IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 27.06.2018

Pushing the Boundaries of Data Services Ecosystem at an Academic Library

Yun Dai
Library, New York University Shanghai, Shanghai, China.
E-mail address: yun.dai@nyu.edu

Copyright © 2018 by Yun Dai. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: http://creativecommons.org/licenses/by/4.0

Abstract:

In academic libraries, data librarians help researchers with data discovery, access and curation. At the Library of New York University Shanghai (NYU Shanghai), we have pushed the boundaries of data librarianship to newer fields of services and initiatives by means of deeper integration with technology, larger roles to take on than a service provider, and extending our services to a wider community via more channels.

This paper first introduces the operation of our data services as an ecosystem. It then explores how the boundaries of data services have been expanded through several successful cases. Finally, it discusses the implications of this program in terms of how it may be applied to a campus with different population and scale than ours, and how it may benefit researchers' data management purposes.

Keywords: data services, technology, data literacy, global education, big data

In academic libraries, data librarians help researchers with data discovery, access and curation. At the Library of New York University Shanghai (NYU Shanghai), we have pushed the boundaries of data librarianship to newer fields of services and initiatives by means of deeper integration with technology, larger roles to take on than a service provider, and extending our services to a wider community via more channels.

This paper first introduces the operation of our data services as an ecosystem. It then explores how the boundaries of data services have been expanded through several successful cases. Finally, it discusses the implications of this program in terms of how it may be applied to a campus with different population and scale than ours, and how it may benefit researchers' data management purposes.

Data Services Ecosystem

A glance at the existing practices of how libraries of several leading universities integrate data and technology into daily services reveals at least three service models. The first long existing model takes the form of the statistical and computing consulting group, functioning as an organic unit within the library. The group offers consulting and/or training on statistical programming and computing, data analysis and interpretation, modeling, and research methodology with the primary goal of facilitating research publication and grants application. Such services place mathematical and computational assistance at core where advising on research design and implementation are indispensable. Princeton University Library's Data and Statistical Services[1] and Yale University Library's StatLab[2] have been embracing this model, along with their partners, which usually are the social sciences and digital research institutes[3] and the academic departments[4].

The second model establishes data services in a full life cycle, including data discovery, data management, data cleaning, data analysis, data visualization, and mapping and GIS. In this model, deep knowledge of statistical and computing methods and software is important yet not as essential as in the first model. NYU Libraries' Data Services[5] and Data and Visualization Services at Duke University Libraries[6] are university libraries that adopt this model. The social sciences divisions of some university libraries have also taken up this model and its variations, such as the Social Science Data and Software unit at Stanford Libraries[7], and the Research Data Services at the Digital Social Science Center, Columbia University Libraries[8].

In the third model, the libraries focus on data management or curation without any involvement in data computing services.

At NYU Shanghai, the data services program stands in the middle ground of statistical and computing consulting and data librarianship, where services on data, technology and research methodology converge. We will see how the three elements work to form the service model in the next section of case studies.

Our data services program started operating in 2017. Since then we have built an ecosystem of data services that is structurally agile, adaptable to the arising needs in scholarly activities, and responsive to the ever-changing external trends. The building blocks of this program are (1) technical support and advising for research, teaching and learning; (2) collaborative data projects with other departments; and (3) data literacy programs. In addition, we are actively developing and integrating the data management component into this ecosystem.

At the Library, data technologists partner with subject reference librarians to cover the needs of data use in stages of data discovery, cleaning, analysis, and visualization. The librarians provide data discovery and access services, while the data technologists take over the rest of the needs. One benefit of this one-stop shop model is smooth flow of information and close coordination. For instance, it is very common for reference librarians to introduce users to data technologists for further technical support; on the other hand, data technologists can easily direct users to the reference librarians if they identify data discovery needs from users.

To enrich the skill sets and range of services that we offer, the Library has been bringing one Technology Enhanced Education Fellow each year with expertise in one area. The person tackles the most urgent need of the Library in response to industry changes. The first year we hired a technology assessment fellow with expertise in statistical programming to work on the pilot massive online course initiative. The following year we had a fellow to apply machine learning and other technologies to enhancing instructional experiences.

The data services program is structured in the organizational settings, university culture and institutional characteristics of NYU Shanghai. It is a portal campus in the NYU global network and a joint venture between NYU and East China Normal University. Founded in 2012, it is by far a campus with around 1200 enrolled undergraduate students from more than 70 countries and more than 200 full-time international faculty members. Students usually spend their third year studying abroad at New York, Abu Dhabi campus or other global academic centers.

