Journal of eScience Librarianship
putting the pieces together: theory and practice

Exploring Workforce Factors in the Data Fields

Heather Soyka, Kent State University, Portage County, OH, USA, hsoyka@kent.edu iD
Angela Murillo, Indiana University Indianapolis, Indianapolis, IN, USA, apmurill@iu.edu iD

Abstract

As the data fields continue to grow and evolve, it is critical to examine factors that impact the workforce. This study explores various workforce development factors through a survey that asked data workers to describe their time spent on data-related activities, training needed for data-related activities, facilitators and barriers to workforce entry, facilitators and barriers to workforce retention, and the impact of diversity, equity, and inclusion efforts. Drawing together these varied factors allows for examination and triangulation of overlapping factors that shape and contribute to workplace experiences for data professionals.

Introduction

Professionals in research data management, data curation, data preservation, data archives, or related fields such as data science, data workflows, and scientific data management, are responsible for the full lifecycle of research data. Across a wide range of responsibilities, tasks, processes, and data types, they perform crucial roles and interventions by ensuring that data is responsibly managed and stewarded, starting from initial research proposal and collection, and supporting appropriate documentation and confidentiality, to providing long-term preservation and access. Making sure that data are findable, accessible, interoperable, and reusable according to the FAIR Principles (Wilkinson et al. 2016; GoFAIR, n.d.) is central to professional data work. Professional data managers are essential to maximizing data impact and reuse, meeting institutional and funder requirements for data sharing, promoting sharing and reproducibility, and ensuring long-term preservation and access to data.

This study explores workforce development factors in research data management, data curation, and related data fields, including facilitators of and barriers to entry, retention, diversity, inclusion, and equity efforts. As the data fields continue to grow and evolve, it is critical to examine challenges and opportunities that impact the data workforce. This research surveyed data workers to gain an understanding of historical and current facilitators and barriers of entry, as well as retention, diversity, inclusion, and equity efforts.

A 27-question survey was distributed to a variety of listservs and through professional social media to recruit professionals in research data management, data curation, data preservation, data archives, or related fields such as data science, data workflows, and scientific data management. The survey questions were designed to address the following research questions:

- RQ1: Which data-related activities does the data workforce spend the majority of their time doing, and in which do they need more training and professional development?

- RQ2: What are the facilitators and barriers to entry in the data workforce?

- RQ3: What are the facilitators and barriers to retention in the data workforce?

- RQ4: What diversity, equity, and inclusion (DEI) initiatives are impacting the data workforce?

The remainder of this paper provides a review of related literature, survey development and distribution, findings, discussion, and conclusion. The data supporting these findings are openly available through the following Figshare repository: https://doi.org/10.6084/m9.figshare.28012625.

Related Literature

Research data management, data curation, and related data fields have become increasingly critical as researchers navigate changes in the techniques, technologies, policies, and practices of research and scientific data management. The importance of well-curated data for science and research cannot be overstated, as

they play a vital role within scientific practice across analytic methods and techniques (Gold 2010) and enable interdisciplinary research (Witt 2008).

Over the past decades, Library and Information Science (LIS) educators have provided various training opportunities to educate future data management and data curation professionals. Within iSchool and ALA-accredited library science programs, data curation programs, such as master's-level specializations and certificate programs, have been available at the University of Illinois at Urbana-Champaign, University of Arizona at Tucson, Syracuse University, University of Michigan at Ann Arbor, University of Tennessee at Knoxville, and the University of North Carolina at Chapel Hill, amongst others, for several decades (Botticelli et al. 2011; Committee on Future Career Opportunities and Educational Requirements for Digital Curation 2015). Along with formal data curation programs within LIS, there have been programs developed in related fields such as chemistry, museum studies, and nursing (Reisner et al. 2014; Tibbo and Duff 2008; Sylvia and Terhaar 2014), where students in specific domains learn about data curation specific to their field. Furthermore, there have been examinations of how the skill sets found within data science programs have some overlap with those of data curation (Virkus and Garoufallou 2019, 2020).

Assessing the data management landscape to mark maturity, growth, and professional needs is important and perennial (Cox et al. 2017; Latham 2017; Bresnahan and Johnson, 2013). This has often occurred in LIS programs along with scoping the changing practices in scholarly communication and liaison librarianship (Corrall 2012; Jagazewski and Williams 2013; Mattern et al. 2015; Cox et al. 2017; Nel 2020). Along with formal education, there have been professional development opportunities for professionals both within LIS and outside of LIS to gain data curation skills. For example, the US-based RDMLA (Research Data Management Librarian Academy, n.d.), and the UK Jisc-funded projects RDMRose (University of Sheffield), DataPool (University of Southampton), and MANTRA (University of Edinburgh) have provided workshops, training, resources, and open, self-paced online professional development curricula on research data management. DataONE (DataONE, n.d.) provides on-demand webinars, training workshops, and open educational resources regarding scientific data management. Lastly, several organizations offer professional services, case studies, and guidance related to research data management freely online, such as the Digital Curation Centre (DCC 2023) and the Data Curation Network (DCN, n.d.).

Defining the types of data curation activities has been important for understanding the skills necessary for building this workforce. Kim et al. (2011) found six data curation activities, including collecting primary data, collecting secondary data, storing data, managing data, analyzing data, and presenting data. Palmer et al. (2014) identified three types of data curation activities: technical duties (i.e., preserving and analyzing data), service duties (i.e., providing training, assisting with data management plans), and administrative/managerial duties (i.e., overseeing collections, developing policies).

Finally, some researchers have examined the demographics and education of the existing data curation workforce. Practitioners who entered the rapidly growing research data management and data curation spaces in the 2010s or earlier did not have the same access to formal educational opportunities tailored to those areas that more recent graduates have encountered. As these areas continue to quickly mature and evolve, opportunities to learn, shape, and participate have been unevenly available. Thompson et al. (2013) surveyed recent graduates from the data curation specialization at the University of Illinois at Urbana-Champaign to assess the value of the program, career placement, and continuing education needs. This study indicated that while graduates often worked in academic libraries, many data curation skills were also transferable to the corporate and non-profit sectors. Additionally, participants indicated a need for continual education, particularly to keep up with changes in data formats, standards, and best practices (Thompson et al. 2013). Bishop et al. (2021) surveyed data managers in the earth science field and conducted follow-up interviews to learn about job tasks and responsibilities related to communication, collaboration, outreach, research, data, and the technologies used for data work.

While each of these studies explores various aspects of the data workforce, none of them have explored the workforce's current research data activities, needs for professional development and training, entry and retention factors, along with the impacts of DEI efforts.

Research Methods

As described in the introduction, this study examines various aspects of the data workforce, specifically to gain an understanding of data-related activities and training needs, facilitators and barriers of entry, as well as retention, diversity, inclusion, and equity efforts.

The survey questions were designed to address the following research questions:

- RQ1: Which data-related activities does the data workforce spend the majority of their time doing, and in which do they need more training and professional development?

- RQ2: What are the facilitators and barriers to entry in the data workforce?

- RQ3: What are the facilitators and barriers to retention in the data workforce?

- RQ4: What diversity, equity, and inclusion (DEI) initiatives are impacting the data workforce?

The online survey instrument was developed during the Spring and Summer of 2023. The researchers piloted the survey with several colleagues to ensure ease of navigation and to gather feedback regarding the survey question design. The researchers updated the preliminary survey to create the final 27-question survey instrument based on the test participants and feedback from the survey design expert. See Appendix A for survey questions.

Both researchers followed the IRB requirements of their respective institutions. As this is a multi-institution collaboration, IRB approval was obtained in accordance with each institution's specific requirements. This study was reviewed and approved for human subject research through the Indiana University Institutional Review Board, IRB#11321, and the Kent State University approved IRB #21-221.

