JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Mapping the Landscape of Research Data: How *JLSC* Contributors View this Rapidly Emerging Terrain

Gail P. Clement, Lisa R. Schiff

Clement, G. P., & Schiff, L. R. (2015). Mapping the Landscape of Research Data: How JLSC Contributors View this Rapidly Emerging Terrain. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1279. http://dx.doi.org/10.7710/2162-3309.1279

ISSN 2162-3309 10.7710/2162-3309.1279

EDITORIAL

Mapping the Landscape of Research Data: How *JLSC* Contributors View this Rapidly Emerging Terrain

Gail P. Clement & Lisa R. Schiff
Issue Editors

Pronouncements that research data have arrived as first class objects of scholarly communication are increasingly common, reflecting a growing consensus that the basic building blocks of knowledge (data, software, algorithms, visualizations, and other outputs of the research process) warrant the same degree of attention as the research papers that synthesize and interpret those raw artifacts. In 1997, the US National Research Council advised that "full and open access to scientific data should be adopted as the international norm for the exchange of scientific data derived from research."[1] The US President's Office of Science and Technology Policy now requires that "the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data."[2] The Research Councils UK Common Principles on Data Policy "set expectations for the systematic and

[1] National Research Council (1997). *Bits of power: Issues in global access to scientific data*. Washington DC: National Academies Press.

[2] United States White House Office of Science and Technology Policy (2013, Februar 22). *Expanding public access to the results of federally funded research*. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

About the editors

Gail P. Clement is Head of Research Services at the CalTech Library System (gclement@library.caltech.edu); Lisa R. Schiff is the Technical Lead for the California Digital Library's Publishing Group (Lisa.Schiff@ucop.edu).

routine management and sharing of research data"[3] and the European Research Council's Data Archiving Policy "encourages deposition in Open Access archives."[4] Other open data policies emerging from around the world articulate comparable visions.[5]

At the same time, international, multi-disciplinary coalitions of stakeholders in the research enterprise convey expectations for careful curation of research data: detailed documentation, formal publication and citation, and long term preservation for indefinite reuse. Chief among these groups are the International Council for Science's Committee on Data for Science and Technology and its Task Group on Data Citation Standards and Practices[6]; the Research Data Alliance's Publishing Data Interest Group and its working groups[7]; the International Federation of Data Organizations' Data Citation Standards and Practices[8]; the Force11 Data Citation Synthesis Group[9]; and DataCite, with its mission to "develop and support methods to locate, identify and cite data and other research objects."[10] Their policy recommendations, proposed principles, and codes of best practice reflect the best thinking of a wide range of open data proponents: government agencies, research funders, publishers, university administrators, legal experts, learned societies, data centers, and libraries.

Librarians, as experienced knowledge workers, and libraries, as long-established knowledge repositories, hold a great stake in establishing themselves as vital players in the research data management enterprise. The Association of Research Libraries (ARL) considers library support for research data an essential strategic direction for the profession, equating it with

[3] Research Councils UK. (2015). *Guidance on best practice in the management of research data*. Retrieved from http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKCommonPrinciplesonDataPolicy.pdf

[4] European Research Council. (2007, December 17). *ERC Scientific Council guidelines for open access*. Retrieved from http://erc.europa.eu/sites/default/files/document/file/erc_scc_guidelines_open_access.pdf

[5] Policies are collected by and made available through the Sherpa/Juliet database of Research Funders Open Access Policies, http://www.sherpa.ac.uk/juliet/

[6] International Council for Science, Committee on Data for Science and Technology, Data Citation Standards and Practices Task Group, http://www.codata.info/taskgroups/TGdatacitation/

[7] https://rd-alliance.org/groups/rdawds-publishing-data-ig.html

[8] International Federation of Data Organizations. (2014, March 13). *Policies for sharing research data in social sciences and humanities*. Retrieved from http://ifdo.org/wordpress/wp-content/uploads/2014/04/ifdo_fact.pdf

[9] Force11, Data Citation Synthesis Group. (2014). *Joint declaration of data citation principles*. Martone M. (ed.) San Diego CA: FORCE11. Retrieved from https://www.force11.org/datacitation

[10] DataCite, "What do we do?," https://www.datacite.org/about-datacite/what-do-we-do

"next-generation librarianship."[11] ARL's 2013 assessment of research data management services, published as *Spec Kit 334*,[12] indicates that the majority of North American research libraries are "expanding or adopting new research data services" yet "are still in the early stages of development and implementation." The ARL study documents increased investment in new library-based data resources, services, programs, and data-savvy staff to lead these initiatives, reflecting determination to preserve the research library's niche in the evolving "knowledge infrastructure" surrounding research data.

