Research Library Issues

The Data Science Revolution

RLI 298 (2019)

ASSOCIATION OF RESEARCH LIBRARIES

In This Issue

Correction, August 1, 2019: *On page 29, the year in which the National Agricultural Library and the US Department of Agriculture were founded was identified incorrectly in an earlier version of this publication. The year of founding was 1862, not 1863.*

Introduction

Judy Ruttenberg, Director, Scholars and Scholarship, Association of Research Libraries

Data analytics and data science programs in research universities in Canada and the United States have grown dramatically over the past decade. An intrinsically interdisciplinary endeavor—drawing from statistics, computer science, engineering, and more—data science education has followed multiple trajectories. Projections of labor market demand for data manipulation skills, growing acceptance of open science and data sharing, the reproducibility "crisis," and the promise of unlocking new scientific discoveries through analyzing massive amounts of data have all fueled enthusiasm for data science education. Funding models for new centers and institutes have multiplied, too—from large, personal, philanthropic donations (University of Virginia), to federal investment (Pan-Canadian AI Strategy and the Big Data to Knowledge, or BD2K, initiative), to private foundations (Gordon and Betty Moore and Alfred P. Sloan) that have invested in creating data science "environments" in three large US research universities.

This issue of *Research Library Issues* looks at the critical role and participation of libraries and librarians in supporting the data science revolution. **Catherine Murray-Rust, with Ameet Doshi, Jay Forrest, Ximin Mi, and Alison Valk**, demonstrates a pathway to data science through teaching core data-literacy skills to students. This is part of a broader library strategy at Georgia Tech to identify service gaps across the university that align with the library's mission, and work with faculty and students to fill those gaps. But the library's data-literacy program is doing much more than that. Through their offerings, including education on basic statistical packages and web scraping tools, the library is providing an informal opportunity for students in non-STEM fields to gain proficiency with these methods for their intellectual, professional, or vocational pursuits.

This is not surprising. "The library is the most interdisciplinary place on campus," offered New York University (NYU) physics professor David W. Hogg, in his reflections on the Moore-Sloan Data Science Environment (MSDSE) at his university and its strong ties to the library. **Jennifer Muilenburg and Judy Ruttenberg** spoke with Hogg and other key personnel at the three MSDSE institutions—NYU, UC Berkeley, and the University of Washington—to highlight the profound growth and transformation of data science education that the MSDSE funding catalyzed. The article charts the instantiation of data science education at the three institutions, their aligned but distinct areas of focus, and key lessons for library leaders and other campus administrators in this arena.

Finally, **Erica Johns, Susan McCarthy, and Cynthia Parr** contribute two companion pieces on the highly collaborative growth of data science services at the US National Agricultural Library (NAL). Working with the University of Maryland iSchool and other land-grant university members of the Agriculture Network Information Collaborative (AgNIC), NAL identified scalable strategies to partner with the Agricultural Research Service (ARS) and the university research community to advance data-driven discovery in agriculture.

Data science is new enough as an endeavor that its relationship to its core disciplinary antecedents, and its infiltration of disciplines across the curriculum, is a dynamic and unfolding story in higher education. Research libraries, however, are prepared for this future as we draw upon our long-standing contributions: creating the conditions for new knowledge discovery, teaching students how to discern validity, and partnering with the research community to prepare and preserve data for science.

"Research libraries...are prepared for this future as we draw upon our long-standing contributions: creating the conditions for new knowledge discovery, teaching students how to discern validity, and partnering with the research community to prepare and preserve data for science."

To cite this article: Judy Ruttenberg. "Introduction." *Research Library Issues*, no. 298 (2019): 3–5. https://doi.org/10.29242/rli.298.1.

Data Literacy as a Pathway to Data Science at Georgia Tech

Catherine Murray-Rust, Dean of Libraries, Georgia Tech

in collaboration with:

Ameet Doshi, Director, Service Experience and Program Design, Georgia Tech

Jay Forrest, Data and Statistical Analysis Manager Librarian, Georgia Tech

Ximin Mi, Data Visualization Librarian, Georgia Tech

Alison Valk, Multimedia Instruction Librarian and College of Computing Liaison, Georgia Tech

A 2019 Gartner report, *Design a Data and Analytics Strategy*, asserts that "data literacy is a growing challenge for most organizations. By 2020, 80% of organizations will initiate deliberate competency development in the field of data literacy, acknowledging their extreme deficiency."[1] Campus collaborations increasingly rely on the ability to curate big data in support of the research community, and undergraduate and graduate student work increasingly requires basic skills in interpreting and presenting data. Librarians in their role as educators—serving the whole campus and the whole person—appreciate that data literacy, as distinct from data science in support of research, is fast becoming a required second language in a digital society.

One of the basic tenets of Georgia Tech Library's strategic initiative, Library Next, is a continuing analysis, based on purposeful interactions with faculty and students, of service gaps that the library can fill. Through Library Next, Georgia Tech is creating a new version of a technologically focused research library for the 21st century. Begun several years ago as a library building renovation, Library Next is now a major transformation of the library's vision. All components of a research library—inspirational physical and digital environments, curated scholarly content, outstanding services, and information

expertise—continue to change as the research and teaching goals and aspirations of Georgia Tech change. Data literacy, a pathway to data science, is one such service.

Mission Continuity

The Georgia Institute of Technology was founded in 1885 to help propel the agrarian South into the industrial economy. Today, the institute is a leading global research university, committed to improving the human condition through advanced science and technology. Georgia Tech's first library building, which opened in 1907, is one of 108 academic libraries (out of the total of 1,687 libraries) that Andrew Carnegie funded in the United States.[2] As the Georgia Tech Library redefines its role in the academic community for a new century, Carnegie's words continue to inspire new programs and services for a new generation of students and faculty: "A library outranks any other one thing a community can do to benefit its people. It is a never failing spring in the desert."[3]

Just as Carnegie libraries promoted reading and writing literacy in the 20th century, the Georgia Tech Library promotes data literacy in the 21st. For the past three years, Georgia Tech has been defining the changing role of higher education in the United States and continuing along its century-old path to becoming what Jan Youtie and Philip Shapira call "a knowledge hub to advance technological innovation and economic development in its region."[4]

The 2018 final report of the Georgia Tech Commission on Creating the Next in Education (CNE) includes a commitment to lifetime education: "a future for college not conceived solely...as a physical place one enters at a particular age and exits when a degree is completed but rather as a platform for an increasingly diverse population of learners."[5] To make good on its promise, the commission states that innovation will be required to close knowledge gaps, develop and pilot new products and services, and build new technological infrastructure. Library Next developed in parallel to the CNE vision. The Georgia Tech Library worked with brightspot strategy on a master plan for

services that informed a transformative design of the physical spaces in the renovation of the two mid-century library buildings. Through this planning effort, library faculty and staff learned how to engage the campus community in discussions about their needs and wants for the future. The result was a *Playbook*[6] that closely aligns institutional and user goals and aspirations. Read together, the CNE report and the *Playbook* connect the tradition of Georgia Tech and its library on a path forward that is both practical and inspirational. Data literacy and data science are core elements to Georgia Tech's future.

Learning Together

Supporting data science and data literacy has prompted the research library community to skill up. The Institute of Museum and Library Services (IMLS) has issued several recent grants to prepare librarians and other educators to teach data literacy, among them the ambitious Library Carpentry program expansion by the California Digital Library (CDL).[7] Similarly, IMLS funded the University of Michigan, Duke University, and the University of Southern California to study data visualization support and usage in libraries, which will culminate in the Visualizing the Future Symposia in 2020.[8] Earlier IMLS grants expanded the capacity of academic and research librarians to teach data literacy, such as at North Carolina State University Libraries, which supports a robust data visualization program on its campus. Additional IMLS-funded projects at Purdue University, Cornell University, University of Minnesota, and University of Oregon have all contributed to our collective insights and experience in developing strategies and techniques for preparing students for a data-centric future.