The survey asks participants about their role at their organization, their educational background, their time spent in certain data-related activities, their training needs for certain data-related activities, facilitators and barriers to entry in their current position, facilitators and barriers to retention in their current position, organizational participation in DEI initiatives, how DEI initiatives influence their workplace, and several demographic questions. The survey included a mixture of multiple-choice, Likert-scale, and open-ended questions to gain multiple levels of nuance in the responses. As with all survey studies, a general limitation is that it is difficult to capture nuance in the responses. To help alleviate this limitation, open-ended questions were included at the end of each survey section.

The survey was administered anonymously online using Qualtrics, with the goal of confidentiality. Although reidentification is always possible, the authors have consciously constructed the survey and dataset to allow for anonymity. The survey was distributed to a variety of listservs and through professional social media to recruit professionals in research data management, data curation, data preservation, data archives, or related fields such as data science, data workflows, and scientific data management. These recruitment channels included JESSE, CODATA, Research Data Access and Preservation Association, DataCure, DataONE, Earth Science Information Partners, Society of American Archivists (SAA), SAA Science Technology and Healthcare Group, Research Data Management Librarian Academy (RDMLA), Center for Scientific Collaboration and Community Engagement (CSCCE), and Research Data Alliance metadata group listservs and Slack channels. Additionally, the recruitment information was shared on LinkedIn. The survey was open for responses from September 28 to November 1, 2023.

The data were cleaned and analyzed using basic descriptive statistics through a combination of pivot tables and standard functions available in Excel, Google Sheets, and SPSS. Open-ended questions were analyzed using open coding to capture the major themes.

Findings

Of the 73 responses that were received, 40 participants completed the entire survey, including the demographic questions. These 40 complete responses were analyzed.

Participant Demographics

Forty participants completed the demographic portion of the survey. Participants indicated their gender, years in the profession, and the highest level of education (See Tables 1 - 3). As shown, a little more than half of the participants were female (52.5%), while twenty-seven percent were male, and several participants

identified as non-binary. Participants represented a range of years in the profession, with most having 10 to 14 years (30%), 0 to 4 years (20%), and 20+ years (20%). Finally, the majority of our participants held a Master's degree (47.5%) or a Ph.D. or higher (35%).

Table 1: Participants' Gender (n = 40)

Gender	Count (%)
Female	21 (*52.5*)
Male	11 (*27.5*)
Non-binary/Third gender	3 (*7.5*)
Other	1 (*2.5*)
Prefer Not to Answer	4 (*10.0*)

Table 2: Participants' Years in Profession (n = 40)

Years in Profession	Count (%)
0 to 4 years	8 (*20.0*)
5 to 9 years	4 (*10.0*)
10 to 14 years	12 (*30.0*)
15 to 19 years	3 (*7.5*)
20 + years	8 (*20.0*)
Prefer Not to Answer	5 (*12.5*)

Table 3: Participants' Highest Level of Education (n = 40)

Highest Level of Education	Count (%)
Bachelor's Degree	2 (*5.0*)
Master's Degree	19 (*47.5*)
Ph.D. or Higher	14 (*35.0*)
Other	1 (*2.5*)
Prefer Not to Answer	4 (*10.5*)

Next, participants also indicated their job location and their race or ethnicity (See Tables 4 and 5). The majority of respondents identified themselves as white (67.5%); 11% reported having two or more races or ethnicities; 6% were Asian; 3% were Latino or Hispanic; 3% were Black, African, or African American; and 3% did not specify their race or ethnicity. Regarding job locations, 81% of the participants were in the

United States (U.S.) (28% in the Northeast, 22% in the South, 17% in the West, and 14% in the Midwest). For the remaining respondents, 8% reported that they were in Australia or the Pacific Islands, 6% in Europe, 3% in Asia, and 3% in Africa.

Table 4: Job Location of Participants (n = 40)

Job Location (n = 489)	Count (%)
U.S. Northeast	10 (*25.0*)
U.S. Midwest	5 (*12.5*)
U.S. South	8 (*20.0*)
U.S. West	6 (*15.0*)
Canada	0 (*0.0*)
Mexico, Central and South America, and the Caribbean	0 (*0.0*)
Europe	2 (*5.0*)
Asia	1 (*2.5*)
Africa	1 (*2.5*)
Australia and the Pacific Islands	3 (*7.5*)
Prefer Not to Answer	4 (*10.0*)

Table 5: Race/Ethnicity of Participants (n = 40)

Race/Ethnicity	Count (%)
American Indian or Alaska Native	0 (*0.0*)
Asian	2 (*5.0*)
Black, African, or African American	1 (*2.5*)
Native Hawaiian or Other Pacific Islander	0 (*0.0*)
White	27 (*67.5*)
Latino or Hispanic	1 (*2.5*)
Arab/Middle Eastern	0 (*0.0*)
Other	1 (*2.5*)
Prefer Not to Answer	4 (*10.0*)

Lastly, participants were asked about their employment status; 86% were employed as full-time employees, 6% were part-time employees, 3% were grant-based contractors, 3% were independent contractors, and 3% were students.

Participant Roles and Educational Background

The survey participants reported their primary roles and secondary roles within their organizations. Primary roles included Librarian (47.5%), Data Manager (12.5%), Faculty/Educator (10.0%), Data Curator (7.5%), Archivist (5.0%), Research Scientists (5.0%), Data Scientist (2.5%), and Other (10%), which included an IT professional, Data Architect, and a Director/Manager. Additionally, participants also provided their secondary roles, which included Data Curator (17.5%), Not Applicable (15.0%), Research Scientist (15.0%), Data Manager (12.5%), Faculty/Educator (10.0%), Librarian (4%), Data Scientists (7.5%), Archivist (2.5%), and Other (10.0%), which included Data Steward, Researcher, and Metadata Manager.

Participants' educational backgrounds included Domain Science (35.0%), Library Science (27.5%), Information Science (20.0%), Archival Science (7.5%), Data Science (2.5%), and Other (7.5%), which included Computer Science, English, and Social Science.

Participant Data-Related Activities and Training Needs

Participants were asked about their time spent and training needs for the following data-related activities. These definitions were modified from the DataONE data lifecycle (DataONE, n.d.) and the UCLA Data Science Center Data Literacy Course Competencies (Peterson and Ali 2021).

- Acquire/Collect: data are captured through fieldwork, sensors, or instruments.

- Assure: data quality is assured through checks and inspections.

- Describe/Organize: data are annotated and described using the appropriate metadata standards, vocabularies, ontologies, or other standards.

- Preserve: data are submitted to an appropriate long-term archive (e.g., data center).

- Find/Discover: potentially useful data are located and obtained for reuse, along with the relevant information about the data (e.g., metadata, code).

- Share/Publish: data are disseminated through portals or repositories, including necessary items for reuse (e.g., accompanying metadata, code, workflows, persistent identifiers).

- Integrate: data from disparate sources are combined to form one homogeneous set of data that can be readily analyzed.

- Analyze: data are analyzed, modeled, or visualized.

