Strategic Planning for a Data-Driven, Shared-Access Research Enterprise: Virginia Tech Research Data Assessment and Landscape Study

Yi Shen

The data landscape study at Virginia Tech addresses the changing modes of faculty scholarship and supports the development of a user-centric data infrastructure, management, and curation system. The study investigates faculty researchers' current practices in organizing, describing, and preserving data and the emerging needs for services and education. The results demonstrate the changing nature of faculty demands regarding data documentation, storage, and archiving and identify opportunities for libraries to develop a coherent service, research, and education system to address the evolving needs.

In an effort to promote data availability, discoverability, and reusability, academic libraries are developing new models of research support, especially taking on the role of archiving and curating digital data. At varying degrees of investment and commitment, a relatively few academic libraries become directly involved in developing data repositories and offering curation support, while a lot more libraries are engaged in the planning for, oversight of, and provision of data management services. As public access and open data movements are gaining greater significance, new development demands systematic analysis and strategic planning for the new environment.

To address increasingly data-driven faculty scholarship, landscape research is necessary to investigate the changing nature of faculty needs in documenting, preserving, and archiving data. Despite the excitement and investment in rapidly evolving data fields, it is unclear to what extent and how much of the activity in data management, sharing, and reuse among faculty researchers involve productive engagement and how much is just fulfilling government mandates.

This project contributes to the practical understanding and methodological framework in scholarly data practices through the design and implementation of a research data landscape study. By applying a newly engineered assessment framework, this study explores what is really happening in a specific research environment and identifies the level of engagement among faculty researchers in different aspects of data-related activities. It demonstrates the full potential of libraries' data and information

Yi Shen is Research Environments Librarian in Newman Library at Virginia Polytechnic Institute and State University; e-mail: yishen18@vt.edu. © 2016 Yi Shen, Attribution-NonCommercial (http://creativecommons.org/licenses/by-nc/3.0/) CC BY-NC.

doi:10.5860/crl.77.4.500

expertise and what they can offer to partner with faculty in different dimensions of data-centric academic work.

This study is grounded in the institutional context of Virginia Tech (VT). With research expenditures of nearly $500 million underway to advance development and foster interaction in science, engineering, medicine, and the arts, VT has a dynamic research data ecosystem that provides a unique opportunity to examine strategies for building the data infrastructure, management, and curation system.[1] The institution's strategic plan for 2012–2018 identifies the needs and challenges of a data-driven networked society. It notes, "the questions that can be asked and the methods and data sets that can be used to solve complex problems are being fundamentally altered by technology and the information sciences. Being effective in this environment means being able to apply and manage information technology while taking advantage of networking, collective intelligence, simulation, data mining, and modeling."[2]

Targeted at supporting changing modes of scholarship, particularly the increasing adoption of new data-driven methods and technologies, this study investigates how data are being stored, managed, shared, and reused by VT faculty and researchers. It examines open access requirements and faculty's attitudes toward data creation and sharing. The results identify core data practices and services necessary, determine challenges and opportunities, and strategize data services, support, and training efforts. This work contributes to the understanding of the research community's emerging need for data services and education. It highlights the various demands of faculty researchers and the critical roles that libraries play in supporting them to effectively manage data.

By offering practical insight into the challenges of data documentation and management, the study stimulates strategic thinking and decision making on what the institution can do to support data stability and maintenance. The implementation of this research also helps increase the awareness among faculty researchers about the wide-ranging data-related topics and stewardship activities that need to be exercised, as well as potential value that may be lost due to inappropriate care or management of data. The objectives are to increase the institutional support for a changing data culture and to promote service innovations and research partnerships in the University Libraries. The long-term goal is to increase research productivity by adding value to data through careful handling, preservation, and curation. Such work will inform future research and development aimed to build the institutional capabilities for sustainable lifecycle data management.

Acting as a building block of the regional and national repository networks, the development of institutional data management capabilities also has broader implications for a sustainable, cohesive, and interoperable global network. Most recently, three major open access repository networks, including OpenAire, LA Referencia, and Shared Access Research Ecosystem (SHARE), along with the Confederation of Open Access Repositories (COAR) and Center for Open Science (COS), have agreed to collaborate on regular exchange of data, common metadata and vocabularies, and technological development.[3] To be better positioned in such broad networks, it is critical to understand local instances to build synergies and make alignments across repositories to achieve an integrated global network.