Most research universities have data science and data analytics programs directed toward students majoring in disciplines such as science, engineering, and business. Georgia Tech offers courses for undergraduates and graduate students in disciplinary data science, and recently added another master's degree in data analytics, on campus and online, organized into three tracks: analytical tools, business

analytics, and computational data. For the past year, Georgia Tech Professional Education has marketed online boot camps, lasting 24 weeks, in which students, who are generally working full time, build skills in several programming languages and tools. This program competes with the growing number of commercially offered boot camps, which are taught online and in person mainly in major cities in the US. These primarily post-graduate level programs are designed to meet the demand for data officers, data scientists, and related positions, primarily in industry.

On many campuses, however, students who are not in data-focused majors or intending to become data science subject-matter experts have few opportunities to learn such new tools and methodologies. Four Georgia Tech librarians, Ameet Doshi, Jay Forrest, Ximin Mi, and Alison Valk, are committed to filling this service gap at the institute. These librarians collaborate to offer non-credit courses, which the library markets to the campus community on its website. Although some classes are taught at the direct request of faculty, librarians create others based on feedback from the community.

Georgia Tech Offerings

In April 2019 alone, the four librarians leading this program offered 18 classes, some repeated, on a variety of data tools and technologies. In addition to Photoshop for Beginners, Illustrator and Adobe Creative Suite, Introduction to Zotero, and EndNoteX9, they offered Maya 3D Modeling, Introduction to R Studio, Grant Funding through Pivot, Python Twitter Scraping and Analysis, Intermediate R Studio: Visualization, Innovation Plus, and Using LaTeX for Advanced Mathematical Formulas. Some classes are online, and most are offered in the late afternoon or evening. Most require advance registration and all are free to participants.

Ximin Mi's goal is for students to effectively choose and apply the right data-visualization tools to tell the stories behind the data they are working with. With her, students learn through hands-on demonstrations with real-life data. In all "data viz" workshops, Ximin

and her graduate student assistants present interesting features of a data set, the right data format for a visualization design, and the right tool to format data. When moving into the visualization phase, Ximin emphasizes the logic behind the tool's functionality. Through design thinking and hands-on exercises, Ximin's expectation is that students will grow to understand effective design appropriate to the audience's social and cultural background.

Data viz-embedded classes are offered in several academic programs at Georgia Tech, including business, industrial and systems engineering, public policy, and computational media. Ximin points out that in embedded classes, instruction also features hands-on exercises to familiarize students with the tools. In these classes, the demonstration projects are designed with data the class instructor provides, or data about the topics of the class projects. Ximin and her team also offer lab hours and flexible follow-up consultation hours to review students' project designs and implementation.

After three semesters of instruction, the libraries' data viz program taught more than 1,500 learners, including students, faculty, and staff. Students are utilizing the tools they learned to create scholarship in innovative ways. In the spring 2019 semester, 30 students changed their class projects from print to digital. Not only did they save the cost of printing, but also their work is easily saved online for further study, sharing, or publishing in future. Ximin and her group keep detailed assessment of the effectiveness of their teaching so they continue to improve.

Ameet Doshi and Jay Forrest teach a popular coding workshop series on R and R Studio. R is an open source statistical software package commonly used for data analysis, visualization, and an increasing array of related scientific purposes. R Studio is an integrated development environment (IDE), which helps R be more user-friendly and expand its capacity. Because R is a popular open source tool, the breadth of R "packages," or niche system capabilities, as well as community support, continues to expand. These workshops have been well received by the

campus community, especially by the graduate students who comprise the majority of workshop attendees. Between fall 2018 and spring 2019, Ameet and Jay held 21 introductory or intermediate-level R workshops with approximately 450 total attendees. A post-workshop assessment indicates that 40% learned "a great deal" and 35% learned "a lot" by attending the R workshops.[9] With 74% of attendees indicating they had a positive learning experience, the data suggest that attendees benefit from this new instructional offering by the library. Drawing from 54 qualitative comments, Ameet and Jay observe that students are eager for the library to offer additional advanced and intermediate-level R workshops, training with more tools and techniques (Python, MatLab, text mining), as well as more emphasis on the syntax of R commands. Many students remarked that the 90-minute session "flew by," so a longer workshop may also be better aligned with graduate student needs.

R has become an increasingly prominent tool for data analysis across academic disciplines and within industry. Because it is open source software, the barrier to entry (cost and availability) is low, but many users still find an in-person introductory workshop helpful for building proficiency. The library has proactively recognized this trend and created a value-added service that reflects well upon the breadth of librarian capabilities, while also evolving the brand of the research library as crucial to academic and research success at Georgia Tech.

Alison Valk, whose specialty is teaching multimedia tools and techniques, believes that no matter the subject material or discipline, the ability to effectively communicate one's research to a broad audience is vital. A skill set including multimedia-based hardware and software to enhance communication skills is no longer optional, but imperative, and interest in this kind of training has grown over the last six years.

Multimedia includes video editing, visual or graphic design, audio editing, and website development. The manner in which information is transmitted and the communication channels used have the potential

to either obscure or enhance the message. Students and faculty alike can effectively communicate their research or data through compelling storytelling in video production or strong visual designs. Distilling a concept down to its key components or main takeaway creates the building blocks for any visually rich project. For example, Alison says, one must consider such decisions as choice of typography or color palettes. With the right choice a presenter can positively emphasize information, or with the wrong one confuse and distract the audience.

Since 2013, Alison has taught more than 500 course-integrated workshops reaching more than 12,000 students from a variety of academic programs including literature, media and communication, biological sciences, and business, as well as drop-in workshops reaching an additional 1,000 students. In the pre/post tests in Alison's classes, on average, students increased their pre-score with no preparation from 50% correct, to 89% correct on the post-test. The pre/post tests were standardized across topics.

In keeping with the overarching goals of Library Next, Alison is working to enhance the library's role in curriculum development. One example is a computational media research section led by Alison and Ximin Mi. Computational Media is a joint program between the Ivan Allen College of Liberal Arts and the College of Computing. Students in the program are heavily involved in the user experience. In this course, students work with emerging technologies to propose a research initiative utilizing available library resources and showcase their work as the semester concludes. Students are challenged to think creatively and critically in concert with the instructors who assess their progress.

To meet the growing demand for data literacy, several librarians at Georgia Tech submitted three proposals for a new program called Minimester Classes. The Commission on Creating the Next in Education proposed a matrix of minimester classes, which are short courses that can overlay the regular semester calendar. The purpose is to create short courses—one credit for five weeks—that could be used in a variety of ways, including future faculty training, whole-

person education, partnerships, and experiments in pedagogy. The proposals submitted by the librarians range from an updated version of a research methods class, which includes statistical literacy, to open data with R within the framework of open science, to an introduction to video editing with the enticing title of "Scare Me, Teach Me, or Make Me Laugh."

Conclusion

Within the context of Library Next, the data literacy program is expected to grow and change as the needs of students and faculty change. The biggest challenge the Georgia Tech Library faces is how to scale up the workshops and course offerings to meet the growing demand. This requires librarian-educators who are not only well versed in data literacy tools and technologies, but also skilled teachers and supervisors of teaching assistants. To develop data-literate students and faculty, the library will need to seek ways to broaden its reach. The library can do this through partnerships, generating and redirecting resources, and perhaps one day realizing its dream of having an intelligent agent, like Jill Watson, Georgia Tech's virtual TA for the Online Master's in Computer Science, who works alongside Ameet, Jay, Alison, and Ximin to train a new generation of data-literate students and faculty.