Participants were asked to report the amount of time they spent conducting data-related activities and their training needs (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal). As shown in Figure 1, participants spend the least time acquiring/collecting, integrating, and assuring data, and most of their time sharing or publishing data and describing/organizing data.
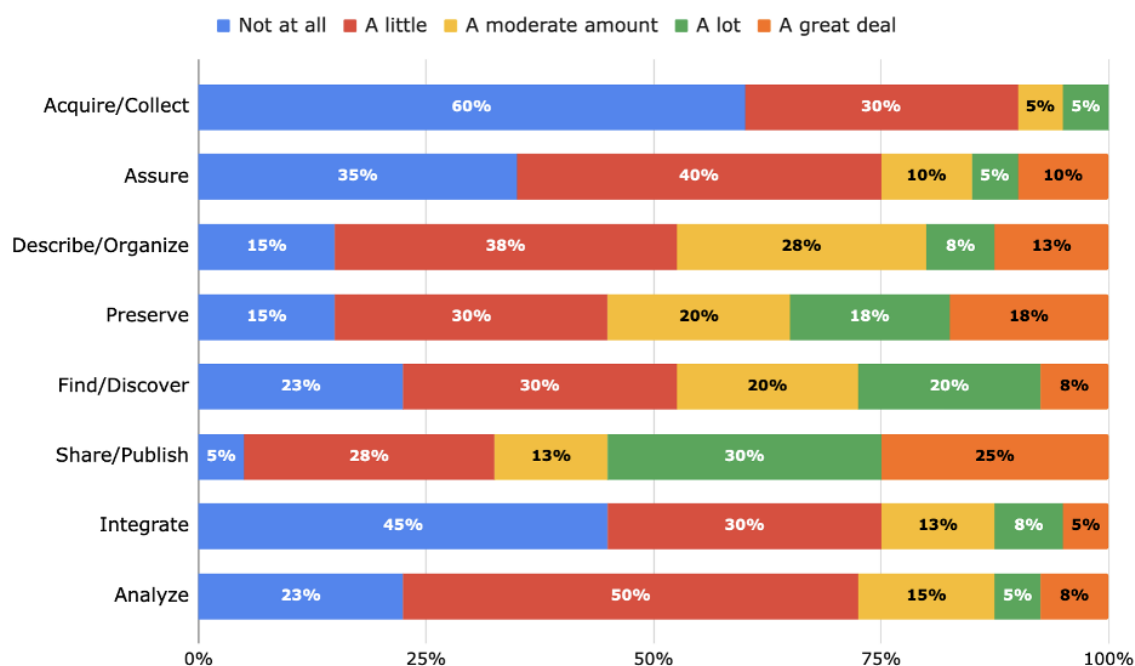
Legend: ■ Not at all ■ A little ■ A moderate amount ■ A lot ■ A great deal

Figure 1 data:

Activity	Not at all	A little	A moderate amount	A lot	A great deal
Acquire/Collect	60%	30%	5%		5%
Assure	35%	40%	10%	5%	10%
Describe/Organize	15%	38%	28%	8%	13%
Preserve	15%	30%	20%	18%	18%
Find/Discover	23%	30%	20%	20%	8%
Share/Publish	5%	28%	13%	30%	25%
Integrate	45%	30%	13%	8%	5%
Analyze	23%	50%	15%	5%	8%

Figure 1: Time spent on data-related activities (n=40)

Legend: ■ Not at all ■ A little ■ A moderate amount ■ A lot ■ A great deal

Figure 2 data:

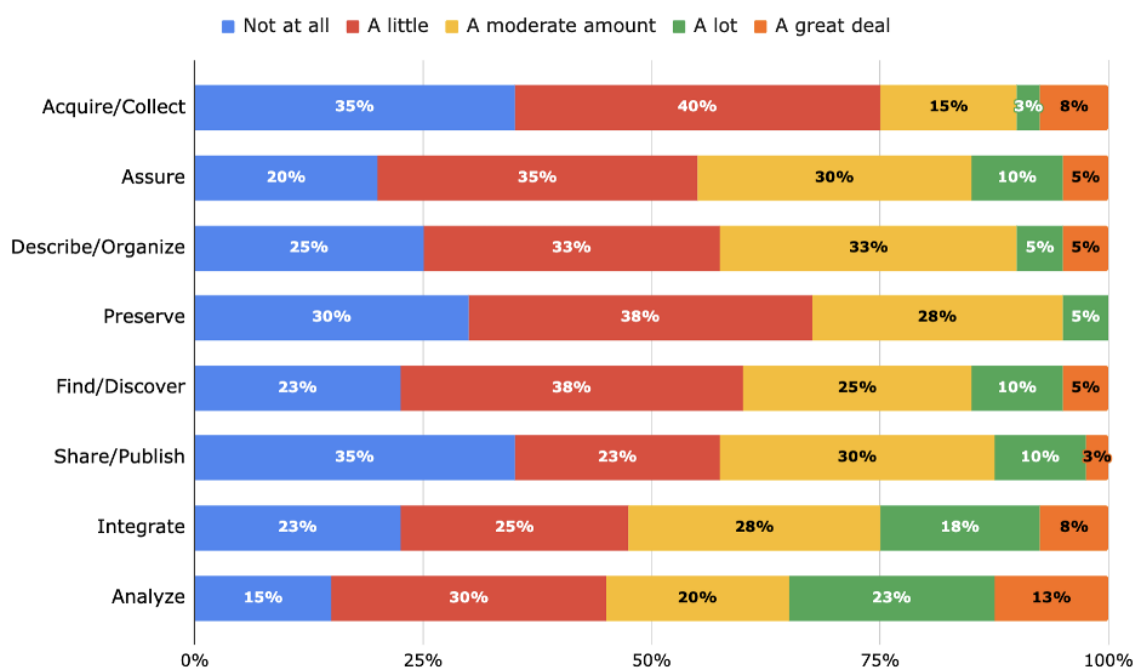| Activity	Not at all	A little	A moderate amount	A lot	A great deal
Acquire/Collect	35%	40%	15%	3%	8%
Assure	20%	35%	30%	10%	5%
Describe/Organize	25%	33%	33%	5%	5%
Preserve	30%	38%	28%	5%	
Find/Discover	23%	38%	25%	10%	5%
Share/Publish	35%	23%	30%	10%	3%
Integrate	23%	25%	28%	18%	8%
Analyze	15%	30%	20%	23%	13%

Figure 2: Training needs for data-related activities (n=40)

Participants were also asked how much additional training or professional development they needed for each data activity (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal). As shown in Figure 2, analyze, integrate, and assure are the data activities for which participants need the most training, while participants need the least training in acquire, collect, and preserve activities.

Tables 6 and 7 provide a comparison of data-related daily activities in terms of time spent and training required. From these tables, we see that participants spend the majority of their time on data activities related to data sharing/publishing, preserving, and describing/organizing. Additionally, they spend significant time on activities associated with finding/discovering and analyzing data. However, participants spend less time on activities related to data integration and acquiring/collecting.

Table 6: Time spent on data-related activities (n=40)

Data-Related Activity	**Time Spent (Mean (SD))**
Share/Publish	3.26 (1.26)
Preserve	2.92 (1.35)
Describe/Organize	2.67 (1.21)
Find/Discover	2.59 (1.26)
Analyze	2.26 (1.10)
Assure	2.15 (1.25)
Integrate	2.00 (1.17)
Acquire/Collect	1.56 (0.81)

As shown in Table 7, participants indicated they have their highest training needs in activities related to data analysis, integration, and assurance. In contrast, their time spent in these activities is relatively low compared to other data-related activities (see Table 6). As shown, while participants spend a considerable amount of time preserving and sharing/publishing data, they have lower training needs for these activities. Lastly, participants indicated having the least training needs in acquiring and collecting data, and this is also the data-related activity where they spend the least time.

Participants were also asked to describe any opportunities or challenges relevant to data-related activities in their current position through an open-ended question. Participants described a general lack of training for many data-related skill sets and noted that the quickly evolving field makes it difficult to keep up with new techniques and tools. Several participants described a need for training in basic programming skills, data analytics, and data visualization. Additionally, several participants noted that keeping up with the sheer number of data types and file formats, as well as standards, can be challenging. Participants suggested that, given the interdisciplinary nature of data curation, there is a need for a greater focus on formalized

management processes. Several participants described the need for more automated QA/QC processes. Keeping up with fast-moving developments in the field while also providing high-level curation services is an ongoing challenge. Additionally, several participants discussed a general need for larger budgets and prioritization for data curation activities. One participant specifically mentioned that since there is no formal definition for "data curator" in the U.S. Government's Federal hiring system, the work of a federal data curator is often hidden or invisible.