Literature Review

Global Efforts in Building Data Infrastructure and Sharing Ecosystems

It is important to preserve research data, both qualitative and quantitative, whether in the natural and social sciences or in the humanities, to enable new discovery, reuse, repurposing, creative integration, and content aggregation. The world's most pressing challenges such as climate change, environmental sustainability, health, and economics require discipline-integrated research and sound science-driven policies that are critically dependent on trusted, global-scale, and interconnected data.[4] The collection, organization, and documentation of research data help realize the essential value of data as important components of the research record and scholarly communication. As such, developing and supporting long-term stewardship of quality-assured research data and related services are becoming widely recognized by governments, scientific communities, and academic institutions worldwide. The dynamic, interdisciplinary, and global nature of data ecosystems are reflected in government efforts to develop public access and sharing policies, as well as to seek international standards and the best practices for data interoperability and exchange.

In recent years, the emphasis on data-driven scientific innovation and economic growth has stimulated many national and international initiatives in building a vibrant, open-data ecosystem by making data more "liquid"—open, widely available, and in shareable formats.[5] For example, the Obama Administration's "Data to Knowledge to Action" event called for harnessing the power of data by forging new partnerships in data innovation.[6] The United Kingdom (UK) government has released the strategy for seizing the data opportunity and building UK data capability.[7] Australian National Data Service (ANDS) established collaboration with Thomson Reuters to include Australian research and data in the Data Citation Index to support the discovery of global data sets.[8] The European Union (EU) partners are working to build international consensus on long-term strategy and policy in the area of research data.[9] The Research Data Alliance (RDA) including the United States, the European Union, Australia, and a growing number of nations is striving to build a global scientific data infrastructure to facilitate the exchange and interoperability of data across disciplines and national boundaries.[10]

Data Infrastructure Development and Research Data Assessment

Charged with the mission to generate, make accessible, and preserve new knowledge and understanding, research universities already own and operate key pieces of the infrastructure in addition to the valuable content created through research and scholarship. These include digital institutional repositories, Internet2, Digital Preservation Network (DPN), and more. To meet federal research requirements for public access, an increasing number of academic institutions have recently invested in data management planning to comply with grant application mandates and foster full-lifecycle data management, discovery, and reuse. Academic libraries in particular have been actively exploring ideas, evolving roles, building expertise, developing tools, and establishing services in the area of research data management (RDM).[11]

In response to the White House Memorandum on "Increasing Access to the Results of Federally Funded Research," higher education entities have started to develop the Shared Access Research Ecosystem (SHARE).[12] These include the Association of American Universities (AAU), Association of Public and Land-grant Universities (APLU), and Association of Research Libraries (ARL). Through SHARE, universities collaborate with the federal government and other commercial sectors to host long-term preservation and cross-institutional digital repositories for public access research and open data.[13] As an active member of these higher education entities, Virginia Tech is committed to becoming a linked node in this federated, consensus-based open data and research system.

As more academic institutions work toward SHARE and invest in RDM, it is increasingly important to identify and assess institutional data assets for effective and sustainable management and long-term preservation. Among the various stakeholders, academic libraries must actively engage in this fast-developing landscape and lead data assessment to champion data services and sharing efforts.[14]

Promoting sharing and deriving valuable insights from shared data will "require new rules and procedures and new attitudes as well as investments in technology and capabilities."[15] Given the recent development of new techniques to collect, archive, and distribute data, it is timely to study the current practices, identify opportunities and gaps in data use, and plan for services at the institution, especially within the broad landscape of national and international data movements. To do so, this research addresses two major questions: 1) what are the practices and needs of faculty scholarship with regard to data? and 2) how could libraries collaborate and develop systems, policies, and training to support such efforts?

Method

This section describes the research method and process. It includes detailed information on the development of the assessment framework and the survey instrument, the participants, and the data collection and analysis process.