The term "knowledge infrastructure" has been employed by leading information and research scientists to describe the system in which knowledge is created, shared, and assessed.[13] The cross-disciplinary Knowledge Infrastructures Research Group (sponsored by the US National Science Foundation and the Sloan Foundation) explains that knowledge production, management, dissemination, and acceptance or dispute occur through "robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds […] [they] include individuals, organizations, routines, shared norms, and practices."[14] The most familiar knowledge infrastructure in today's research enterprise is the refereed publishing system centered around the peer reviewed journal in the sciences and the scholarly monograph in the humanities. These centuries-old vehicles for registration, quality control, and credit have formed the basis of the scholarly record.

Now enter research data. Adherents to the traditional scholarly publishing ecosystem contemplate whether the established "knowledge infrastructure" it represents is extensible to the newly-recognized and less well-understood research data object. Research data represent artifacts from the entire research life cycle, not just the final outcomes of scholarly inquiry. They are generally less fixed than the published literature (though of course that published literature is also mutable) and they can grow and change over time and with reuse. Their legal status is often less well understood. Normative practices for crediting data contributors are still in the formative stages of community deliberation, though the

[11] Hswe, P., & Holt A. (n.d.). *NSF guide: A new leadership role for libraries*. Association of Research Libraries. Retrieved from http://www.arl.org/focus-areas/e-research/data-access-management-and-sharing/nsf-data-sharing-policy/241-a-new-leadership-role-for-libraries#.VbVMYf7JDCM

[12] Fearon, D., Gunia, B., Lake, S., Pralle, B. E., & Sallans, A. L. (2013). *Research data management services, SPEC Kit 334*. Washington, D.C.: Association of Research Libraries.

[13] Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge infrastructures: Intellectual frameworks and research challenges*. NSF/U Michigan. Retrieved from http://pne.people.si.umich.edu/PDF/Edwards_etal_2013_Knowledge_Infrastructures.pdf

[14] Ibid., page 13

need and expectation to do so is increasingly encouraged and expected across disciplines. In considering these and other characteristics that distinguish research data from the artifacts of the scholarly literature, librarians may find the writing of UCLA Information School Professor Christine Borgman particularly illuminating. In *Big Data, Little Data, No Data,*[15] Dr. Borgman replaces the term "knowledge infrastructure" with the richer, more descriptive metaphor "ecology" in recognition of the massively transformed 21st century information environment and the resulting disruption to established scholarly networks. Use of the term "ecology" affirms that knowledge infrastructures are diversely populated social constructs —modes of handling and regarding knowledge objects which are in themselves products of human endeavor. Situating research data in their own ecology recognizes that these knowledge objects "have no value or meaning in isolation; they exist within a (system) of people, practices, technologies, institutions, material objects, and relationships."[16]

Important distinctions between the ecologies of data and literature noted by Borgman include (among others):

- Differing standards and practices around peer review that validate the accuracy and replicability for each genre of research output;
- Differing modes of documenting provenance, because published papers self-contain information about authorship, affiliation, methodologies, etc., while data lose their meaning when extracted from the context in which they were produced; and
- Legal regimes and attribution practices that govern research papers, as original expressions fixed in a tangible medium, but which may not apply to dynamic compilations of facts that lack sufficient originality to warrant copyright protection.

Given that research data inhabit their own ecology separate and apart from the mature (yet continually evolving) system surrounding research publications, it is not always clear what the librarian's roles and responsibilities are with these new components of the scholarly record. This was the question driving us to launch the *Journal of Librarianship and Scholarly Communication*'s (*JLSC*) special issue on "Data Sharing, Data Publication and Data Citation." We invited librarians engaged in research data support, and specifically data sharing, to describe their vision and experience in order to paint a picture of this new ecosystem. And contribute they did. The fifteen papers published in this issue cover many

[15] Borgman, C. (2015). *Big Data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

[16] Ibid., page 4

of the components of research data knowledge infrastructure defined by Edwards et al. and Borgman.[17,18] Perhaps more remarkably, all of the research papers submitted to this issue are accompanied by the underlying data either as supplements to the article or as cited datasets published elsewhere. In this way, the 33 contributors to this special issue not only provide critical analysis of an evolving ecosystem; many also model exemplary data sharing attitudes and behaviors that are essential to its survival.