Table 7: Training needed for data-related activities (n=40)

Data-Related Activity	Training Needed (Mean (SD))
Analyze	2.88 (1.28)
Integrate	2.63 (1.23
Assure	2.45 (1.08)
Find/Discover	2.38 (1.10)
Describe/Organize	2.33 (1.07)
Share/Publish	2.23 (1.12)
Preserve	2.08 (.89)
Acquire/Collect	2.08 (1.14)

Participant Entry in the Data Workforce (Facilitators and Barriers)

Participants were asked about facilitators for entering the data workforce. Specifically, they were asked to rate the importance of certain factors in facilitating their entry into their current position (1 = not at all important, 2 = slightly important, 3 = moderately important, 4 = very important, 5 = extremely important).

As shown in Figure 3, education, job availability, and networks/personnel connections ranked the highest, while DEI initiatives, recruitment, and internships ranked the lowest.

Next, participants were asked about barriers to entry into their workplace. Specifically, they were asked to rate the importance of certain items as barriers to entry into their current position (1 = not at all important, 2 = slightly important, 3 = moderately important, 4 = very important, 5 = extremely important). As shown in Figure 4, only job availability was considered moderately important as a barrier for participants' entry into the data workforce.

Tables 8 and 9 provide a comparison of the facilitators and barriers to entering the data workforce. As shown in Table 8, education, job availability, and the hiring process are the most impactful when entering the data workforce. Additionally, networks and personal connections, as well as post-graduate work, have some impact. Recruitment and DEI initiatives have the least impact on facilitating workforce entry.
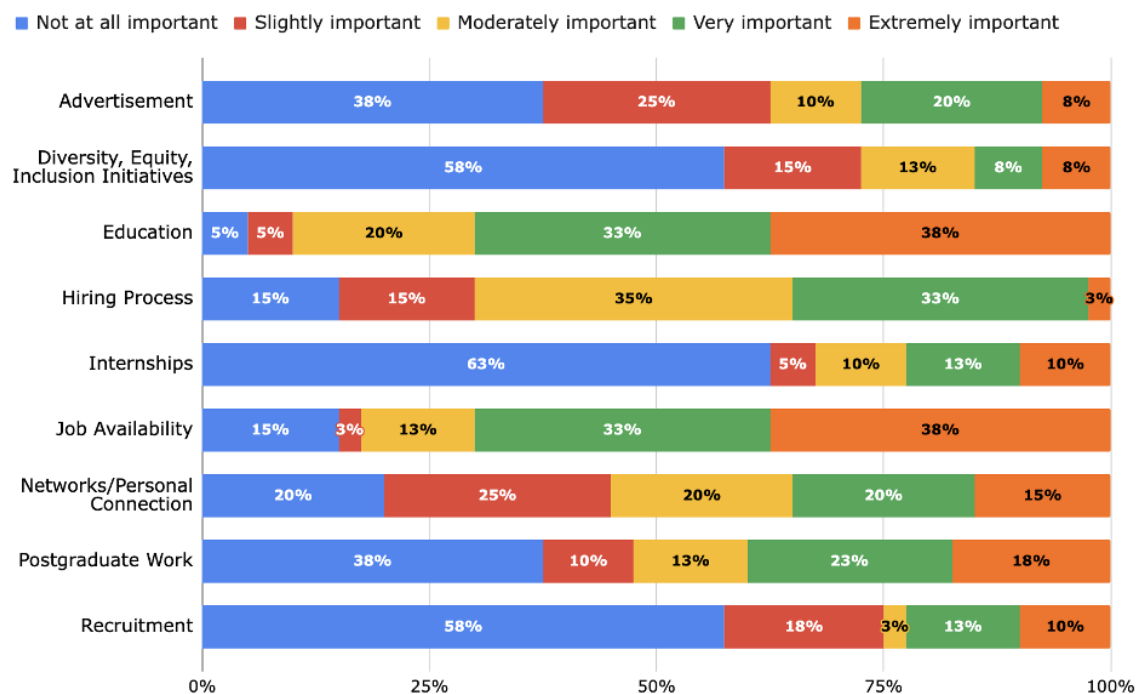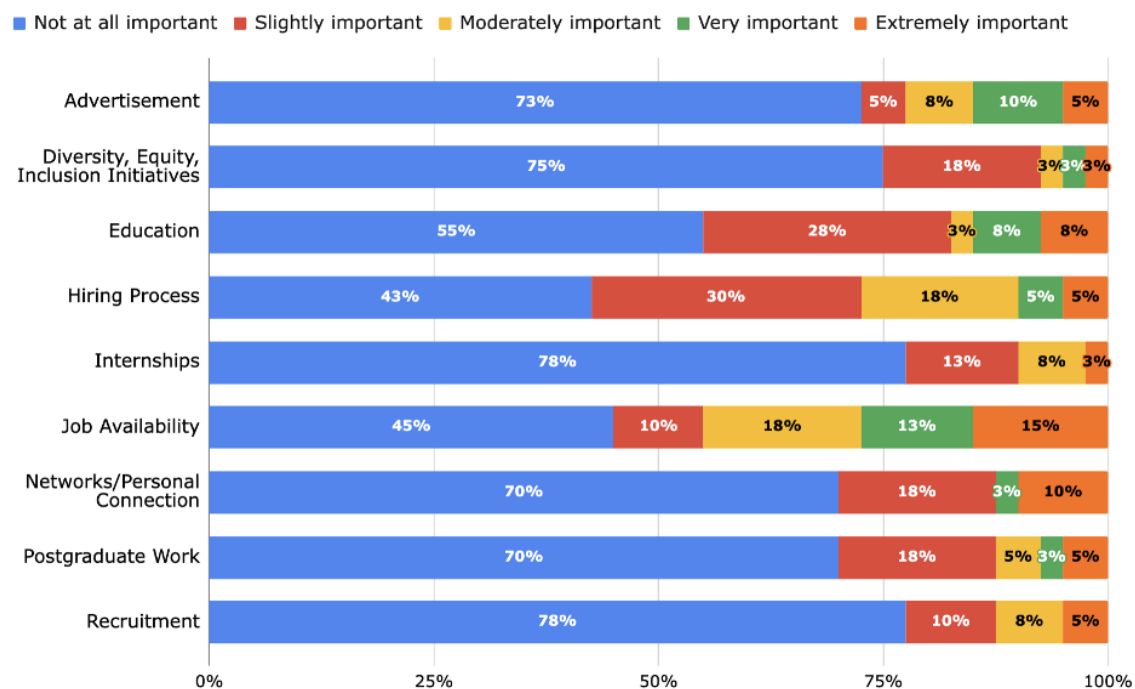
Figure 3: Facilitators of Workforce Entry (n = 40)

Legend: Not at all important | Slightly important | Moderately important | Very important | Extremely important

Category	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Advertisement	38%	25%	10%	20%	8%
Diversity, Equity, Inclusion Initiatives	58%	15%	13%	8%	8%
Education	5%	5%	20%	33%	38%
Hiring Process	15%	15%	35%	33%	3%
Internships	63%	5%	10%	13%	10%
Job Availability	15%	3%	13%	33%	38%
Networks/Personal Connection	20%	25%	20%	20%	15%
Postgraduate Work	38%	10%	13%	23%	18%
Recruitment	58%	18%	3%	13%	10%

Figure 4: Barriers to Entry Workforce Entry (n=40)