Endnotes

1. *Design a Data and Analytics Strategy*, ed. Andrew White (Stamford, CT: Gartner, 2019), 9, https://www.gartner.com/en/publications/data-analytics-strategy.

2. "A History of US Public Libraries," Digital Public Library of America, accessed June 26, 2019, https://dp.la/exhibitions/history-us-public-libraries/carnegie-libraries.

3. Andrew Carnegie Birthplace Museum (@carnegiemuseum), "'A library outranks any other one thing a community can do to benefit its people. It is a never failing spring in the desert' - Andrew Carnegie,"

Twitter, October 8, 2018, https://twitter.com/carnegiemuseum/status/1049269759060189189.

4. Jan Youtie and Philip Shapira, "Building an Innovation Hub: A Case Study of the Transformation of University Roles in Regional Technological and Economic Development," *Research Policy* 37, no. 8 (2008): 1188, https://EconPapers.repec.org/RePEc:eee:respol:v:37:y:2008:i:8:p:1188-1204.

5. Commission on Creating the Next in Education, *Deliberate Innovation, Lifetime Education: Final Report* (Atlanta, GA: Georgia Tech, April 2018): 5, http://www.provost.gatech.edu/sites/default/files/documents/deliberate_innovation_lifetime_education.pdf.

6. *User Research Project: Part 1: Research Report & Playbook* (Atlanta, GA: Georgia Tech, March 18, 2014), https://www.library.gatech.edu/sites/default/files/2019-01/part1.pdf.

7. John Chodacki, "Skills Training for Librarians: Expanding Library Carpentry," University of California Curation Center (UC3), November 6, 2017, https://uc3.cdlib.org/2017/11/06/skills-training-for-librarians-expanding-library-carpentry/.

8. Visualizing the Future Symposia website, accessed June 13, 2019, https://visualizingthefuture.github.io/.

9. n=81 respondents; 18% survey response rate.

To cite this article: Catherine Murray-Rust, Ameet Doshi, Jay Forrest, Ximin Mi, and Alison Valk. "Data Literacy as a Pathway to Data Science at Georgia Tech." *Research Library Issues*, no. 298 (2019): 6–15. https://doi.org/10.29242/rli.298.2.

New Collaboration for New Education: Libraries in the Moore-Sloan Data Science Environments

Jennifer Muilenburg, Research Data Services Librarian, University of Washington, and Visiting Program Officer, Association of Research Libraries

Judy Ruttenberg, Director, Scholars and Scholarship, Association of Research Libraries

In 2014, the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation partnered to invest $37.8 million across three US universities to build what they called Data Science Environments in order to "demonstrate how an institution-wide commitment to data science can deliver dramatic gains in scientific productivity and lead to significant new discoveries."[1] The Moore-Sloan Data Science Environments (MSDSE) were a five-year experiment "to better understand how best to bring together interdisciplinary people within institutional environments in order to provide them with the resources, freedom, and interconnected networks necessary for science to flourish." As that five-year experiment and its funding wind down, how can these and other research institutions—and their libraries—advance data science education? What did the MSDSE sites learn about the pedagogical relationship between data science methods and traditional disciplines, and intra-institutional partnerships? ARL staff spoke with key personnel at the three MSDSE sites: New York University (NYU), UC Berkeley, and the University of Washington (UW).

Of course, these three institutions are not unique in dedicating resources to data science; many other universities in Canada, Europe, and the United Kingdom have also launched degree and research programs focused on data science and analytics.[2] This grant, however, structured as it was across three institutions, provided a unique opportunity to assess each university's goals against the outcomes achieved. This article highlights not only their individual learning, but also draws overarching conclusions of value to all research libraries engaged with data science or looking for a pathway to doing so.

Common Themes and Distinctive Paths

Each MSDSE institution focused on interdisciplinarity and collaboration—attributes that data science fosters and requires. Each also initially established or built up a data science education program outside of existing departments (but in collaboration with them) in order to encourage institution-wide cohesion. In addition to local endeavors, the three institutional groups also came together frequently, through focused meetings on a particular topic, and in an annual summit to talk about research, education, careers, and the progress of the grant itself.

There were core themes, goals, and structure to the three-site design. Each institution participated in six cross-institutional working groups focused on recognized challenges to data-driven science, including (1) careers, (2) education and training, (3) tools and software, (4) reproducibility and open science, (5) working spaces and culture, and (6) data science studies. NYU, UC Berkeley and UW approached curriculum-building differently—some starting with graduate and others with undergraduate courses—and each emphasized the six themes to greater or lesser degree. NYU's focus on openness and reproducibility; UC Berkeley's development of Data 8, an entry-level course designed for students in any major; and UW's widely adopted data science "options" across the curriculum are examples of their individual distinction.

> "Core to the development of data science at the three MSDSE universities was a posture of non-competitiveness with existing programs and disciplines...and a belief that the set of tools and methods that comprise data science are critical to unlocking new discoveries across all research domains."

All three institutions have made many of their research and education materials open to the public. For example, the Data 8 course can be viewed online and freely reused at http://data8.org/, UW has three massively open online

courses available via Coursera at https://escience.washington.edu/education/mooc/, and all three institutions contributed to the open access book *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*, available at https://www.practicereproducibleresearch.org/.[3] Core to the development of data science at the three MSDSE universities was a posture of non-competitiveness with existing programs and disciplines, particularly in computer science, engineering, and statistics, and a belief that the set of tools and methods that comprise data science are critical to unlocking new discoveries across all research domains.

Abt Associates conducted an assessment of the grant in 2015 (at the midway point), and published its findings in 2019. Abt found that the MSDSE funding strategy was effective at increasing positive conditions for data-driven discovery at academic institutions, and noted that "the culture of partnership and experimentation adopted by the program facilitated mutual learning and growth."[4] The Abt report also stressed the importance of strong management and staffing, the success at establishing promising career tracks, and new programs that led to collaboration. In a secondary review of 20 other data science entities, Abt also found that there is no one way to create a data science center: culture, administration, physical space, funding, and people all combine in different ways to work toward supporting data-intensive science on campuses.

> "There is no one way to create a data science center: culture, administration, physical space, funding, and people all combine in different ways to work toward supporting data-intensive science on campuses."

Finally, a concentration on ethics was not a formal objective of the funding agencies providing MSDSE funding, but NYU, UC Berkeley, and UW have all developed an emphasis on data science and ethics, or data science and the public good. As the sites developed informal and formal curricula, and started working with students, it was immediately clear how important ethics education is to these efforts.

Research libraries are keenly interested in supporting data-intensive science under conditions of privacy, human subjects protection, and inclusivity. The commitment to ethics within the MSDSE and other academic data science programs bodes well for the library as a critical campus partner in the future of data science education.

NYU and the Center for Data Science

> As we transition to data-intensive scientific discovery, we have the opportunity to address these issues through software tools and practices that support the sharing, preservation, provenance tracking, and reproducibility of data, software, and scientific workflows.[5]
>
> —"Reproducibility and Open Science," Moore-Sloan Data Science Environments

Professor David W. Hogg (Physics and Data Science) served as executive director of the MSDSE at NYU in 2013–2015. Hogg has been involved with the project and the Center for Data Science (CDS) from the beginning. He emphasized interdisciplinarity as the main driver of the Moore-Sloan grant, and of the resulting transformation to data science education at NYU. While the NYU Libraries were not an original partner on the grant, Hogg acknowledged, he quickly learned that the library is "the most interdisciplinary place on campus," and credited the library for its influence on the CDS over the life of the MSDSE grant. In fact, in 2015, Vicky Steeves was appointed librarian for research data management and reproducibility, reporting jointly to the NYU Libraries and the CDS.