Assessment Framework and Survey Instrument

The research design adopted multiple theoretical and practical frameworks. First, the study referred to the Data Asset Framework (DAF) developed by the UK Digital Curation Center (DCC) and adopted elements from the example questionnaires and interview frameworks already tested by the DAF pilot studies and early exemplars.[16] Second, the study incorporated the Community Capability Model Framework (CCMF) developed by UKOLN of the University of Bath and Microsoft Research Connections.[17] By doing so, the survey instrument not only identifies individual data habits but also profiles institutional and community readiness and capability for performing data-intensive research. For example, throughout the survey, questions were asked about whether there are established community standards, policies, and practices for data access and preservation and whether there are shared metadata frameworks and broadly adopted documentation schemes. Finally, the current survey also referenced DataOne's scientists and research data survey, as well as other institutional data management surveys (such as Emory University's survey of faculty RDM practices and perspectives) and planning questionnaires (like Johns Hopkins University's data management planning questionnaire).[18]

Being a multifaceted and multilevel assessment, the survey integrated value judgments, risk analysis, and potential return to better understand the potential of data for deposit. In terms of value judgments, for example, there are questions about data with potential to be reused or repurposed, concerns associated with sharing, and priority placed on services, training, and support. As for risk analysis, examples include data most at risk of loss or cases with penalties for misuse. Regarding potential return, questions relate to whether data could be ingested into repositories for sharing.

Building upon previous frameworks, models, and templates, the current survey includes six specific data-related practice areas ranging from production, storage, description, sharing and access, use and reuse, to training and services. This framework provides the guidelines necessary to understand data characteristics and management needs and informs the process of developing data services and educational programs. The survey includes 32 questions and uses skip patterns to tailor the survey to respondents' experiences.

Participants

The targeted population includes Teaching and Research (T&R) faculty and Research faculty as defined in the Virginia Tech's faculty handbook.[19] They are from the following colleges:

- College of Agriculture and Life Sciences (CALS)
- College of Architecture and Urban Studies (CAUS)
- Pamplin College of Business
- College of Engineering (COE)
- College of Liberal Arts and Human Sciences (CLAHS)
- College of Natural Resources and Environment (CNRE)
- College of Science (COS)
- Virginia-Maryland College of Veterinary Medicine (VA-MD Vet Med)

The full list of faculty names and e-mail addresses was collected from the Virginia Tech Department of Human Resources and then imported into the Qualtrics web survey software as a CSV file for survey distribution. A total of 2,532 e-mail invitations were sent and 652 responses were received, among which are 423 completed entries. The respondents could skip any questions while taking the survey. The college distribution of the survey respondents who clearly identified their affiliations is shown in figure 1, and the demographic characteristics of the survey respondents who answered the relevant questions are described in table 1. All appropriate human subjects procedures were approved and followed under VT IRB-14-825.