Legend: Not at all important | Slightly important | Moderately important | Very important | Extremely important

Category	Not at all important	Slightly important	Moderately important	Very important	Extremely important
Advertisement	73%	5%	8%	10%	5%
Diversity, Equity, Inclusion Initiatives	75%	18%	3%	3%	3%
Education	55%	28%	3%	8%	8%
Hiring Process	43%	30%	18%	5%	5%
Internships	78%	13%	8%		3%
Job Availability	45%	10%	18%	13%	15%
Networks/Personal Connection	70%	18%		3%	10%
Postgraduate Work	70%	18%	5%	3%	5%
Recruitment	78%	10%	8%		5%

Table 8: Workforce Entry Facilitators (n=40)

Workforce Entry Factors	Facilitators (Mean (SD)
Education	3.93 (1.12)
Job Availability	3.75 (1.39)
Hiring Process	2.93 (1.10)
Networks/Personal Connections	2.85 (1.37)
Post-Graduate Work	2.73 (1.58)
Advertisement	2.35 (1.37)
Internships	2.03 (1.48)
Recruitment	2.00 (1.43)
DEI Initiatives	1.93 (1.31)

As shown in Table 9, job availability and the hiring process are the greatest barriers to workforce entry. Many factors, including postgraduate work, recruitment, DEI initiatives, and internships, have a minimal impact as barriers to workforce entry.

Table 9: Workforce Entry Barriers (n=40)

Workforce Entry Factors	Barriers (Mean (SD))
Job Availability	2.43 (1.53)
Hiring Process	2.00 (1.13)
Education	1.85 (1.25)
Advertisement	1.70 (1.26)
Networks/Personal Connections	1.65 (1.27)
Post-Graduate Work	1.55 (1.06)
Recruitment	1.45 (1.01)
DEI Initiatives	1.40 (0.87)
Internships	1.38 (0.84)

Participants were also asked through an open-ended question to describe any opportunities or challenges that had influenced their decision to enter into their current position. Several participants described their blend of expertise in domain science, technology, LIS/archives, and data work that made them qualified for their current position. Several participants also mentioned that without the benefit of residency programs and technology internships, they would likely not be qualified for their current roles. Multiple

participants expressed that they still do not feel qualified and need additional professional development. Lastly, several participants described how their data management role was an alternative to tenure-track or research positions, but that they also found that a separation between data management and research was an unfortunate reality in practice.

Participant Retention in the Data Workforce (Facilitators and Barriers)

Participants were asked about their workplace retention. Specifically, they were asked to rate the importance of certain factors in facilitating their retention in their current position (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal).

As shown in Figure 5, job security/salary/benefits, work/life balance, and culture ranked the highest, while DEI initiatives and family-friendly initiatives ranked the lowest.

Additionally, participants were asked about barriers to retention in their workplace. Specifically, they were asked to rate the importance of certain items as barriers to retention in their current position (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal).

As shown in Figure 6, culture, training/mentoring, and job security ranked the highest, while family-friendly initiatives, geographic location, and DEI initiatives ranked the lowest.
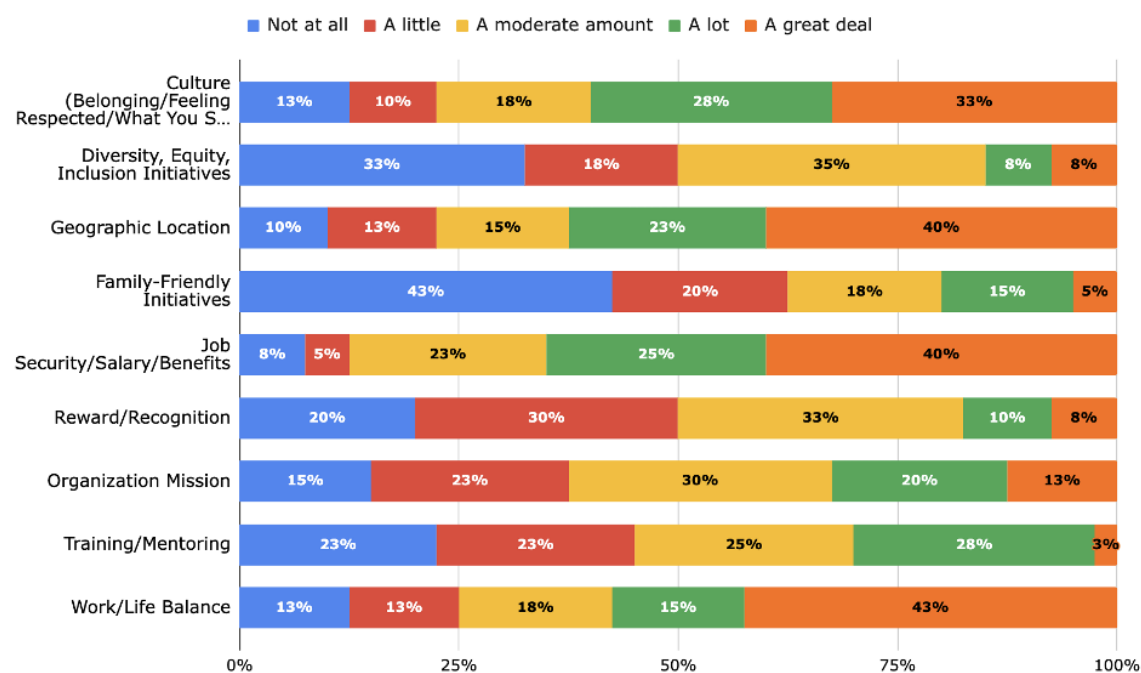

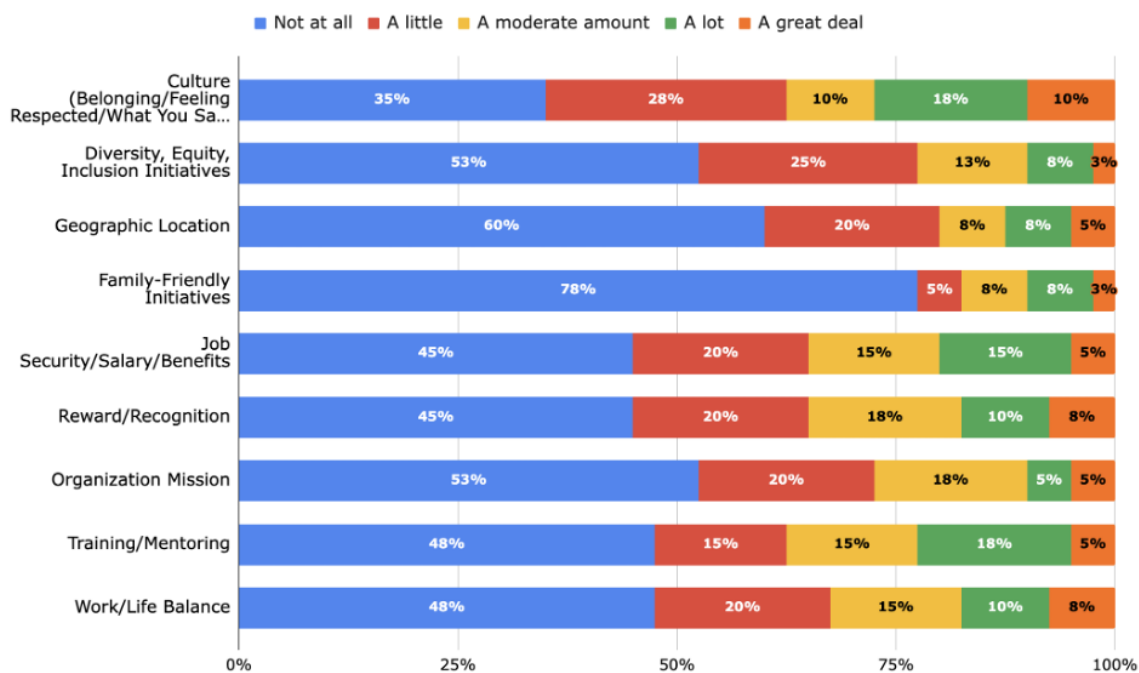**Figure 5**: Facilitators of Workplace Retention (n=40)

Figure 6: Barriers to Workplace Retention (n=40)

Tables 10 and 11 provide a comparison of the facilitators and barriers to workforce retention. As shown in Table 10, job security, geographical location, and work/life balance have the most impact on facilitating workforce retention. Additionally, culture and organizational mission also impact the retention of the data workforce. DEI and family-friendly initiatives rank the least important as facilitators for workforce retention.