Steeves talked to ARL staff about how the norms and practices within data science—namely openness and reproducibility—are helping to transform disciplines that interface with CDS as well as the library. "Data science is a set of methods involving lots of computing power, lots of fast access to data, and the norms are open," said Steeves. Hogg also talked about openness as a cultural norm, and explained that by working with Moore and Sloan, along with NYU's lawyers and a legal team at UW, the CDS was able to construct an intellectual property

(IP) policy whereby any of the center's inventions would be released under an open license. UC Berkeley's Institute for Data Science (BIDS) and UW's eScience Institute also have working groups to advance reproducibility and open science. According to Steeves, the MSDSE has contributed to a wider conversation about openness at NYU in general.

NYU is currently launching an undergraduate Data Science course, which may lead ultimately to an undergraduate major. But Hogg pointed out that the MSDSE project did not set out to create formal degree programs at its three sites. Rather, it was used to create informal educational opportunities that would draw from and aid the disciplines. For example, one of the accomplishments Hogg is most proud of is CDS's deployment of Hack Weeks—informal opportunities to learn by doing in the company of experts. He was a co-author of "Hack Weeks as a Model for Data Science Education and Collaboration," along with colleagues from UC Berkeley and UW, which assesses this work at the three institutions in promoting interdisciplinarity, methods rigor, skill building, and positive attitudes toward open science in general.[6] Geohackweek, Oceanhackweek, and Waterhackweek were created and sponsored at UW's eScience Institute after the initial Astrohackweek offering.

The MSDSE provided all three site institutions the opportunity to experiment with different approaches to education and different partnership arrangements. While UC Berkeley and UW Libraries each housed the initial MSDSE sites, at NYU, the CDS started out in temporary space, and then worked closely with the libraries when it designed its new permanent space. The library, Hogg came to find out, had a deep understanding of how people across the campus and across disciplines use space—for writing, reading, interacting with technology, and collaborating. The library had experience with modularity and flexibility in space design that greatly informed CDS's new environment.

Likewise, the growth of data science throughout the university has influenced the library's collecting, such as purchasing more vendor-

produced data sets, responding to students' need for big data (for example, large social media feeds), and integrating APIs into their collection and discovery environment. The library has also ramped up its teaching of data literacy, and there is a new CDS-IT-library joint program called DS3 (data science and software services).

UC Berkeley and the Berkeley Institute for Data Science

> What faculty needed their future graduate students to know, no undergraduate program in the world taught.
>
> —David Culler, UC Berkeley

David Culler and Cathryn Carson were among the original co-investigators on the MSDSE grant to UC Berkeley. Culler is the interim dean for Data Sciences, and Carson is the faculty lead on UCB's undergraduate Data Science program. In 2012, when data science was still an emerging field, Culler, Carson, and their colleagues began an extensive inventory of related practice throughout the university, and began planning to launch the Berkeley Institute for Data Science (BIDS) in order to be competitive as an MSDSE site. Hundreds of faculty participated in the inventory that led to BIDS and eventually their successful proposal. "What faculty needed their future graduate students to know, no undergraduate program in the world taught," said Culler. The process, he added, "revealed a crack in the foundations of the research university. The world had changed and we had not gotten out ahead of it." Carson, a trained historian, was even more pointed: "UC Berkeley was ready when the Moore-Sloan funding was on the table." Undergraduates were moving in greater numbers to "real-world facing" computer science, statistics, and applied math programs. "Students were ready, and they were basically trying to do it on their own without the university," Carson said.

Once the university launched BIDS, based in Doe Library, as a new center of data science activity, other programs became involved, including the College of Engineering and the I School. With MSDSE investment, UC Berkeley built a data science curriculum from the first-year-student class up. "As we saw it, this was a moment to rethink

at a fundamental level what every educated person must know about quantitative reasoning: how to effectively understand, process and interpret information, to inform decisions in their professional and personal lives and as citizens of the world in the 21st century,"[7] said A. Paul Alivisatos in testimony to the US House of Representatives Science Committee in 2017. From the outset, the library was considered a key partner by virtue of their convening and collecting functions, enabling easy, broadly distributed opportunities for students to experiment with data and obtain peer consulting.

Data 8

The UC Berkeley Foundations of Data Science course combines three perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. It delves into social issues surrounding data analysis such as privacy and design.

—"Data 8: The Foundations of Data Science," http://data8.org/

Both Culler and Carson described the MSDSE experiment at UC Berkeley as transformative, and indicated that its enduring effects transcend data science and serve as a model for how faculty can come together around transdisciplinary research. Having established mechanisms for transdisciplinary practice as a new academic structure through BIDS, data science will become part of a larger engine at Berkeley. Culler observed that just as "Moore-Sloan wanted to create new career paths, UC Berkeley created new institutional paths." Carson hopes that "the new integrative structure is a good model for other [programs], that are conventionally schools and colleges—to build core strength [in those programs] and have solid connections outward." Might a new school or college of data science emerge at Berkeley or the other MSDSE sites? Perhaps, Carson conceded, but they would be

different: less competitive and with strong connections throughout the university.

In spring 2019, the UC Berkeley Libraries launched a Library Data Initiatives Plan. The plan will accelerate the library's focus and resources on data literacy, data acquisition, and housing data in a local repository. "For the Library, one constant goal is to demystify data science for the campus community, building new pipelines into the field from all directions."[8]

The University of Washington and the eScience Institute

> People come here [to the library] and are more likely to collaborate across disciplines than they might if they were all going to somebody's particular department.[9]
>
> —2017 Abt Associations Site Visit

ARL staff spoke with Sarah Stone, executive director of the University of Washington's eScience Institute—a campus unit that describes itself as UW's "hub of data-intensive discovery on campus."[10] Jennifer Muilenburg, a visiting program officer at ARL, has worked closely with Stone for the past five years as a data librarian and member of the eScience Institute Steering Committee at UW. Stone describes the library's role as central, and more aligned with the university's formal educational programs than the informal opportunities created by the MSDSE. They spoke in similar terms to UC Berkeley about the new program's non-competitiveness on campus, and its mission to integrate with all disciplines. These latter features are also characteristics of the library, which provide both physical space and expertise to the endeavor.

> Partnering with our libraries has been an important component of these efforts. For example, at UW several librarians who specialize in data management are data science fellows, and efforts toward a new institutional data repository, clarification of intellectual property rights concerning software and data, and a proposed Open

> Access Policy for faculty publications have all been informed by DSE efforts.[11]

Outside of a professional Master's Program in Data Science, geared toward retraining students returning to the workforce, the UW MSDSE did not set out to create a separate degree program. Rather, it took the approach that data science education should reside within the departments. At both the undergraduate and graduate levels, departments have ownership of the tools and techniques of data science that enhance their programs, and have the options and flexibility to implement what is most beneficial in their curricula. In contrast to UC Berkeley, UW started developing its data science curriculum at the doctoral level.

At UW, there was a pre-existing category referred to as a "transcriptable option," similar to a minor, that could be recognized on a student's transcript. This option, used at both the graduate and undergraduate level, allows a student to take a small number of data-focused courses in key data-science areas (such as statistics, machine learning, data management, data visualization, and ethics/privacy) in addition to their chosen major. Upon completion, the student receives a statement on his or her transcript—for example, "Bachelor of Science in Bioengineering, Data Science Option."

In looking for ways to expand data science techniques into the disciplines, UW has put much of its energy into its Incubator Program and Data Science for Social Good. Both programs match a researcher who is studying an existing problem with a group of data scientists who can help with the data-intensive part of the work. And both programs bring together a diverse set of participants who learn from one another, across disciplines, to help solve problems.

