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.*