Table 10: Workforce Retention Facilitators (n=40)

Workforce Retention Factors	Facilitators (Mean (SD)
Job Security/Salary/Benefits	3.85 (1.23)
Geographic Location	3.70 (1.38)
Work/Life Balance	3.63 (1.46)
Culture	3.58 (1.38)
Organizational Mission	2.93 (1.25)
Training/Mentoring	2.65 (1.19)
Reward/Recognition	2.55 (1.15)
DEI Initiatives	2.40 (1.24)
Family-Friendly Initiatives	2.20 (1.29)

As shown in Table 11, culture and training/mentoring have the greatest impact on barriers to workforce retention. Additionally, job security/salary/benefits, and reward/recognition are also impactful on workplace retention. DEI and geographic location are ranked the least important as barriers for workforce retention.

Table 11: Workforce Retention Barriers (n=40)

Workforce Retention Factors	Barriers (Mean (SD))
Culture	2.40 (1.39)
Training/Mentoring	2.18 (1.34)
Job Security/Salary/Benefits	2.15 (1.29)
Reward/Recognition	2.15 (1.31)
Work/Life Balance	2.10 (1.32)
Organizational Mission	1.90 (1.17)
DEI Initiatives	1.83 (1.08)
Geographic Location	1.78 (1.19)

Participants were also asked to describe any opportunities or challenges that could influence their decision to leave their current position through an open-ended question. Respondents further discussed how management and colleagues were influenced by both their decision to stay in their current position and their potential eagerness to leave. They also described a general lack of understanding among the higher administrative levels of their organizations regarding the importance of their work, particularly in how their work supports federal data management policies. Additionally, participants discussed that while many of their skills could be transferable to industry, it is often challenging to convey data management expertise in terms that make sense to support that move.

Participant Organizations and DEI Initiatives

When participants were asked if their organization participated in DEI initiatives, all but one participant indicated affirmatively. Additionally, participants were asked about the extent to which their organization provides various types of DEIA initiatives (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal).

As shown in Figure 7, nearly all participant organizations had issued DEI statements, some professional development and mentoring programs that targeted diverse populations, and professional development workshops to raise awareness on DEI-related issues. Lastly, most organizations have provided at least some specific policies and procedures related to DEI efforts.
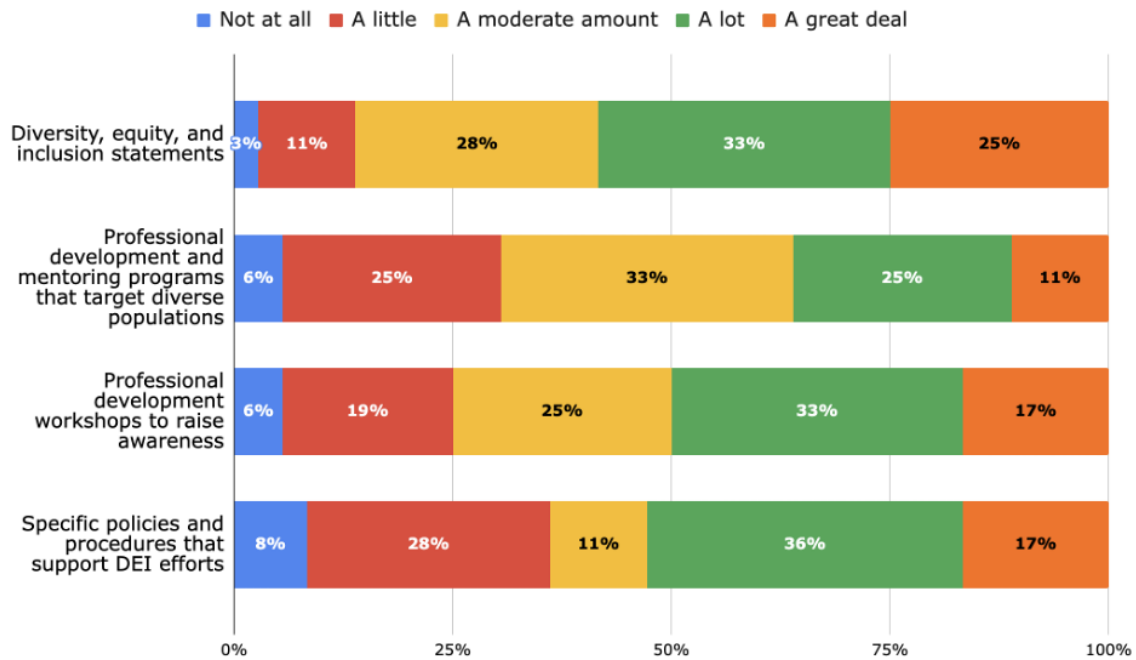
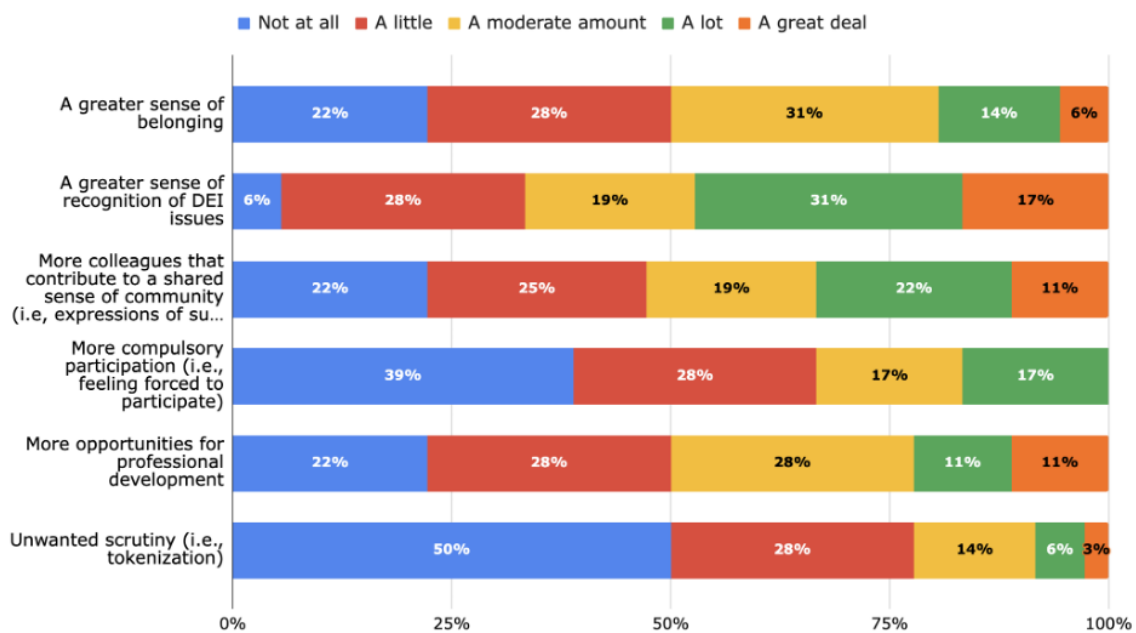**Figure 7**: DEI Initiatives in the Workplace

Figure 8: Impact of DEI Efforts in the Workplace

Participants were also asked to what extent DEI initiatives influenced their workplace (1 = not at all, 2 = a little, 3 = a moderate amount, 4 = a lot, 5 = a great deal). As shown in Figure 8, the results were mixed regarding a greater sense of belonging and the creation of a shared sense of community. Many participants stated there was a greater sense of recognition of DEI issues, as well as more opportunities for professional development. Lastly, many participants did not feel forced to participate in DEI-related efforts, nor did many participants feel unwanted scrutiny.