Conclusion

Each of the three MSDSE sites has had significant interaction and collaboration with their university library. At both NYU and UW, the newly designed data science centers were established in library

spaces. As stated in the Abt report, at UC Berkeley, by being in a library space, "The large investment in renovation and the choice of popular location signaled a commitment to data science by the administration and elevated the status of BIDS." At UW, "The open and modular layout of the space was intended to signal inclusivity to the university community and to foster collaboration." Libraries are seen as discipline-agnostic and welcoming to everyone, a feeling that carried over into the motivations of the data science centers themselves.

During one of the Abt visits to UW in 2017, one of the participants said, "One thing we talk a lot about and I think has been verified, is that having a neutral space on campus is important. We're not viewed as part of the computer sciences department or another department in particular. There's this Switzerland effect, when you are outside of the departmental silos. People come here and are more likely to collaborate across disciplines than they might if they were all going to somebody's particular department."

As part of the evaluation, Abt also looked at 17 other universities that have data science efforts underway, and found that each shared some characteristics with UC Berkeley, NYU, and UW, including the fact that nearly all were established in a location outside an established department. This helped signify inclusiveness and openness to collaboration. It's not known how many of these involved their library in the decision on location, but collaboration between data science entities and the libraries has proven beneficial for both parties in the MSDSE context, as well as the campus communities as a whole.

In *Creating Institutional Change in Data Science*, the authors emphasized the library's unique position with respect to the disciplines in hosting data science endeavors in their space:

> We have found that the ideal space is untethered to any department and is in a central location on campus where students and faculty habitually visit. In these respects, former library spaces are excellent choices. Our university libraries and librarians have been integral

partners to our efforts. Intellectually, libraries and librarians are in the information business.

Libraries at NYU, UC Berkeley, and UW, and many other research libraries in Canada and the US have incorporated data science education in their collecting and service frameworks. "Library spaces have been transformed—among other things—into campus centers for data science research, training, and services, with open floor plans and furnishings that are adaptable to a range of activities that promote and support data science research and learning.[12]

Endnotes

1. "About," Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/about/.

2. A partial list is available under "Other Data Science Centers" on "Environments," Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/environments/#others.

3. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*, ed. Justin Kitzes, Daniel Turek, and Fatma Deniz (Oakland, CA: University of California Press, 2018), https://www.practicereproducibleresearch.org/.

4. Luba Katz, *Evaluation of the Moore-Sloan Data Science Environments: Final Report*, (Abt Associates, 2019), iii, https://web.archive.org/web/20190501190246/http://msdse.org/files/ABT_MSDSE_Eval_Report_Feb2019.pdf.

5. "Reproducibility and Open Science," Moore-Sloan Data Science Environments, accessed June 26, 2019, http://msdse.org/themes/#reproducibility.

6. Daniela Huppenkothen, Anthony Arendt, David W. Hogg, Karthik Ram, Jacob T. VanderPlas, and Ariel Rokem, "Hack Weeks as a Model for Data Science Education and Collaboration," *Proceedings of the National Academy of Sciences* 115, no. 36 (Sep 2018): 8872–8877, https://doi.org/10.1073/pnas.1717196115.

7. *Hearing: STEM and Computer Science Education: Preparing the 21st Century Workforce, before the Research and Technology Subcommittee, House Committee on Science, Space, and Technology*, 115th Cong. 12 (2017) (testimony of A. Paul Alivisatos, executive vice chancellor and provost and vice chancellor for research, University of California, Berkeley), https://docs.house.gov/meetings/SY/SY15/20170726/106330/HHRG-115-SY15-Wstate-AlivisatosA-20170726.pdf.

8. Virgie Hoban, "Library Launches Initiative to Boost Data Science Expertise, Services at UC Berkeley," *UC Berkeley Library News*, April 25, 2019, https://news.lib.berkeley.edu/data-initiative.

9. Katz, *Evaluation of the Moore-Sloan Data Science Environments*, 27.

10. "About Us," University of Washington eScience Institute, accessed June 14, 2019, https://escience.washington.edu/about-us/.

11. *Creating Institutional Change in Data Science: The Moore-Sloan Data Science Environments: New York University, UC Berkeley, and the University of Washington*, Moore-Sloan Data Science Environments, accessed June 14, 2019, http://msdse.org/files/Creating_Institutional_Change.pdf.

12. *Creating Institutional Change in Data Science*.

To cite this article: Jennifer Muilenburg and Judy Ruttenberg. "New Collaboration for New Education: Libraries in the Moore-Sloan Data Science Environments." *Research Library Issues*, no. 298 (2019): 16–27. https://doi.org/10.29242/rli.298.3.

Building Capacity for Data Science with Help from our Friends

Cynthia Parr, Technical Information Specialist, National Agricultural Library, Agricultural Research Service, US Department of Agriculture

Susan McCarthy, Associate Director, Knowledge Services Division, National Agricultural Library, Agricultural Research Service, US Department of Agriculture

Introduction

Among the latest corporate and academic fads is "data science," that often ambiguously defined collection of data analytics activities that promises to take us all to the next level of efficiency and knowledge. Especially when combined with other buzzwords like "big data" and "open data," data science appears to be somewhere between the "peak of inflated expectations" and the "trough of disillusionment" in Gartner's hype cycle.[1] One might wonder how, and even why, a research library should dip its toes into these murky waters.

At the United States Department of Agriculture (USDA) National Agricultural Library (NAL), we believe that the answer to "why explore data science?" is that institutional experience with core data-science activities will inform the larger set of data and data management services the library performs.[2] Moreover, engaging information science students and library managers in data science projects builds capacity both for us and for the communities we serve. The answer to "how?" is "not alone." In this article, we describe how NAL, in collaboration with the University of Maryland College of Information Studies (UMD iSchool), the USDA Agricultural Research Service (ARS), and university librarians (the "friends" in the title), is using lean start-up methodology to enable us all to continue a long tradition of supporting agricultural knowledge generation, dissemination, and preservation.

At NAL, we consider data science to include, but not be limited to, the core analytical activities necessary for deriving insight from data. We recognize that data science activities are practiced not only by

those with a title "data scientist," but also by our data curators and even by the scientists who originally collected, analyzed, and then prepared their data for publication. How did we get to this place? The story begins with NAL and its data services, starting with data curation. We describe how we are developing and testing prototype services in data science and data management planning. In each data service discussion, we explain the key partnerships that have been instrumental to iterative service development.

History of NAL's Data Services

NAL has a deep history. Founded in 1862 by the same legislation that founded the Department of Agriculture, NAL's mission has been to acquire, describe, provide access to, and preserve the literature of agriculture and related sciences. The library has grown to manage nearly 2.5 million physical volumes, as well as digital collections, information centers, and a thriving reference service. NAL currently is part of the USDA Agricultural Research Service, USDA's agency for intramural research, which employs nearly 2,000 scientists. It has long had a special relationship with the United States land-grant university system and, together with four other libraries in that system, formed the Agriculture Network Information Collaborative (AgNIC) in 1995. AgNIC has since grown to more than 50 participating groups,[3] with partners in both Canada and Mexico along with the United States, and taken on a number of data-related initiatives that are further described by Erica Johns in this issue of *RLI*.

In 2012, NAL created the Knowledge Services Division to support the advanced research data needs of agricultural researchers and the broader community. However, "data needs" encompass a very broad range of requirements. Cooper et al.[4] and Hanscom et al.[5] found that the needs range from guidance on data management, to data storage infrastructure, to analytical skill building. Moreover, data policy and practice are rapidly changing, and there are calls for improved infrastructure to serve the needs of emerging fields such as precision agriculture.[6]

The first efforts at NAL to build infrastructure and services to support data analysis and management included the i5K workspace@NAL[7] and the Life Cycle Assessment Commons.[8] These two projects were designed to support users in specific disciplines: insect genomics and cradle-to-grave modeling of the inputs and outputs of agricultural production processes. Later, NAL responded to calls for US public access to data as a product of federally funded research by initiating the Ag Data Commons,[9] a general catalog and repository for open research data funded by USDA.