This has affected the ways we deliver the data services program. Being a small population allows us to take the advantage of centrality of communication, and cross the disciplinary borders in research and outreach activities more easily than larger campuses. Being a very young university allows us to create in a semi-startup environment without much of historical burdens. Being part of the NYU global network allows us to leverage the global resources when it goes beyond the capacity of local service offerings. It also means we serve a mobile student body - our contact with our own students relies on face-to-face interactions but also digital channels while they are travelling; meanwhile, our services extend to the study away students from other global sites. However, those traits should not limit our data services program only to campuses sharing similar demographic and environmental features. I will return to this point in the final section of implications.

Case Studies

In this section, I will explore the approaches and attempts towards forming the current data services ecosystem through the success cases for each piece of our service model.

(1) Creating an environment where technology is embedded in daily data services.

The libraries have always been primary service providers of data reference, data collection, and increasingly data curation. Less discussed is the data computing services[9], although in the era of big data and AI the landscape may be shifting.

Our data services program is part of the research and instructional technology services under the Library. Technology, therefore, is in the gene of the program. The program offers support to and input in research projects and teaching on statistical programming, data visualization, data cleaning, data analysis, GIS mapping, and machine learning with tools such as *R*, *Stata*, *Python*, *D3.js*, *ArcGIS*, *Carto*, *R Shiny App*, *Tableau*, *Gephi* and *TensorFlow*.

One key element is services on programming and computing where data technologists' skills in the latest technology is an asset to the project team's domain knowledge. We have several successful stories in this regard. In the first case, the GIS specialist visualized NYU Shanghai students' studying away pattern with D3.js for the global affairs office, quantifying the big picture of students travelling and studying within the global network otherwise less visible to the administrators[10]. In the second case, a data technologist built an interactive web application with R's Shiny, allowing for deeper examination of the user needs by survey questions, student groups, and plot types. In another case, the data technologist supported a faculty member's research with Stata programming for needs arising from complicated data structures and statistical modeling. In the last case, the technology fellow used his knowledge in machine learning to train criminal images for a faculty member's investigative psychology project. In all cases, be they scholarly or administrative, data technologists complemented the project team's domain knowledge with their computing and programming skills.

The second key element is workshop delivery, custom instruction or guest lecturing. Data technologists deliver regular workshops on tools of statistical programming, GIS mapping, machine learning and data visualization to faculty and students. We also offer custom in-class instructions to students by faculty request, including web scraping with Python for a business analytics class, Stata programming for a business honors seminar class, and ArcGIS and Story Maps for a Chinese history class.

To stay connected with our users after classroom instructions and when they travel globally, we have developed websites or contents on social media to maintain the contact. For instance, following each data workshop, we would publish a post on WeChat (an ubiquitous social networking app in China) for further self-paced learning pertaining to a recently concluded workshop, each as part of the "Data Resources Mini Series". For instance, a post on data visualization books would be released following a Tableau workshop. This is pushed to every student, faculty, and staff on their mobile devices. Another example is the Stata website[11] initially developed for managing the workshop materials. But it has grown to be larger than a workshop site, when local and exchange students from other campuses keep accessing the tutorials on the website for self-paced learning when they travel globally. The site attracted traffic from more than 30 countries/regions since its launch date.

The third key element of our data services program is consulting, advising, or occasionally some coaching, on statistical programming, modeling, and debugging in the context of research design implementation. In one case, a senior student came to the data technologist to get help for her capstone project on a sociology topic within a tight timeline. She came with basic knowledge of statistics, little experience with statistical software, and limited training in research methodology. In this case, some individual coaching on a bit of everything - statistical modeling, programming, and research design implementation - was inevitable. Technical assistance and statistical consulting were offered in each step of research workflow to the advisee.

The boundary of the Library's data services has thus been expanded by facilitating research and teaching as a technical partner to faculty, students and administrators who conduct research. Leveraging social media and digital platforms also allows us to magnify the reach of services to a larger group of library service users than if we offer only face-to-face services.

(2) Partnership with other departments to develop data products and research projects.

In the increasingly more data-driven world of research, opportunities abound for libraries to become more than data services providers but collaborators and partners, locally with departments and administrators and even with the larger community[12].

Our *Chinese Datasets Archive*[13], an ongoing collaboration between our Library and the Data Science Center, is an example in this regard that seeks to tackle the challenges in discovering datasets[14]. This portal is a searchable catalogue of open Chinese datasets. The portal features a variety of data types (statistical, GIS, textual, images etc.), data sources (survey data, web scraped data, administrative data etc.), and subject areas. It serves as a starting point for researchers and students to search for open data on China. Faculty members also use the portal in teaching data science courses, such as Introduction to Machine Learning.