Participants were also asked to describe how DEI initiatives impacted entry into their workplace through an open-ended question. Many participants did not feel able to answer the question, either because of their personal experience or because they were not involved in the hiring process. However, some participants discussed the fact that there is still a significant systemic issue regarding inclusion and diversity in the workplace. Some stated that these initiatives raise awareness and potentially could encourage more diversity in the application process, as well as reduce bias in hiring practices. Further, a few participants stated that they have seen the direct impact of DEI initiatives through fellowship programs, mentoring, training, and improved job conditions.

Finally, participants were asked how diversity, equity, and inclusion initiatives impact employee retention at their workplace through an open-ended question. As with the previous question, many participants did not feel able to answer the question, either because they lacked personal experience or were not involved in HR processes. However, some participants discussed how certain DEI initiatives have contributed to retention by creating a more welcoming environment and fostering a sense of belonging and openness. While some participants felt that the DEI initiatives were effective, several believed that they were not helping diverse individuals, either due to structural issues in the workplace that continue to 'hold marginalized people back' or because they 'draw attention to minority colleagues without offering support'. There were mixed results in general, with some participants reporting very positive efforts in their workplace and others stating that the efforts were ineffective.

Discussion

Based upon our 2023 survey of data professionals from around the world, we can determine that data workers in a range of positions and areas have similar and overlapping concerns that are centered on three major areas: the work duties and responsibilities that they perform in their current positions; the pathways and incentives that facilitate and retain data workers in their posts and in the field more broadly; and the impacts, needs, and potential for DEI initiatives on data workers within organizations and across the data workforce.

Activities, Skills, Training, and Development

From our 2023 survey data, participants report spending most of their time sharing/publishing data and describing/organizing data, and noted that they needed the most training in the areas of data analysis and

data integration, including with modeling and visualization (Table 6, Table 7). Additionally, participants called the lack of data management skills "a key challenge" and noted that their major concerns were focused around keeping up, conceptually and technologically, with data formats and standards. The many tools, methods, and data types across different domains are challenging to manage (Rod et al. 2023), and lack of collaboration or standardization are mentioned as barriers to success. Due to resource constraints, research data workers feel pulled in many directions and have lots of different demands on their time. A wide range of services are necessary to support the scale and complexity of data at different parts of the research data lifecycle, but the skills and capacity are not necessarily in place, congruent with previous studies about the maturity of the research data management landscape in academic libraries and beyond (Cox et al. 2017).

Preliminary responses indicated that formal education was useful for the respondents initially, but that much of their relevant development has occurred as part of fieldwork, internships, workshops/professional development, and hands-on training. Many of the survey participants' entry to the field predated formal academic programs (see Table 2) that centered on research data management work, which likely has impacted and shaped their professional experiences. Additional research in this area is needed to further assess whether deficits in training, skill building, and development are the key concerns, as the open responses suggest. This work can also inform recommendations around ongoing education (Thompson et al. 2013), understanding current tasks and competencies (Bishop et al. 2021), making data infrastructure work visible to stakeholders and administrators (Suchman 1995; Gray et al. 2018), and extrapolating future workforce and development needs.

Entering and Staying in the Data Workforce

Entry into the data management workforce has been through a diverse range of pathways over the previous decades. Along with formal data management and curation programs in LIS and professional development opportunities within LIS (the RDMLA and MANTRA are examples), workshops and open learning courses (numerous UK Jisc-funded projects, including RDMRose and DataPool), and adjacent to LIS (such as the DataONE webinars and Carpentries training workshops), many data workers enter into a professional data management or data administration role through their work in a broad range of scientific or other fields. Specific research domains maintain their own methods, tools, languages, and data types. This results in a diverse range of experiences, needs, challenges, and opportunities for data workers.

Participants in our study described a combination of education and expertise in domain science, LIS/digital archives, technology and coding courses, and other data management work that prepared them for their current positions. Some respondents described the importance of internships, fieldwork, and residency programs for meeting the qualifications necessary for their current roles. Overall, respondents indicated on the Likert scale that job security/salary/benefits, work/life balance, and culture were important for retention in their workplace. Participants noted that 'culture, opportunities to make significant contributions, training, and mentoring' are factors that will influence their choices about whether to stay in a position. Anxiety

about keeping up, learning new skills, and finding meaningful opportunities in work are concerns that were raised throughout the survey. Future research can build on these insights by further examining key and secondary opportunities, challenges, and barriers faced by those seeking entry and working to stay in the data workforce.

DEI Initiatives and Workforce Opportunities

There has been increased attention to DEI in professional spaces, including in data science and data management work, over the past decade (D'Ignazio and Klein 2020). Examining what impact organizational DEI initiatives have where data professionals work seems important to understanding culture, belonging, and organizational self-reflection.

The survey data regarding organizational DEI initiatives were relatively positive about the promise of the efforts, but provided mixed responses about the results or long-term effects of DEI statements, programs, workshops, or organizational policies. While participants noted some positive concrete results, including mentoring and fellowship programs, as well as a greater sense of recognition and opportunity for inclusive practice, others felt that structural issues meant that not much real change had occurred. More broadly, these findings indicate a clear need for more examination of both the positive and negative outcomes of DEI efforts that impact the data workforce.

Conclusions and Future Work

This study investigated the following research questions: *Which data-related activities (from the DataONE data life cycle and the UCLA Data Science Center Data Literacy Course Competencies) does the data workforce spend the most time doing, and, of those, in which do they need more training and professional development? What are the facilitators and barriers to entry in the data workforce? What are the facilitators and barriers to staying in the data workforce? and What DEI initiatives are impacting the data workforce?* Through analysis of data from the 2023 survey (mixed multiple choice, Likert scale, and open-ended questions), we have identified key areas of focus to explore as factors for workforce participation and development in the data fields.

As with all survey studies, limitations include self-reporting and limited participants. While we sent out several recruitment emails to diverse recruitment channels, the sample size is still relatively small. However, the field itself is small and we did receive international participation and strong North American participation. Future work could expand both the participant population and complete smaller focus groups or interviews for more detailed, rich data.

This research builds upon previous foundations for understanding the activities that data management professionals perform in their daily work. Competency-based assessments are useful for identifying training needs and informing both formal research data management education and ongoing professional

development opportunities (Bishop et al. 2021). However, this project also diverged from detailing needs, skills, and competencies to further explore how and why data workers enter the field, and what helps them to stay, thrive, and actively participate. Understanding the opportunities that shape how current data professionals enter the workforce, along with the barriers to entry and retention, can help organizations identify, calibrate, and construct supportive policies, infrastructures, and other resources. Further exploration of what data workers see as the impacts and promise of DEI efforts within the workplace will inform future initiatives for creating equity, inclusive spaces, and belonging, all of which are important for ongoing work toward a more diverse data workforce. Drawing together these three components (data activities in daily work; entry and barriers to the field; and impact of DEI efforts) has allowed us to triangulate overlapping factors that all contribute to the work conditions of data management professionals in the field. Finally, the study and analysis present practical insights and actionable data for future work that will support current and future data workers.