Lean Start-up Methodology

Having made these investments in software product infrastructure and the associated data curation services, we next turned our attention to complementary services: data science services and data management planning. With these two newest services, and particularly with its data science services, NAL's Knowledge Services Division is beginning to apply lean start-up methodology.[10] In this method, products and services undergo iterative cycles of build–measure–learn, followed by either pivoting or perseverance. With specific hypotheses in mind, one quickly creates a minimum viable product (MVP) and tests it with stakeholders. For these services, we are testing hypotheses about whether the services are needed, who can best provide these services, and how to develop the appropriate capacity and resources to deliver these services effectively. Essential to our application of this method is partnership with land-grant university libraries, information schools, and researchers at both universities and USDA ARS who are our key stakeholders. While our cycles are not yet as rapid as is recommended, we are constantly refining our methodology as we go.

Defining Data Science within the Full Data Life Cycle

As of this writing, data science is defined in Wikipedia as "a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data."[11] Donoho offers six buckets of data science activities: (1) Data Exploration and Preparation, (2) Data

Representation and Transformation, (3) Computing with Data, (4) Data Modeling, (5) Data Visualization and Presentation, and (6) Science about Data Science.[12] Gartner lists 33 relevant hot topics ranging from crowdsourcing and ethics to machine learning and visualization.[13]

In most paradigms, the activities of planning for data management, data collection, and making data available (for data science or other purposes) are not included in the definition of data science. However, the lines cannot be clear when choices made during planning, data collection, or data sharing will impact any of Donoho's six buckets. For example, the availability of data and its discoverability and suitability for data exploration and modeling may rely on tidy data[14] and standards-driven metadata APIs. Schutt and O'Neil[15] took a data-driven approach to capturing the nature of data science, in other words, asking data scientists what they do in their work, and concluded that (1) data science is a real thing worth teaching, and (2) in industry, that includes not only programming and statistical skills, but also basic data wrangling, driving curiosity, and the ability to communicate effectively.

"Schutt and O'Neil took a data-driven approach to capturing the nature of data science, in other words, asking data scientists what they do in their work, and concluded that 1) data science is a real thing worth teaching, and 2) in industry, that includes not only programming and statistical skills, but also basic data wrangling, driving curiosity, and the ability to communicate effectively."

Data Curation Services

Soon after NAL began developing services in data curation and preservation, we initiated cooperative agreements with the University of Maryland iSchool. Under these agreements, we have sponsored four master's of library and information science students to work with us on digital data curation projects and two postdoctoral scholars to assess our readiness for trusted data repository certification and make

recommendations for workflow process improvement and digital preservation. These fellowships (part of a larger cohort throughout the library) provided practical, on-the-ground work experience for these individuals, enabling them to build on their education and prepare for their careers. They also provided NAL with our first data collection policy documents and protocols for curation at the Ag Data Commons. And they helped conduct an important assessment of federal responses to US public access policy.[16]

The partnership with UMD taught us valuable lessons about how to offer data curation services. We quickly realized that a combination of data set self-submission and expert data curation would be necessary to scale our services without a reduction in quality. We are now preparing for a distributed curation model.[17] Expertise in metadata standards and tools is critical and requires different knowledge than traditional library metadata. We realized it was important to include our data curators as stakeholders in agile software development—they are critical links between end users and developers, and they are the power users of our software. One former student is now one of our professional data curators, and two others are working in related positions (including management) elsewhere.

Data Science Services

In 2017, master's of information management students from University of Maryland joined us to launch the first iteration of an experiment in providing data science services. These students considered themselves data scientists. We tested the hypothesis that an information school such as UMD can adequately prepare students to apply their skills to real-world problems. We tested the idea that data scientists need not have domain knowledge to conduct their work. Finally, we tested the idea that staff, such as that in a library, could supervise data science work of individuals working on distinct projects and produce results within a year. In collaboration with a high school intern, one student built an impact tracking dashboard prototype for the Ag Data Commons—we later refactored it and it is still in use.[18] The same

student conducted sentiment analysis on historical dietary guidelines documents as a test of how to apply these methods to the digital library collections. Another student used data from the Global Biodiversity Information Facility[19] to find geographic and temporal patterns in agricultural biodiversity data reporting.

In 2018, when University of Maryland created a new Data Science specialization within their undergraduate information science degree, we learned that students need not be graduate students to have the necessary skills. Our fellowship could offer juniors and seniors an opportunity to practice data science in a meaningful context. A college senior, our first such student, developed scripts to automate the process of checking for public access policy compliance and establish a baseline for future tracking.[20]

“We learned that students need not be graduate students to have the necessary skills. Our fellowship could offer juniors and seniors an opportunity to practice data science in a meaningful context. A college senior, our first such student, developed scripts to automate the process of checking for public access policy compliance and establish a baseline for future tracking.”

Later in 2018, in a third iteration, we began testing the hypothesis that we could offer data science services to a client outside the library, staffing that service cost-effectively by hiring recent data science graduates as contractors, and managing that service using the same model developed with Maryland iSchool Fellows. Both the fellows and the contracted data scientists now meet regularly with a growing group of library staff who, in the course of their jobs, interface with or do data analytics. This interest group includes software developers, web analytics experts, data policy analysts, and indexers who are developing the NAL thesaurus.[21] Capacity is growing and self-reinforcing.

We developed a simple project template for data science projects that guides young data scientists to work closely with their “clients” to

establish shared goals and questions and propose their own technical approaches before they get too far into the exploration, insight, and predictive phases of their work. These project proposals are also discussed with the data science interest group. We learned that some early-career data scientists can quickly learn how to use a high-performance computing system on their own, but that managers need to work hard to encourage them to work with domain-savvy clients, and to cultivate them to draw their own insights from data. Finally, we can confirm what others describe: data scientists work best in small teams (teams of two plus a manager worked for us) to which members bring complementary skills.[22] If the library continues to offer this service, it will be important to provide an environment in which teams can tackle projects together and regularly share their work with interested colleagues.

Taken together, these experiences have shaped a functioning data science program that serves the needs of clients both within the library and with the Agricultural Research Service's Office of National Programs. We used the experience we gained managing data science fellows to work with ARS and other USDA agencies to prepare a menu of data science position description snippets that could be used to effectively tailor job postings or statements of work that would be effective in recruiting talent. These position description snippets were tested when recruiting the data science contractors, and they are designed to be included in any new position inside or outside the library where data science skills are needed.

While the two teams of data curators and data scientists typically work separately, the data scientists have provided valuable feedback to the features of our data catalog and of our APIs. They represent an emerging community of library users—developers and analysts who will use library products and services to do analytics and derive new knowledge.

Data Management Planning Services

The final service emerging at NAL is a data management planning service. Although planning for data management has long been recognized as an essential part of the data life cycle,[23] many researchers in agriculture have only recently encountered funder requirements for formal data management plans (DMPs) to be submitted with proposals (for example, the National Institute of Food and Agriculture[24]).

As we develop this service, the hypotheses to test are as follows:

1. Many agricultural researchers need a review service, but some of them have access to such services through their institutional libraries.
2. Data curators are well positioned and have the capacity to perform this service.
3. The service will improve the quality of data management plans and improve NAL's ability to anticipate the infrastructure it must supply as part of the plans.

