O Brave New Print Collection, That Has Such Data Science Books in It!

Heidi Tebbe
North Carolina State University Libraries, hjtebbe@ncsu.edu

Mira Waller
North Carolina State University Libraries

Author ORCID Identifier: https://orcid.org/0000-0002-5482-4035

Heidi Tebbe and Mira Waller, "O Brave New Print Collection, That Has Such Data Science Books in It!" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316691

O Brave New Print Collection, That Has Such Data Science Books in It!

Heidi J. Tebbe, North Carolina State University Libraries

Mira Waller, North Carolina State University Libraries

Abstract

The field of data science exists at the intersection of several disciplines, including statistics, social science, information science, computer science, and visualization. This can make collection development for data science challenging, but it's a field that has become increasingly important in industry and academia. Data scientists, and increasingly researchers, academics, and others, work with large amounts of data, complex computation, and data visualization to solve real-world problems. Those working in or studying data science may need to learn new skills and tools to be successful.

North Carolina State University (NCSU) recognizes the importance of this growing field, as shown in the establishment of the Data Science Initiative (DSI); courses taught by faculty in computer science, statistics, advanced analytics, and management; and research conducted at interdisciplinary centers and institutes.

This poster session will describe how librarians from the Collections & Research Strategy department at NCSU Libraries conducted a project to build a niche data science print collection. Information shared in the poster will include the sources that were used to compile an initial list of books, including recommendations from fellow librarians, a curated GOBI notification, websites, suggested reading lists, and course syllabi. Criteria for narrowing this initial list will be provided. The poster will also show an analysis of how this collection overlaps with more established collection areas.

Wherefore Data Science?

The field of data science exists at the intersection of several disciplines, including statistics, social science, information science, computer science, and visualization. This can make collection development for data science challenging, since purchasing systems and protocols for libraries are based upon discipline- and domain-specific classifications. Data science spans several of the main classes of the Library of Congress Classification system, requiring more nuanced and frequent human intervention to ensure that content and materials are appropriately purchased. Despite the collecting challenge presented by data science's interdisciplinary nature, it is a field that has become increasingly important in industry and academia and is thus an important field to consider for our collections.

Both human- and machine-created data continue to grow at a rapid rate, "doubling in size every two years" (Turner, 2014). In order to analyze and work with these increasingly large sets of data, new technologies and systems are being created. Data scientists, researchers, academics, and others work with large amounts of data, complex computation, and data visualization to advance science and scholarship while solving real-world problems. Those working in or studying data science may need to learn new skills and tools to be successful.

North Carolina State University recognizes the importance of this growing field, as shown in the establishment of the Data Science Initiative (DSI) in 2014. The DSI was "sponsored by the Office of Research, Innovation and Economic Development at NC State University to create a nationally recognized hub of excellence in data science and analytics" ("About: Data Science Initiative [DSI]," n.d.). The importance of data science at NC State University is also demonstrated by research conducted at interdisciplinary centers and institutes and courses taught by faculty in computer science, statistics, advanced analytics, and management.

E-Preferred (Usually)

NCSU Libraries operates on an electronic preferred approval plan with electronic purchasing priorities, meaning that we prefer to purchase electronic books in most cases. However, the format focus of this niche data science book collection is print for a number of reasons. The collection was originally created for inclusion in a renovated physical

https://doi.org/10.5703/1288284316691

location to highlight both the space and the books. Since we also get requests for print books in related subject areas such as statistics and mathematics, we believed this would be a great way to show our understanding of users' needs and desires to read and reference certain kinds of texts in print. The print collection would also be easy to access as a quick reference point. Additionally, this collection could be used to showcase our support of the growing importance of data science and help to reinforce the alignment of a particular space as a data experience hub in the Libraries.