The Library was approached by the Data Science Center with a prototype of the data portal, at the time ambiguously defined regarding its scope, loosely configured regarding its structure, and limited in dataset listings. The project team was formed consisting of a library data services technologist, a Data Science Center administrator and a university web services developer. The Library's role evolved soon from a consultant to an equal partner deeply involved in every aspect of its creation - restructuring the datasets organization, identifying datasets, reviewing the datasets' quality, describing datasets, tracking the status of hosting websites, and making sure the listed datasets comply with laws and regulation governing datasets acquisition and other activities to avoid any risks and controversies.

At the first stage, we revamped the portal with what was most feasible and solicited feedback from scholars globally. The reference librarians also offered critical comments for improvement, based on which in the second stage of development we enhanced several features, including the more complete metadata labels for users to reliably decide whether to make efforts to access a dataset hosted elsewhere.

The Chinese Datasets Archive is but one example of how we work with other departments to develop products that benefit communities local and larger. There are other projects, on a lesser scale, which made use of data technologists' skills. For instance, the data technologist once worked with the academic advising department to evaluate admission metrics in relation to students' first-year academic outcomes. The result was a compiled report submitted to the university's admissions committee.

In all cases, the data technologists' knowledge and expertise were deeply valued and contributed to the projects with their technical skills and insights into data. The boundary of the data services has been again expanded by establishing ourselves not only as service providers but creators of useful resources.

(3) Collaborating with university-wide initiatives to lead data literacy campaigns and events.

One challenge of libraries' data services is to find its niche in the larger "data ecosystem" of the university. For instance, what is the role of the data services program in relation to the activities and events hosted by the research institutes, such as the data science center or the business analytics center? They are the knowledge hubs with groups of interdisciplinary scholars conducting the most advanced research with quantitative and computational methods. Besides, what is the comparative advantage of our program, such as the workshops, to the formal courses teaching data science? Between the two extremes of the "data spectrum", where are we?

As the data services program evolved, we have found ourselves a spot somewhere in the middle: our program can be and should be complementary to the course offerings and the formal scholarly activities. One way to categorize our data services program is to view it as data literacy efforts. Let the data scientists be data scientists, and we can be the "ambassadors" bridging the gap between those with little interest and knowledge in data science and the seasoned data practitioners. Academic libraries are the ideal places to host such events. For instance, the New York University Health Sciences Library[15] developed a series of classes to help attendees overcome gaps in acquiring data skills, which tackled one data skill each time in data management and visualization. The program also increased the library data services' visibility.

In the design of our workshops and other services, we have been offering technical support where not covered in classes; programming and computing support where a researcher looked for technical input to complement his or her domain knowledge; and workshops with important topics often neglected in formal academic training but sought after in industries. As such, we also turn potential competitors in service offerings to collaborators.

In addition to the technical workshops and consulting we have been providing to the university, we were collaborating with the university's Committee of Critical Inquiry for a live conversation on "Lying with data" to explore how studies, reports, visualizations and narratives built upon numbers, statistics, and algorithms conceal and mislead as much as they reveal. It brings various perspectives together through case studies and open discussions on, for instance, what biases could trap us in statistical reasoning and what lessons we could learn from cases of falsified and fabricated data in scientific studies. The Committee was convened by the university leadership with members from faculty, staff and students to address critical thinking, which is a larger platform to mobilize resources and call attention to the data literacy topics. We expand the boundary of data services, therefore, by extending our services to a wider group of audience.

Implications

I discussed our data services program in the context of NYU Shanghai campus. However, the service model can be applied to other academic libraries, if we translate the benefits of a small university into the organizational settings. This would mean: (1) Create a flatter structure where information exchange is at lower costs. This would include communication within the library between librarians and technologists, and externally with the academic departments to be sensitive to their needs. (2) Find your niche in the university's larger data services ecosystem. In our case, we place ourselves in the position of data literacy advocates; we maintain a relationship with the data science research hubs where we can be a collaborator on data projects; we also work with university-wide platforms to promote literacy programs to the more general audience. (3) Be adaptive and responsive to the fast-changing needs in the research world and the industries, and develop skill sets that are up-to-date to serve the people who are often smarter than ourselves. These three methods should be applicable in campuses independent of their sizes.