Our initial incarnation of the service requests that researchers prepare a draft according to the instructions provided by their funder, and submit that draft to us by email. We then circulate the draft among data curators and relevant subject-matter experts, and provide recommended changes and comments by email. In addition to assistance with specific funder instructions, we provide guidance (with examples) on the NAL website.[25]

This service is being developed in close cooperation with University of Maryland, other AgNIC members, and the Office of National Programs in USDA ARS. Together we are developing and delivering educational webinars tailored to agricultural researchers. As with data curation and data science, data management planning is recognized as part of the data life cycle and all three activities benefit from consideration of the others. For example, a recent data science project was to use natural language processing to establish a baseline in DMP content in a pilot round of USDA ARS project plans. By reviewing DMPs,

our data curators are able to better understand the specific needs of disciplinary research projects before they are funded. They can establish relationships with project personnel that will in coming years bear fruit in well-described, highly accessible, data submissions to the Ag Data Commons resulting from the funded work.

Recent analysis by Smale et al. finds that supposed benefits of DMPs have not yet been supported by evidence.[26] However, they suggest that placing DMP preparation into a researcher-centric context of education and sound program management may be useful. As this experimental service progresses we can determine if this is the case.

Conclusion

In summary, we have two primary conclusions:

1. It is important for a library like NAL to team up with universities and science organizations to explore new services and create talent pipelines.
2. It is productive to explore new services in the context of the entire data life cycle.

These efforts allow us to gather the requirements for future programs that may or may not need to be housed in the library. In closing, let us consider the benefit to the library of developing these services here rather than being embedded, for example, in an IT services organization or a scientific department.

Developing data science services in a research library—in the context of other data management services—provides an excellent way to attract and develop technical talent to the field of library and information science. It may even be the case that data science can help increase diversity in the library pipeline. People of many backgrounds, genders, and races may be excited by a career in innovative science and technology and find satisfaction in practicing those skills in a mission-driven service context rather than profit-driven organization. Finally, given the uncertain future of traditional scholarly publishing, the role

of research libraries must change to support scholarly activity in new ways. Delivering data science services can be the next iteration in a progression first noted in the 1894 *Yearbook of Agriculture*:

> A reading room has been arranged and increased facilities provided for the convenience of investigators. The Library has been made in this manner a working laboratory instead of a miscellaneous storehouse.[27]

Endnotes

1. "Gartner Hype Cycle," Gartner, accessed June 14, 2019, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.

2. Vicki Boykis, "Data Science is Different Now," *Data, Tech, and Sometimes Nutella*, February 13, 2019, http://veekaybee.github.io/2019/02/13/data-science-is-different/.

3. Livia Olsen, Julie Kelly, and Noël Kopriva, "The Agriculture Network Information Collaborative (AgNIC): Building on the Past, Looking to the Future," *Library Trends* 65, no. 3 (2017): 279–292, https://doi.org/10.1353/lib.2017.0002.

4. Danielle Cooper et al., *Supporting the Changing Research Practices of Agriculture Scholars* (New York: Ithaka S+R, June 7, 2017), https://doi.org/10.18665/sr.303663.

5. Scott Hanscom, Adam Kreisberg, Emily Marsh, and Cynthia Parr, *Agricultural Researcher Support Services: A Study Conducted by the National Agricultural Library in Cooperation with Ithaka S+R*, (Washington, DC: USDA, June 2017), https://www.nal.usda.gov/sites/default/files/ithaka_report_with_logo_revised.pdf.

6. Sylvie Brouder, Alison Eagle, Naomi Fukagawa, John McNamara, Seth Murray, Cynthia Parr, Nicolas Tremblay, *Enabling Open-source Data Networks in Public Agricultural Research* (Ames, IA: Council for Agricultural Science and Technology, 2019), https://bit.ly/2EW0stL.

7. Monica Poelchau, Christopher Childers, Gary Moore, Vijaya Tsavatapalli, Jay Evans, Chien-Yueh Lee, Han Lin, Jun-Wei Lin, and Kevin Hackett, "The i5k Workspace@NAL—Enabling Genomic Data Access, Visualization and Curation of Arthropod Genomes," *Nucleic Acids Research* 43, no. D1 (January 28, 2015): D714–D719, https://doi.org/10.1093/nar/gku983.

8. "LCA Commons," Ag Data Commons, USDA National Agricultural Library, 2015, https://doi.org/10.15482/USDA.ADC/1173236.

9. "Ag Data Commons," re3data.org - Registry of Research Data Repositories, editing status November 13, 2018, accessed June 14, 2019, https://doi.org/10.17616/R3G051.

10. Eric Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (New York: Crown Books, 2011).

11. Wikipedia, s.v. "Data science," last edited June 8, 2019, https://en.wikipedia.org/wiki/Data_science.

12. David Donoho, "50 Years of Data Science," MIT Computer Science and Artificial Intelligence Laboratory course server, September 18, 2015, http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

13. Peter Krensky and Jim Hare, *Hype Cycle for Data Science and Machine Learning* (Stamford, CT: Gartner, July 23, 2018), https://www.gartner.com/en/documents/3883664.

14. Hadley Wickham, "Tidy Data," *Journal of Statistical Software* 59, no. 10 (2014): 1–23, https://doi.org/10.18637/jss.v059.i10.

15. Cathy O'Neil and Rachel Schutt, *Doing Data Science: Straight Talk from the Frontline* (Sebastopol, CA: O'Reilly Media, 2013).

16. Adam Kriesberg, Kerry Huller, Ricardo Punzalan, and Cynthia Parr, "An Analysis of Federal Policy on Public Access to Scientific Research

Data," *Data Science Journal* 16 (2017): 27, https://doi.org/10.5334/dsj-2017-027.

17. Lisa R. Johnston, Jake R. Carlson, Patricia Hswe, Cynthia Hudson-Vitale, Heidi Imker, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart, "Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services," *Journal of eScience Librarianship* 6, no. 1 (2017): e1102, https://doi.org/10.7191/jeslib.2017.1102.

18. "Ag Data Commons Metrics," USDA National Agricultural Library, accessed June 14, 2019, https://data.nal.usda.gov/ag-data-commons-metrics.

19. Global Biodiversity Information Facility, accessed June 14, 2019, https://www.gbif.org/.

20. "Public Access (PA) Plans of U.S. Federal Agencies," CENDI, accessed June 26, 2019, https://cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html.

21. National Agricultural Library, USDA, Inter-American Institute for Cooperation on Agriculture, and Agriculture Information and Documentation Service of the Americas (SIDALC), "Thesaurus and Glossary Home," USDA National Agricultural Library, accessed June 14, 2019, https://doi.org/10.15482/USDA.ADC/1503889.

22. Gendron et al., "Data Transforming Governmental Data Science Teams in the Future," chap. 13 in *Federal Data Science*, ed. F. Batarsey and Yang (Academic Press, 2018), 211–221.

23. Line Pouchard, "Revisiting the Data Lifecycle with Big Data Curation," *International Journal of Digital Curation* 10, no. 2 (2015): 176–192, https://doi.org/10.2218/ijdc.v10i2.342.

24. "Data Management Plan for NIFA-Funded Research, Education, and Extension Projects," USDA National Institute of Food and Agriculture, October 29, 2018, https://nifa.usda.gov/resource/data-management-plan-nifa-funded-research-projects.

25. "Data Management Resources," USDA National Agricultural Library, accessed June 14, 2019, https://www.nal.usda.gov/ks/data-management-resources.

26. Nicholas Smale, Kathryn Unsworth, Gareth Denyer, and Daniel Barr, "The History, Advocacy and Efficacy of Data Management Plans," bioRxiv preprint, posted October 17, 2018, https://doi.org/10.1101/443499.

27. United States Department of Agriculture, *Yearbook of Agriculture, 1894* (Washington, DC: Government Printing Office, 1895), 61, https://www.biodiversitylibrary.org/item/96795#page/9/mode/1up.

To cite this article: Cynthia Parr and Susan McCarthy. "Building Capacity for Data Science with Help from our Friends." *Research Library Issues*, no. 298 (2019): 28–40. https://doi.org/10.29242/rli.298.4.

Correction, August 1, 2019: *On page 29, the year in which the National Agricultural Library and the US Department of Agriculture were founded was identified incorrectly in an earlier version of this article. The year of founding was 1862, not 1863.*

The AgNIC Data Working Group: University Collaboration with the National Agricultural Library

Erica M. Johns, Head of Research Services and Scholarly Engagement, Albert R. Mann Library, Cornell University

Cynthia Parr, Technical Information Specialist, National Agricultural Library, Agricultural Research Service, US Department of Agriculture

The Agriculture Network Information Collaborative (AgNIC) is a group of member institutions "dedicated to enhancing collective information and services...for all those seeking agricultural information."[1] AgNIC member institutions are predominantly US land-grant universities,[2] but include Agriculture and Agri-Food Canada, and the collaborative has established strategic partnerships with the International Association of Agriculture Information Specialists (IAALD), the International Food Policy Research Institute (IFPRI), and the United States Agricultural Information Network (USAIN).[3] AgNIC is a project-based initiative that accomplishes tasks through working groups related to established interest areas, such as agriculture data.[4]

The AgNIC Data Working Group (DWG) started in 2017 as a response to the DataRefuge movement, a campaign to safeguard federal data (particularly environmental and climate data) from politically motivated removal.[5] The DWG was concerned about continued, long-term access to data sets from the United States Department of Agriculture (USDA), in addition to the climate and environmental data that DataRescue events[6] focused on saving. These open, public-facing events, also called "archive-a-thons," were highly mobilized and efficient, with community-sourced efforts conducted at a speed that allowed for fairly comprehensive and fast scraping of endangered government data. Before an AgNIC effort could advance towards rescuing the researcher-identified, highly impactful data sets, the working group found that most data sets were already cataloged and saved within the archive-a-thon framework. Despite the reduced urgency of participating in relevant data rescue activities, however, the

DWG realized a need for ongoing conversations around the diverse makeup and management of agriculture data.

The DWG initially wanted to capitalize on existing studies by agriculture researchers, attempting to mine data from the Ithaka S&R interview transcripts on agriculture data publishing venues.[7] Unfortunately, the varied institutional review board contracts from the 19 participating universities made this effort overly cumbersome. During the mining process, however, the National Agricultural Library (NAL) reviewed both the repositories that accept agricultural data, and agricultural journals' data-sharing policies, to determine whether there was an obvious home for agriculture data.[8] NAL found that while many sub-disciplinary repositories exist for agriculture data, there was not one clear, all-encompassing, agriculture data repository. The DWG hopes that NAL's Ag Data Commons can become this repository, as it is already cataloging all USDA-funded data and accepting extramurally funded USDA data into its repository. Currently, Ag Data Commons also considers accepting agriculture data funded externally by state offices and other sources on a case-by-case basis. Ag Data Commons aims to "foster innovative data re-use, integration, and visualization to support bigger, better science and policy;" and the DWG's support for Ag Data Commons includes outreach to university faculty about its webinar series and policies for data inclusion and submission—all in an effort to build Ag Data Commons content, educate researchers, and facilitate data sharing.

"Ag Data Commons aims to "foster innovative data re-use, integration, and visualization to support bigger, better science and policy;" and the DWG's support for Ag Data Commons includes outreach to university faculty about its webinar series and policies for data inclusion and submission."

In late 2017, DWG members from the University of Maryland created a survey to understand the data management practices of diverse and

interdisciplinary agriculture researchers internationally, working with the DWG to review the survey instrument and assist with its distribution.[9] The University of Maryland team also worked with NAL to create USDA data management plan (DMP) guidance on the NAL website, including encouragement for researchers to make use of data management planning services offered through their university libraries.[10]

Another subset of the DWG organized a workshop in 2018 called "Driving Innovation through Data in Agriculture (DIDAg),"[11] funded by the USDA National Institute of Food and Agriculture (NIFA). DIDAg, a Food and Agriculture Cyberinformatics and Tools (FACT) Initiative workshop for researchers and librarians, focused on data management and publication for agricultural economics and dairy agroecosystems. The workshop addressed long-term goals and supporting objectives for the two research domains, including the following:

- Shared understanding of existing policies and resources related to public access to data and agricultural data management
- Clear expectations for research data management and publication in selected research domains, both so that researchers can plan for them and so that information professionals can support them
- Improved cyberinfrastructure, training materials, and business models
- A road map for supporting the next generation of data-intensive research in agricultural economics and dairy agroecosystems

Most recently, the DWG has created a reviewer checklist associated with USDA DMP guidelines. Some USDA program officers have vetted the checklist, and the DWG hopes that the checklist will be provided to grant proposal reviewers as a resource, much like the data management planning guidance is referenced in grant application guidelines, in addition to being used as a training tool for researchers.[12]

Next steps for the AgNIC DWG is a follow-up to the DIDAg workshop in late summer 2019, issuing best practices for data management within the sub-disciplines of agriculture, and continued training for agriculture librarians in research data management.

Endnotes

1. "About AgNIC," Agriculture Network Information Collaborative, accessed June 15, 2019, https://www.agnic.org/about.

2. "List of Partners," Agriculture Network Information Collaborative, accessed June 15, 2019, https://www.agnic.org/partners.

3. "AgNIC Strategic Partnerships," Agriculture Network Information Collaborative, accessed June 15, 2019, https://www.agnic.org/sites/default/files/partnerships.pdf.

4. "Check Us Out and Get Involved!," Agriculture Network Information Collaborative, accessed June 15, 2019, https://www.agnic.org/sites/default/files/AgNIC_JoinUS.pdf.

5. "About," DataRefuge, accessed June 15, 2019, https://www.datarefuge.org/about.

6. "Archiving Data," Environmental Data & Governance Initiative, accessed June 15, 2019, https://envirodatagov.org/archiving/.

7. Danielle Cooper et al., *Supporting the Changing Research Practices of Agriculture Scholars* (New York: Ithaka S+R, June 7, 2017), https://doi.org/10.18665/sr.303663.

8. Cynthia Parr, Erin Antognoli, and Jonathan Sears, "How Agricultural Researchers Share Their Data: A Landscape Inventory," *Biodiversity Information Science and Standards* 1, e20434 (2017), https://doi.org/10.3897/tdwgproceedings.1.20434.

9. Adam Kriesberg, Richard Punzalan, and Morgan G. Daniels, "Data Practices of Agricultural Scientists: Global Insights" (presented at the USAIN 16th Biennial Conference, Pullman, WA, May 14, 2018).

10. "Guidelines for Data Management Planning," USDA National Agricultural Library, accessed June 15, 2019, https://www.nal.usda.gov/ks/guidelines-data-management-planning.

11 "Driving Innovation through Data in Agriculture (DIDAg)," USDA National Agricultural Library, accessed June 15, 2019, https://www.nal.usda.gov/ks/didag2018.

12. "Guidelines for Data Management Planning."

To cite this article: Erica M. Johns and Cynthia Parr. "The AgNIC Data Working Group: University Collaboration with the National Agricultural Library." *Research Library Issues*, no. 298 (2019): 41–45. https://doi.org/10.29242/rli.298.5.

ASSOCIATION OF RESEARCH LIBRARIES

Research Library Issues

Research Library Issues (*RLI*) focuses on current and emerging topics that are strategically important to research libraries. The articles explore issues, share information, pose critical questions, and provide examples. Suggestions for potential themes, articles, and authors are welcome. Please submit suggestions via this online form.

ISSN 1947-4911 https://doi.org/10.29242/rli

Current and back issues are available on the ARL Digital Publications website, publications.arl.org/rli. The website is also where you may sign up for alerts to new releases of *Research Library Issues*.

Association of Research Libraries
21 Dupont Circle, NW
Suite 800
Washington, DC 20036
T 202.296.2296
F 202.872.0884

ARL.org
pubs@arl.org