Our Process

In the spring semester of 2016, we were asked to create a data science print collection with an allocation of $5,000. Our first objective was to create a broad set of criteria for the collection. As part of this process, we started with a very large set of keywords, narrowed down to the following: data viz; dashboards; web scraping, python; altmetrics; big data; big data analytics; big data, data science, management; business data science; code art; comics/viz; data—private vs. public; data analysis, physics; data analysis as art; data analysis; predictive analysis; data analytics; data and research ethics; data and text viz; data fundamentals; data management; and data, general. We also indicated areas that we would not include, such as big data infrastructure, and areas that we would review on a case by case basis, such as books geared toward a librarian audience. Since we already have strong collections in some areas, such as statistics, mathematics, business, computer science, and human medicine/biomedical, we made sure to touch base with the selectors responsible for those areas.

In order to cast a wide net for the collection, we searched for potential books using both internal and external sources. Since data science spans a number of different disciplines, we solicited recommendations from librarians whose collection responsibilities touched upon data science. We also reached out to noncollections librarians whose responsibilities included aspects of data science such as visualization or programming. In addition to checking with internal experts, we searched through websites about data science, combed through subject and recommendation lists on Amazon and Goodreads, and reviewed syllabi for data science courses and programs. We used EBSCO's GOBI Library Solutions to set up a notification scoped to new books in data science. Since our fund codes were too broad to be used in setting up the notification, and GOBI did not have an interdisciplinary topic for data science, we had to create our notification using subject terms. The terms we used included data mining, information visualization, big data and visualization, data protection, data structures, and distributed processing.

Based upon our criteria, we created an initial list of books for review. We removed books from this initial list that the libraries already owned in print. We did not remove books that we already owned in an electronic format (about one-third of the books on our initial list). Once we had refined our list, we associated each book with an existing fund code. Then we sent the list to our Acquisitions & Discovery department for purchase. As part of the cataloging process, a unique identifier was added to the catalog records in order to internally identify books in the collection.

The Collection

The initial list of data science monographs included 133 books, of which 75 books were actually purchased. The books span the collecting areas of 7 selectors and the following 16 subject fund codes:

- Computing & Electronics
- General Science
- General Humanities & Social Sciences
- Economics
- Library Science
- Statistics
- Business Management
- Education
- Language & Linguistics
- Mathematics
- Psychology
- Graphic & Information Design
- Industrial Design
- General Technology
- Human Communication
- Physics

The books can be roughly divided into four categories:

- Data and Data Management
- Data Processing and Analysis
- Programming
- Visualization

Between June 2017 and October 2017, more than one-third of the newly purchased books had already been checked out at least once, with those about programming and visualization being the most popular of the books checked out.

Present and Future

The physical space that this print collection was intended for has not been completed yet. SirsiDynix Symphony Workflows provides item category fields that can be used to indicate the collection and/or acquisition statuses of an item. The presence of a "DATASCI" identifier in an item category field will allow us to easily locate the books when the space is ready for them.

This print collection provided a way to refine future search efforts from using keywords to using call numbers. Books about visualization can be found in call numbers such as BF241, HF5718.22, P93.5, QA76.9.I52, and QA276.3. Books about data processing and analysis can be found in call numbers including H61.3, H61.4, HA29, HD38.7, HF5415.125, HF5548.2, Q183.9, and QA276.4. Data and data management books can be found in call numbers including QA76.9.B45, QA76.9.D32, Z666.7, and ZA4080. Data science–related programming books can be found in Q325.5, QA76.73.A-Z (e.g., QA76.73.P98 for the Python programming language or QA76.73.R3 for the R programming language), QA276.45.A-Z (e.g., QA276.45.R3 for the R programming language), QA76.9.A43, and QA76.9.D343. A new GOBI notification has been created using these call numbers; this will allow us to continue to keep our data science offerings up-to-date. In the future, if it seems necessary, we may create a subject fund specifically for data science materials.

Circulation statistics of this print collection will be monitored. Additionally, the call numbers listed in the previous paragraph may also be used as a means to investigate the circulation of data science–related books in our collection that are not part of this print collection, to better inform future collecting in this area.

References

About: Data Science Initiative (DSI). Retrieved from https://research.ncsu.edu/dsi/about/

Turner, V. (2014). *The digital universe of opportunities: Rich data and the increasing value of the Internet of things*. Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm