visualized is with a line graph. The blue line indicates the actual line graph while the orange line is a trend line that is fit using third degree polynomial kernel regression.

3.2 Big Deal Pricing Tracker



SearchBar

Institution: (All)

Publisher: Elsevier

Carnegie Classification: (All)

Collection: (All)

FTE: 125 – 87,400 Not Specified

Year: 2020 – 2024

Values: 332,108 – 4,499,438

USD GBP EUR CAD

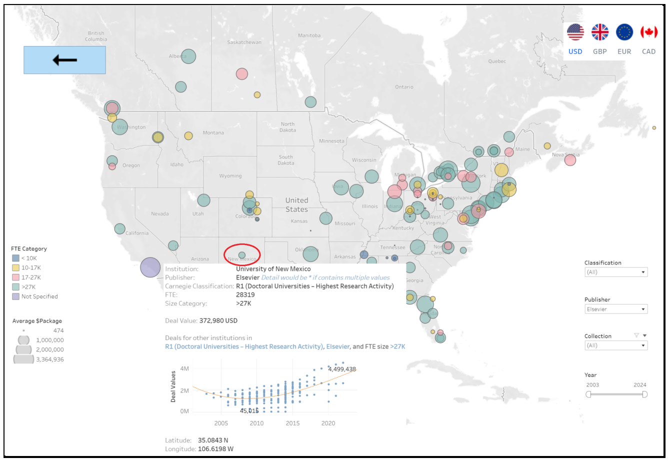
Deal Legend: 332,108 – 4,499,438

Institution	Publisher	Collection	FTE	Classification	Years	
East Carolina University	Elsevier	ScienceDirect online	30363	R2 (Doctoral Universities – High research activity)	2024	$345,525
					2023	$338,750
					2022	$332,108
					2021	$335,463
Grand Valley State University	Elsevier	ScienceDirect - Complete Collection	25831	M1 (Master's Colleges and Universities – Larger programs)	2022	$443,846
					2021	$428,837
					2020	$414,335

Figure 9: Big Deal Pricing Tracker Table View

Figure 9 above represents the Big Deal Pricing Tracker Table view. This visualization is similar to what the SPARC team displays on their web page currently but it adds more filters so that the user can gain insights more easily.

3.3 Geographic Pricing Tracker



Institution: University of New Mexico

Publisher: Elsevier *Detail would be * if contains multiple values*

Carnegie Classification: R1 (Doctoral Universities – Highest Research Activity)

FTE: 28319

Size Category: >27K

Deal Value: 372,980 USD

Deals for other institutions in R1 (Doctoral Universities – Highest Research Activity), Elsevier, and FTE size >27K

Latitude: 35.0843 N

Longitude: 106.6198 W

Figure 10: Geographic Pricing Tracker

Figure 10 represents the Geographic Pricing Tracker dashboard. This Proportional Symbol Map shows each institution on a map of North America and Great Britain. This visualization shows each institution as a circle centered at its approximate Longitude and Latitude. The area of the circle corresponds to the average price of all the purchases that institution has made. The color of each circle corresponds to the size of the institution, measured by FTE. This visualization can be filtered by Carnegie classification, publisher, collection and year.

4 How Our Dashboards Meet Insight Needs

We are highly confident our dashboards meet the SPARC's insight needs. The SPARC team wants visualizations that others can use to find a large list of desired insights rather than conducting research and analysis to discover select insights for buyers of academic resources. The three dashboards can enable these buyers to extract a wide variety of insights about the market for academic resources, examples of which will be discussed in the following three sections.

4.1 How to utilize the Comparative Visualization Tracker

The Comparative Visualization Tracker answers questions about relationships between FTE, Carnegie Classification, time, and distributions of prices. Here is an example of a use case that illustrates how users of SPARC's open data could use this dashboard for analysis. Let's say you're the Dean of the Indiana University School of Medicine, and you want to purchase the ScienceDirect Freedom Package from Elsevier. You're going to want to know what other institutions have paid for this resource, especially those institutions that are also R1 Carnagie Classified and somewhat close to 50,000 FTE. Using the interactive filters on the dashboard, a user can query the data so that only transactions where the buyer has R1 Carnegie Classification and an FTE between 40,000 and 60,000 are used as inputs to the visualization. Additionally, the user can use the filters from Figure 6 to specify the ScienceDirect Freedom Package from Elsevier. Doing so yields figure 11 and 12, shown below.:

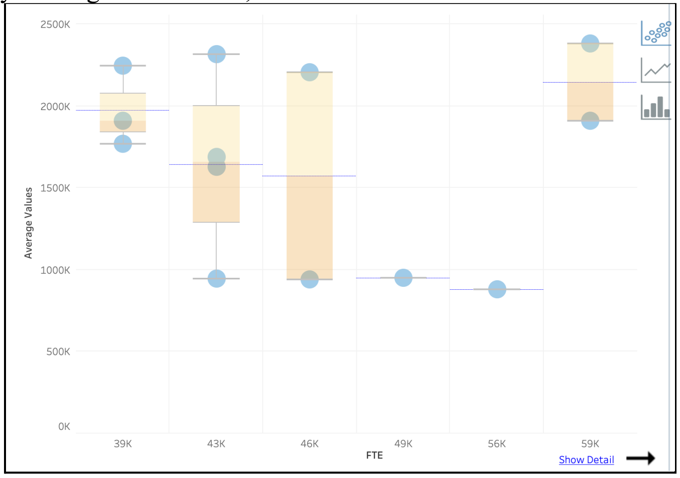


Figure 11:Example Boxplot

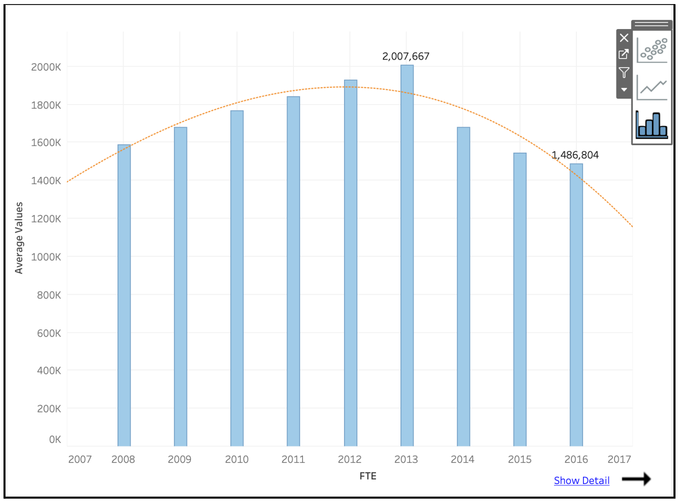
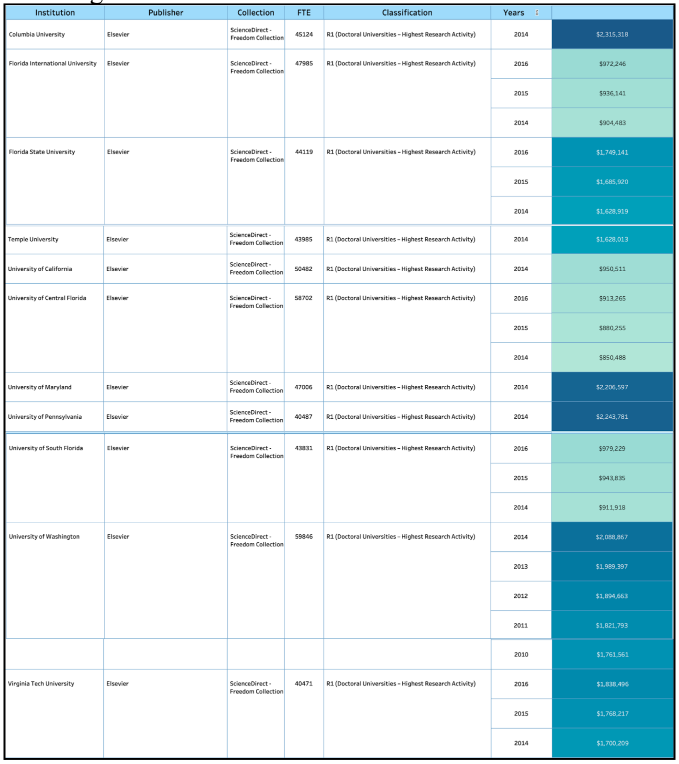


Figure 12: Example Bar Graph

The graph on the bottom tells us that the deal size, averaged over all FTE ranges, for each year given the filtering peaked in 2013 and has since gotten down. The graph on top suggests that there is not much of a relationship between deal size averaged over all years and FTE. So, it does seem that year explains more variation in deal size. However, the distribution of deal sizes averaged over all years for each institution almost looks bimodal with peaks at 1 and 2 Million US Dollars. This may just be noise, or this may mean there are two different packages sold within this collection where the fair prices for each are 1 and 2 million US Dollars respectively. Either way, the Dean of the Indiana University School of Medicine would know from looking at these visualizations that she should expect to pay at least 1 Million US Dollars for the Freedom Collection but should probably avoid paying any more than 2 Million US Dollars.

4.2 How to utilize Big Deal Pricing Tracker

The Big Deal Pricing Tracker answers questions that are similar to those of the Comparative Visualization Tracker, but the Comparative Visualization Tracker shows transaction details more granularly in a tabular view. This dashboard includes all of the same filtering capabilities as the Comparative Visualization Tracker and includes a search bar. Instead of showing prices averaged over all years for a buying institution or showing the price for each year averaged over all institutions, the Big Deal Pricing Tracker shows individual transactions in a tabular view. Going back to the previous use case, we can use the Big Deal Pricing Tracker to view individual transactions that comprise the averages we saw on the Comparative Visualization Tracker. Using the same filtering as we did with the last dashboard we can see the following table:



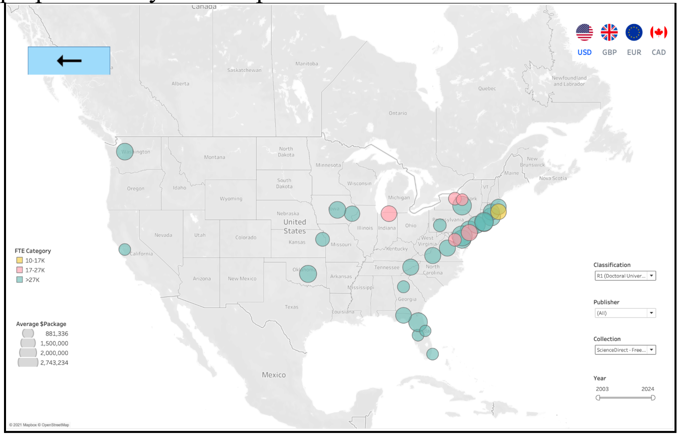
Institution	Publisher	Collection	FTE	Classification	Years	
Columbia University	Elsevier	ScienceDirect - Freedom Collection	45124	R1 (Doctoral Universities – Highest Research Activity)	2014	$2,315,318
Florida International University	Elsevier	ScienceDirect - Freedom Collection	47985	R1 (Doctoral Universities – Highest Research Activity)	2016	$972,246
					2015	$936,141
					2014	$904,483
Florida State University	Elsevier	ScienceDirect - Freedom Collection	44119	R1 (Doctoral Universities – Highest Research Activity)	2016	$1,749,141
					2015	$1,685,920
					2014	$1,628,919
Temple University	Elsevier	ScienceDirect - Freedom Collection	43985	R1 (Doctoral Universities – Highest Research Activity)	2014	$1,628,013
University of California	Elsevier	ScienceDirect - Freedom Collection	50482	R1 (Doctoral Universities – Highest Research Activity)	2014	$950,511
University of Central Florida	Elsevier	ScienceDirect - Freedom Collection	58702	R1 (Doctoral Universities – Highest Research Activity)	2016	$913,265
					2015	$880,255
					2014	$850,488
University of Maryland	Elsevier	ScienceDirect - Freedom Collection	47006	R1 (Doctoral Universities – Highest Research Activity)	2014	$2,206,597
University of Pennsylvania	Elsevier	ScienceDirect - Freedom Collection	40487	R1 (Doctoral Universities – Highest Research Activity)	2014	$2,243,781
University of South Florida	Elsevier	ScienceDirect - Freedom Collection	43831	R1 (Doctoral Universities – Highest Research Activity)	2016	$979,229
					2015	$943,835
					2014	$911,918
University of Washington	Elsevier	ScienceDirect - Freedom Collection	59846	R1 (Doctoral Universities – Highest Research Activity)	2014	$2,088,867
					2013	$1,989,397
					2012	$1,894,663
					2011	$1,821,793
					2010	$1,761,561
Virginia Tech University	Elsevier	ScienceDirect - Freedom Collection	40471	R1 (Doctoral Universities – Highest Research Activity)	2016	$1,838,496
					2015	$1,768,217
					2014	$1,700,209

Figure 13: Example Table

From this table, we can see that schools like Florida International University, University of Central Florida, and University of South Florida, got the freedom collection for less than 1 Million US Dollars while Johns Hopkins, Columbia, Penn, and NYU paid well over 2 Million US Dollars. Geography may play a role in these differences, but it's hard to tell since the University of Florida appears to be paying more than their peers for this resource. If geography influences these differences, we will be able to determine that using the Geographic Pricing Tracker.

4.3 How to utilize the Geographic Pricing Tracker

The Geographic Pricing Tracker helps users discover insights related to where buyers are located. Do buyers in one part of the country get better deals on a collection than buyers in another part of the country? Are there two collections that are similar, and buyers in the US overwhelmingly prefer one collection, while buyers in the UK overwhelmingly prefer another? Questions like these are easy to explore with the Geographic Pricing Tracker. To illustrate its utility, we're going to use the same example we have for the last two dashboards. Using insights from the Big Deal Pricing Tracker, we found that of schools that are of R1 Carnegie Classification with an FTE within 10,000 of Indiana University, three schools in Florida appeared to get relatively good deals on the Freedom Collection with the exception of the University of Florida. Using the same filtering we used for this use case on the last two dashboards yields the following proportional symbol map:



Classification
R1 (Doctoral Univer...

Publisher
(All)

Collection
ScienceDirect - Free...

Year
2003 2024

Figure 14: Example Proportional Symbol Map

From this map, it doesn't appear that there are any insightful links between deal size averaged over all years and geography. This dashboard does tell us that most of the R1 classified institutions that purchased the Freedom Collection are on the east coast. All in all, using these three dashboards to analyze this use case doesn't provide enough insight for the Dean of the Indiana University School of Medicine to negotiate an appropriate price for the Freedom Collection, but he or she can easily compile a list of schools who got the best and worst deals on this resource, reach out to them, and hopefully gain all the necessary insights to negotiate a deal for the Freedom Collection.

5 Economic Recommendations to SPARC

- We recommend that SPARC build a pipeline that can scrape the internet for all transaction data related to academic resources. Manually entering thousands of records guarantees errors, and errors can render insights meaningless. We also speculate that there is a lot of missing data that could be helpful.
- We recommend that SPARC incorporate more independent variables. Using our dashboards, it doesn't seem that FTE explains any variation in deal quality that Carnegie Classification doesn't already explain. We think FTE should be split up into more granular attributes like number FTEs of undergraduate students, number FTEs of graduate students, and number of FTEs of staff. We also speculate that prestige and endowment characteristics explain a significant amount of variation in deal quality and should be included in the data set. Overall, it's very difficult to extract insights that are useful from just two independent variables.
- We recommend that SPARC categorize each collection. For example, Elsevier's ScienceDirect Freedom Collection is a comprehensive collection of medical journals and biological research. In this case, it's probably more helpful to look at the size of a buying institution's medical school or relevant department. We aren't sure what level of granularity is appropriate for classification, but it's helpful to know, for example, that the Freedom Collection is primarily used by pre-med students and medical students.

Acknowledgments

The authors wish to thank the SPARC organization and Indiana University for their partnership to offer students with a real world project experience. Additional thanks to course instructors Katy Borner, Andreas Bueckle, and Michael Ginda for their guidance and assistance.

References

[1] SPARC Big Deal Knowledge Base, https://sparcopen.org/our-work/big-deal-knowledge-base, 2007-2021.