Author Contributions

Authors made equal contributions.

Data Availability

The data supporting these findings are openly available through the following Figshare repository: https://doi.org/10.6084/m9.figshare.28012625.

Survey questions are available under the article Supplementary Files:

Appendix A: Data Workforce Survey Questions

References

Bishop, Bradley, Matthew Cowan, Hannah Collier, Matthew Mayernik, and Peter Organisciak. 2022. "Job Analyses of Earth Science Data Managers: A Survey Validation of Competencies to Inform Curricula in Research Data Management Education." *Journal of Education for Library and Information Science* 64 (2): 104-119. https://doi.org/10.3138/jelis-2021-0023.

Botticelli, Peter, Bruce Fulton, Richard Pearce-Moses, Christine Szuter, and Pete Watters. 2011. "Educating Digital Curators: Challenges and Opportunities." *International Journal of Digital Curation* 6 (2): 146-164. https://doi.org/10.2218/ijdc.v6i2.193.

Bresnahan, Megan and Andrew Johnson. 2013. "Assessing scholarly communication and research data training needs." *Reference Services Review* 41 (3): 413-433. https://doi.org/10.1108/RSR-01-2013-0003.

Byatt, Dorothy, and Wendy White. 2013. "Research data management planning, guidance and support: a DataPool Project report." Southampton, GB. University of Southampton. http://eprints.soton.ac.uk/id/eprint/351027.

Corrall, Sheila. 2012. "Roles and Responsibilities: Libraries, Librarians and Data." In *Managing Research Data*, edited by Graham Pryor, 105–134. Facet. https://doi.org/10.29085/9781856048910.007.

Cox, Andrew M., Eddy Verbaan, and Barbara Sen. 2014. "A Spider, an Octopus, or an Animal Just Coming into Existence? Designing a Curriculum for Librarians to Support Research Data Management." *Journal of eScience Librarianship* 3 (1): e1055. https://doi.org/10.7191/jeslib.2014.1055.

Cox, Andrew M., Mary Anne Kennan, Liz Lyon, and Stephen Pinfield. 2017. "Developments in Research Data Management in Academic Libraries: Towards an Understanding of Research Data Service Maturity." *Journal of the Association of Information Science & Technology* 68 (9): 2182-2200. https://doi.org/10.1002/asi.23781.

Data Curation Network. (n.d). "Data Curation Network." Retrieved from https://datacurationnetwork.org.

Data Management Skillbuilding Hub. (n.d.). "DataONE." Retrieved from https://dataoneorg.github.io/Education.

DataONE. (n.d.). "DataONE, Data Observation Network for Earth." Retrieved from https://www.dataone.org.

Digital Curation Centre. 2023. "Digital Curation Centre." Retrieved from https://dcc.ac.uk.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Gold, Anna. 2010. "Data Curation and Libraries: Short-Term Developments, Long-Term Prospects." California Polytechnic State University, San Luis Obispo. https://digitalcommons.calpoly.edu/lib_dean/27.

Gray, Jonathan, Carolin Gerlitz, and Liliana Bounegru. 2018. "Data Infrastructure Literacy." *Big Data & Society* 5 (2): 205395171878631. https://doi.org/10.1177/2053951718786316.

Kim, Youngseek, Benjamin K. Addom, and Jeffrey M. Stanton. 2011. "Education for eScience Professionals: Integrating Data Curation and Cyberinfrastructure." *International Journal of Digital Curation* 6 (1): 125-138. https://doi.org/10.2218/ijdc.v6i1.177.

Latham, Bethany. 2017. "Research data management: defining roles, prioritizing services, and enumerating challenges." *The Journal of Academic Librarianship* 43 (3): 263-265. https://doi.org/10.1016/j.acalib.2017.04.004.

Mattern, Eleanor, Jeng, Wei, He, Daqing, Lyon, Liz and Brenner, Aaron. 2015. "Using participatory design and visual narrative inquiry to investigate researchers' data challenges and recommendations for library research data services." *Program: electronic library and information systems* 49 (4): 408-423. https://doi.org/10.1108/PROG-01-2015-0012.

National Research Council, Committee on Future Career Opportunities and Educational Requirements for Digital Curation. 2015. "Preparing the Workforce for Digital Curation." Washington, D.C.: National Academies Press, April 22. https://doi.org/10.17226/18590.

Nel, Marguerite. 2020. "Information behaviour and information practices of academic librarians: a scoping review to guide studies on their learning in practice." 2020. Paper presented at *ISIC, the Information Behaviour Conference*, Pretoria, South Africa, September 28, 2020 - October 1, 2020. https://doi.org/10.47989/irisic2020.

Palmer, Carole L, Cheryl A. Thompson, Karen S. Baker, and Megan Senseney. 2014. "Meeting Data Workforce Needs: Indicators Based on Recent Data Curation Placements." In *iConference 2014 Proceedings*, iSchools, 2014. https://doi.org/10.9776/14133.

Peterson, Ashley and Ibraheem Ali. 2021. "Data Literacy Core Competencies-UCLA Library." Retrieved from https://osf.io/fb7hg.

Reisner, Barbara A., K.T.L. Vaughan, and Yasmeen L. Shorish. 2014. "Making Data Management Accessible in the Undergraduate Chemistry Curriculum." *Journal of Chemical Education* 91 (11): 1943-1946. https://doi.org/10.1021/ed500099h.

Rod, Alisa Beth, Biru Zhou, and Marc-Étienne Rousseau. 2023. "There's no "I" in Research Data Management: Reshaping RDM Services Toward a Collaborative Multi-Stakeholder Model." *Journal of eScience Librarianship* 12 (1): e624. https://doi.org/10.7191/jeslib.624.

RDMLA. (n.d.). "RDMLA Research Data Management Librarian Academy." Retrieved from https://rdmla.github.io.

Soyka, Heather. 2025. "Exploring Workforce Dataset A". figshare. https://doi.org/10.6084/m9.figshare.28012625.v1.

Stanton, Jeffrey M., Youngseek Kim, Megan Oakleaf, R. David Lankes, Paul Gandel, Derrick Cogburn, and Elizabeth D. Liddy. 2011. "Education for eScience Professionals: Job Analysis, Curriculum Guidance, and Program Considerations." *Journal of Education for Library & Information Science* 52 (2): 79-94.

Suchman, Lucy. 1995. "Making work visible." *Communications of the Association for Computing Machinery* 38 (9): 56-64. https://doi.org/10.1145/223248.223263.

Tibbo, Helen R., and Wendy Duff. 2008. "Toward a digital curation curriculum for museum studies: A North American perspective." In *Proceedings of 2008 Annual Conference of the International Documentation Committee of the International Council of Museums*, Athens, Greece, September 15-18, 2008. https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2018/12/70_papers.pdf.

Thompson, Cheryl A., Megan Senseney, Karen S. Baker, Virgil E. Varvel, and Carole L. Palmer. 2013. "Specialization in Data Curation: Preliminary Results from an Alumni Survey, 2008–2012." *Proceedings of the American Society for Information Science and Technology* 50 (1): 1-4. https://doi.org/10.1002/meet.14505001151.

University of Edinburgh. 2022. "MANTRA Research Data Management Training." Retrieved from https://mantra.ed.ac.uk.

Virkus, Sirje, and Emmanouel Garoufallou. 2019. "Data Science from a Library and Information Science Perspective." *Data Technologies and Applications* 53 (4): 422-441. https://doi.org/10.1108/DTA-05-2019-0076.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Witt, Michael, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin. 2009. "Constructing Data Curation Profiles." *International Journal of Digital Curation* 4 (3): 93-103. https://doi.org/10.2218/ijdc.v4i3.117.