Defining the Contours of "Research Data"

Ecosystems occupy a given terrain, defined in place and time by geographic borders, elevation contours, physical features, and the living organisms that populate them. The research data ecology also occupies a landscape defined, at its core, by a negotiated and evolving understanding among the inhabitants of that landscape of what research data means within the context of the scholarly record. As Borgman notes in *Big Data, Little Data, No Data*,

> data rarely are things at all. They are not natural objects with an essence of their own. Rather, data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. Those representations vary by scholar, circumstance, and over time. Across the sciences, social sciences, and the humanities, scholars create, use, analyze, and interpret data, often without agreeing on what those data are. Conceptualizing something as data is itself a scholarly act.[19]

Contributors to this issue have conceptualized research data in myriad ways, reflecting an incisive understanding of the disciplinary practices and needs of varied research domains. Hélène Prost, Cécile Malleret, and Joachim Schöpfel recognize the potential of supplementary materials associated with humanities and social sciences dissertations as "windows for the scientist" inviting further integration, extension, enrichment, aggregation, and updating to advance knowledge creation. In their view, archival text samples may represent data reflecting the common characteristics of a historical group; a curated database of Egyptian steles could be useful for data mining or meta-analysis. The key to unlocking these potential data assets for future reuse and application is handling them as more than incidental supplements tightly coupled to the research thesis.[20]

---

[17] Edwards et al. (2013), see Footnote 13

[18] Borgman (2015), see Footnote 15

[19] Ibid., page xviii

[20] Hélène Prost, Cécile Malleret, and Joachim Schöpfel. *Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities*.

Andrew Gordon, David Millman, Lisa Steiger, Karen Adolph, and Rick Gilmore curated videos on child development and learning as datasets for behavioral scientists, noting that the videotaped session represents "a basic unit of analysis" for observational studies.[21] Stacey Knight Davis, Todd Bruns, and Gordon Tucker recognized the value of herbarium sheets as fugitive data of great value for biodiversity informatics research. Their carefully planned data rescue project advances knowledge discovery about changes to the natural environment when species are disappearing at an accelerating rate.[22] Human population survey data attracted the attention of Astrid Recker, Stefan Müller, Jessica Trixa and Natascha Schumann, impelling them to enhance the value of archived survey data through georeferencing.[23] In these diverse ways, *JLSC* contributors demonstrate that librarians engaged in data sharing initiatives hold a broad and boldly pragmatic view of what should at this moment be considered data and are particularly sensitive to the authentic differences in data genres across different disciplines.

Recognizing Community Composition and Community Interactions

Beyond the physical (abiotic) components of a terrain, ecosystems are also defined by the organisms populating the landscape, their interactions, and the services they offer to each other and to the larger system. The ecology of research data represented in this *JLSC* special issue comprises well-delineated communities engaged in complex interactions, including the process of forging new relationships and interactions across the boundaries of those communities. The living, breathing scholars populating this landscape provide and receive a diverse set of services as libraries and librarians seek to develop richer understandings of the scholars they aim to support, and in doing so are revamping the fundamental skill sets and roles required to thrive in this new world. Hilary Davis and William Cross discuss their collaborative Data Management Plan review service as an opportunity to "share best practices, learn from each other, and form the broad network of research support on campus."[24] They identify librarian subspecialists with the expertise to meet the needs of researchers managing data and they enumerate a set of core competencies necessary to audit a Data Management Plan. Andrew Johnson and Megan Bresnahan determined that a full-

[21] Andrew Gordon, David Millman, Lisa Steiger, Karen Adolph, and Rick Gilmore. R*esearcher-Library Collaborations: Data Repositories as a Service for Researchers*.

[22] Stacey Knight-Davis, Todd Bruns, and Gordon C. Tucker. *Big Things Have Small Beginnings: Curating a Large Natural History Collection—Processes and Lessons Learned.*

[23] Astrid Recker, Stefan Müller, Jessica Trixa, and Natascha Schumann. *Paving the Way for Data-centric, Open Science: An Example from the Social Sciences*

[24] Hilary M. Davis and William M. Cross. *Using a Data Management Plan Review Service as Training Ground for Librarians.*

day training program on research data management alleviated subject librarians' anxiety and lack of confidence in providing research data support—an area of job performance they recognize as increasingly important to their success. The ability to practice and improve research data support skills in a safe setting before attempting to apply them in the wild, with real-life researchers, measurably improved the confidence of these librarians.[25]

These practice articles indicate that a specialized subset of information professionals (instruction experts, data specialists, and subject liaisons) form alliances to improve the capacity for librarians as a whole to succeed in the research data ecology. Even more powerful is when this mix includes librarians who also occupy the niche of researcher, bringing their first-hand understanding of the investigative process into services that support scholars. They cooperate with each other, share resources, and learn each other's habits and practices. They possess survival skills, if you will: a combination of core competencies, confidence, and mutual benevolence toward each other and to the other species in the community.