Another implication of our service model in data management is less formal but more organic. Data curation is not crystallized yet in our service model but embedded in many designs. In the technical workshops and consulting, research workflow and reproducibility have been at the heart of our concern. For instance, in many statistics courses, concepts, equations and models are taught but not code and data reproducibility. We make sure this is included and emphasized in the statistical programming or machine learning workshops we offer. Additionally, workshops on GitHub, Zotero and Google Scholar profile touch upon the management needs directly regarding versioning and file management. Moreover, as part of the NYU global network, the researchers can always turn to NYU New York for data management needs, where the data librarians assist with data management planning for grant applications and reproducibility purposes.

Acknowledgments

I would like to thank Edward Junhao Lim, Xiaojing Zu, Adrian Hodge, Jennifer Anne Wood Stubbs, Caitlin MacKenzie Mannion, Qinghua Xu and Fan Luo for all your advising, guidance and support. Thank you Edward for urging me to submit my work. Thank you Jennifer for spending time reviewing my paper word by word.

References

[1] "Data and Statistical Services," Princeton University Library, accessed 26 May, 2018. https://dss.princeton.edu/

[2] "StatLab," Yale University Library, accessed 26 May, 2018. http://statlab.stat.yale.edu/

[3] "Data Science Services," The Institute for Quantitative Social Science, Harvard University, accessed 26 May, 2018. https://dss.iq.harvard.edu/; "about us," UCLA: Statistical Consulting Group, Institute for Digital Research and Education, accessed 26 May, 2018. https://stats.idre.ucla.edu/ucla/about/; "SSCC Statistical Consulting," Social Science Computing Cooperative, University of Wisconsin - Madison, accessed 26 May, 2018. https://www.ssc.wisc.edu/sscc/statconsult.htm; "Statistical Consulting," Institute for Social and Economic Research and Policy (ISERP), Columbia University, accessed 26 May, 2018. http://iserp.columbia.edu/node/302; "Faculty: Data & Research Services," Institute for Research in the Social Sciences (IRiSS), Stanford University, accessed 26 May, 2018. https://iriss.stanford.edu/faculty-research-data-services; "Students: Data & Research Services," Institute for Research in the Social Sciences (IRiSS), Stanford University, accessed 26 May, 2018. https://iriss.stanford.edu/students-data-research-services; "Connection Bar," The Social Science Research Institute, Duke University, accessed 26 May, 2018. https://ssri.duke.edu/connection/connection-bar.

[4] "Consulting Information", Department of Statistics, Columbia University, accessed 26 May, 2018. http://stat.columbia.edu/consulting-information/; "Consulting Services," Department of Statistics, Stanford University, accessed 26 May, 2018. https://statistics.stanford.edu/resources/consulting; "The Data Science Drop-in is a free educational and consulting service for members of the Stanford community," Department of Management Science and Engineering, accessed 26 May, 2018. https://5harad.com/drop-in.

[5] "Data Services: Home," New York University Division of Libraries, accessed 26 May, 2018. https://guides.nyu.edu/c.php?g=277095&p=1847020.

[6] "Data and Visualization Services," Duke University Libraries, accessed 26 May, 2018. https://library.duke.edu/data/.

[7] "Social Science Data and Software (SSDS)," Stanford Libraries, Stanford University, accessed 26 May, 2018. https://ssds.stanford.edu/.

[8] "Research Data Services," Digital Social Science Center, Columbia University Libraries, accessed 26 May, 2018. http://library.columbia.edu/services/research-data-services.html.

[9] Minglu Wang, "Supporting the research process through expanded library data services." *Program* 47, no. 3 (2013): 282-303.

[10] Fan Luo, "New York University, Shanghai - Study Away Data Visualization," accessed 26 May, 2018. http://nyush.hosting.nyu.edu/StudyAway/NYUSH_StudyAway.html.

[11] Yun Dai, "Stata Starters," accessed 26 May, 2018. http://shanghai.hosting.nyu.edu/stata/.

[12] Matt Burton et al., "Shifting to Data Savvy: The Future of Data Science In Libraries," University of Pittsburgh, last modified 19 March, 2018. http://d-scholarship.pitt.edu/33891/.

[13] "Chinese Datasets Archive 2.0," New York University Shanghai, accessed 26 May, 2018. https://datascience.shanghai.nyu.edu/datasets.

[14] Aaron Tay, "Datasets as a first class entity - preliminary musings," *Musings about librarianship* (blog), April 1, 2018, http://musingsaboutlibrarianship.blogspot.hk/2018/04/datasets-as-first-class-entity.html.

[15] Alisa Surkis et al., "Data Day to Day: building a community of expertise to address data skills gaps in an academic medical center," *Journal of the Medical Library Association* 105, no. 2 (2017):185-191.