Librarians also survive in the research data ecology by working cooperatively and serving the needs of neighboring inhabitants: information technologists, faculty, students, researchers, administrators, and others. Their efforts to reach across the boundaries defining professions and roles produce mutually beneficial outcomes throughout the research data lifecycle. Amanda Whitmire developed and delivered a graduate level, credit-bearing Research Data Management course for graduate students to improve the data literacy core competencies of these new researchers. She notes that "providing instruction in data information literacy is one of the primary areas of engagement" for libraries. As with librarians, graduate students expressed a preference for applying concepts to real-world cases to gain confidence in increasingly important research skills they are anxious about mastering.[26] Jennifer Doty, Melanie Kowalskie, Bethany Nash, and Simon O'Riordan launched a data deposit service for thesis authors, leveraging tools from the Dataverse Network developed in a neighboring university community. Two keys to this successful alliance were: navigating new legal terrain to facilitate resource sharing across institutions and attracting participation of early career researchers by casting the service as critical professional development in the responsible conduct of research.[27] Gordon et al. discuss the inner workings of a library-hosted observational video repository program to support the research of a specific scholarly community. They note that

[25] Andrew M. Johnson and Megan M. Bresnahan. *DataDay!: Designing and Assessing A Research Data Workshop for Subject Librarians*.

[26] Amanda L. Whitmire. *Implementing a Graduate-Level Research Data Management Course: Approach, Outcomes, and Lessons Learned.*

[27] Jennifer Doty, Melanie Kowalski, Bethany Nash, and Simon O'Riordan. *Making Student Research Data Discoverable: A Pilot Program Using Dataverse.*

researcher-librarian interactions must take place throughout the research process to ensure well-curated (and ultimately useful) behavioral data.[28] Kerstin Helbig, Brigitte Hausstein, and Ralf Toepfer describe how a DOI allocation service helps data producers, providers, and users register, discover, and credit research data in a reliable and normative fashion. Their service requires that libraries, in their role as data providers, demonstrate critical capacity to use the DOI service—authority to register the research data and long-term commitment to maintain the dataset in a useable and accessible form. This library responsibility assures that the registered data is of likely interest to other researchers and users and thus has citation potential. Helbig et al. caution libraries to seriously consider the challenges of dataset registration (e.g., versioning, granularity, assurance of metadata quality, handling dynamic datasets) before taking on this essential service role in the research data ecology.[29]

Contributors to this *JLSC* special issue demonstrate how librarians share their particular values, principles, and skills to help the research data ecology flourish and grow. Librarians develop and apply resources, techniques, technologies, and standards that serve the system in a productive way. These characteristics may lead some to conclude that librarians serve as *keystone species* in the research data ecology:

> A keystone species is a species that exerts an impact on its community that is both strong and disproportionate to its abundance. The keystone analogy refers to the architectural element at the apex of an arch that locks the other pieces into position, and is used colloquially to refer to the supporting element of a larger structure.[30]

Librarians have long held a keystone position to ensure the stability of the scholarly record as traditionally understood. Now, as the very terrain of the scholarly communication landscape transforms around us, the authors contributing to this special issue demonstrate how librarians are successfully adapting to and thriving in the emerging ecology of research data and are poised to play a similarly stabilizing role in this new environment.

Understanding the Flow of Data Resources

In ecological communities, the flow of energy through the food web is an essential dynamic undergirding the success (or endangerment) of a community and the relative health of

[28] Gordon et al., see Footnote 21.

[29] Kerstin Helbig, Brigitte Hausstein, and Ralf Toepfer. *Supporting Data Citation: Experiences and Best Practices of a DOI Allocation Agency for Social Sciences.*

[30] Environmental Information Coalition (EIC) & National Council for Science and the Environment (NCSE). (2015). Keystone species. *The Encyclopedia of the Earth*. Retrieved from http://www.eoearth.org/view/article/51cbee487896bb431f696afe/

its constituent populations. Consumers and producers, predators and prey, decomposers, parasites, and pathogens: these are the roles and relationships that characterize the niches that each species occupies in the network. In the ecology of research data, data providers, aggregators, and users interact with research funders, institutional administrators, information professionals, publishers, and other scholars to ensure that the outputs of the research process are retained, managed, shared, and reused with the greatest efficacy. The complex mix of behaviors, attitudes, values, and relationships among these "species" govern the ease or friction with which data assets flow. Expectations and mandates to curate and share data; willingness or reluctance to do so; imposition of legal and ethical restrictions curtailing sharing; and normative, benevolent practices that widen the flow of assets govern the energy flow of data resources through the research data ecology.

Contributors to this issue make it clear that the attitudes, behaviors, and practices around data sharing across the academy writ large are in an early stage of evolution. Authors Carolyn Bishoff and Lisa Johnston observe, in their analysis of Data Management Plans at one top-tier research university, that the primary means of data sharing occurs through "the traditional journal publication…or more accurately, via components of a journal publication (tables, graphs, images, etc.)." The second most common form of sharing detected in their analysis is "providing data on request."[31] The findings of Cunera Buys and Pamela Shaw from their survey of a different university with very high research activity also reflect the predominance of residual behaviors and practices inherited from the "traditional" scholarly publishing ecology. Their respondents store data using older technologies such as computer and external hard drives, flash drives, departmental servers, or internal capacity within their research instrumentation. A small minority of respondents share data in external repositories, with the majority indicating reluctance to share data publicly before formal publication of research results.[32] Contributor Philip Herold's investigation of data sharing practices of ecologists and biologists at a research university found that the vast majority of publicly shared data was made available as supplemental files via journal websites. A smaller percentage of study participants placed data in disciplinary repositories, suggesting that community norms and availability of repository technologies in the biological sciences may be more mature than those in some other related fields.[33]

[31] Carolyn Bishoff and Lisa Johnston. *Approaches to Data Sharing: An Analysis of NSF Data Management Plans from a Large Research University.*

[32] Cunera M. Buys and Pamela L. Shaw. *Data Management Practices Across an Institution: Survey and Report.*

[33] Philip Herold. *Data Sharing Among Ecology, Evolution, and Natural Resources Scientists: An Analysis of Selected Publications*.

The data sharing guidelines and requirements of universities with high research activity were the focus for Kristin Briney, Abigail Goben, and Lisa Zilinksi's study of institutional data policies. The authors report that less than half of the universities studied have a data policy and that the existence of a standalone data policy correlates with higher research expenditures and higher Carnegie classification. Perhaps not surprisingly, the existence of institutional data policies also correlates with the presence of a dedicated data librarian on staff. The inference from this study may be that data librarians are an "indicator species" in the research data ecology—a marker of sorts of the most serious library engagement in an institution's research enterprise.[34]

Finally, the matter of which species in the research data ecology are sustained by the most common sharing practices is carefully considered in the commentary by Wendy Walker and Teressa Keenan. Their call to make research data not only available but "truly accessible" is an important reminder that some of the practices well-honed in the traditional scholarly publishing system are worth emulating in the research data ecology. These contributors' vision of data sharing to enrich the future work of all prospective researchers and users conveys the value of "Universal Participation: everyone must be able to use, reuse and redistribute."[35] Their suggested practices help fulfill the objective of truly accessible data: providing raw data that can be interpreted using assistive technology, associating rich metadata to datasets to serve as rich alternative text, and avoiding proprietary data rendering technologies that can be barriers to access and reuse.[36] Their suggestion that librarians acquire and then offer expertise and assistance in the area of data accessibility offers yet another vision for how our profession's genera can nourish and serve the research data ecosystem.

Conclusion

Contributors to this *JLSC* special issue enthusiastically argue for the pragmatic need to hold a somewhat elastic view of data, especially in this transitional period and especially when working with legacy content. While some may caution that too capacious a definition of data risks losing sight of distinctive characteristics about this type of scholarly object, in these relatively early days we are more inclined towards Michele Hayslett's cautions—which she articulates in a critique of the very call for papers for this special issue—against scoping

[34] Kristin Briney, Abigail Goben, and Lisa Zilinksi. *Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies*

[35] Open Knowledge Foundation. (2012). *What is open data? – Open data handbook*. As cited in Walker & Keenan from http://opendatahandbook.org/en/what-is-open-data/#what-is-open

[36] Wendy Walker and Teressa Keenan. *Going Beyond Availability: Truly Accessible Research Data.*

concepts too narrowly or making presumptions without knowing.[37] To invoke Borgman once again, in this vast, "elephantine" research data ecosystem, the perspectives and experiences of any single professional are necessarily limited. The contributors to this *JLSC* special issue teach us that it is through broad-minded seeing, careful analysis, authentic engagement, much-needed listening, and thoughtful discussion that librarians can fulfill their potential to serve as a keystone species in the research data ecology, playing a unique and crucial role in the way it functions.

[37] Michelle Hayslett. *Data World Does Not Lack Standards.*