FIGURE 1
College Distribution of Survey Respondents

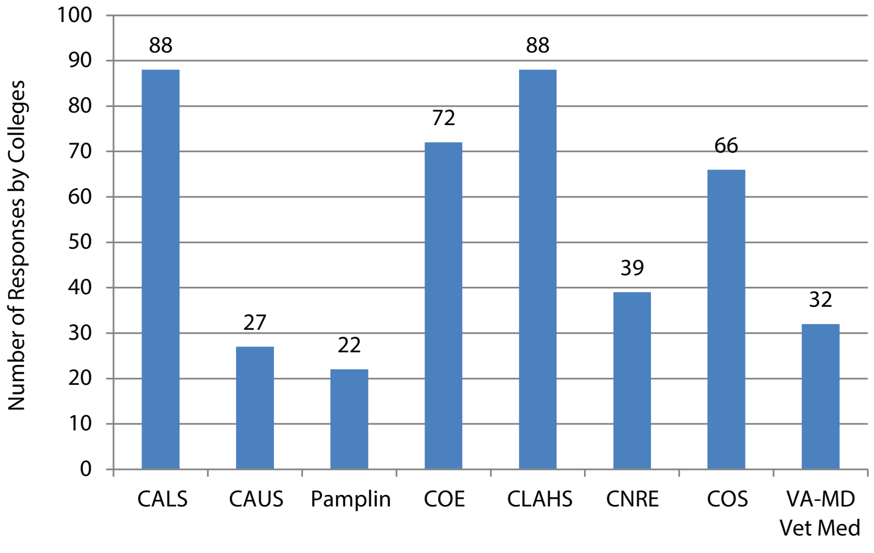


TABLE 1
Demographic Characteristics of Survey Respondents

	Gender	Rank	Age
Total Responses	609	569	603
Subgroups	Male: 365 (60%) Female: 210 (34%) Prefer not to respond: 34 (6%)	T&R Assistant Prof.: 115 (20%) T&R Associate Prof.: 138 (24%) T&R Prof.: 146 (26%) Research Faculty: 170 (30%)	< 36: 142 (24%) 36–45: 171 (28%) 46–60: 187 (31%) >60: 103 (17%)

Data Collection and Analysis

The formal data collection took place in November 2014. To promote interest and boost responses, the University Libraries garnered support from the Virginia Tech Office of the Vice President for Research (OVPR) and adopted various communication and outreach channels. These included meeting with the Associate Vice President for Research Programs and other OVPR personnel and a presentation to the Commission on Research. The Scholarly Communication Librarian also dedicated a blog post on Open@VT to raise awareness and advocate the significance of this project.[20]

Before the formal launch of the survey, extensive review and feedback were received from the library faculty in the Research and Informatics division and from liaison librarians. The survey instrument was then pilot-tested by the cross-campus faculty representatives serving on the University Library Committee. Several iterations of review and feedback helped refine the survey questions and polish the wording. For a final check, the principal investigator consulted with the Director of the Virginia Tech Center for Survey Research for further suggestions and improvement. Such a rigorous process of reviewing, testing, and editing supported the effective design and robust development of the research instrument.

Statistical analysis and modeling were performed to identify patterns and to inform the current state and future development of data-related practices and services. The results provide insights into the socio-technical dimensions of VT research environment and data landscape.

Findings

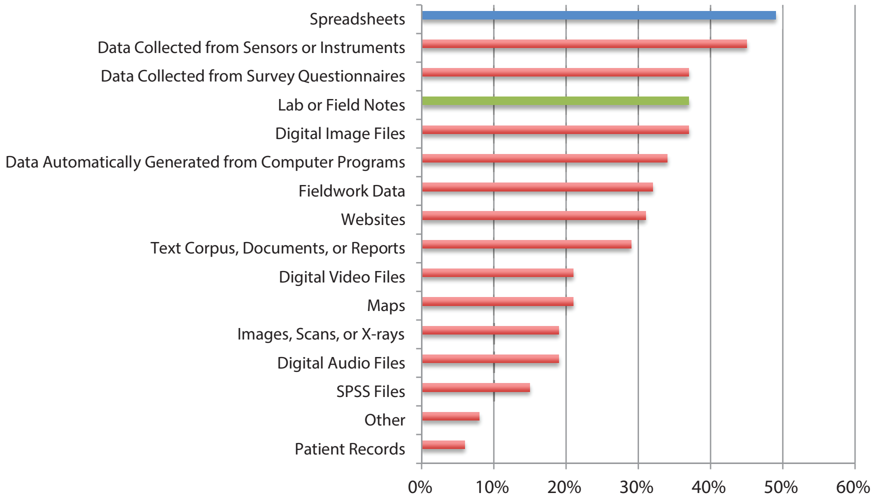
The research aimed to take stock of the data being created and held within the institution, to characterize the data, and to understand data-sharing practices and expectations of VT faculty and researchers. Accordingly, the survey questions concern research data holdings, current data management practices, as well as educational needs and service requirements related to data. Selected results are reported below.

Research Data Characteristics

The survey asked about the basic characteristics of digital research data that faculty have created and maintained in the course of their research during the previous 18 months. For the purpose of this study, research data are interpreted broadly and can be quantitative or qualitative, including (but not limited to) numerical data produced by computational experiments, output from experimental equipment, calibration information, simulation outputs, images or audiovisual files, survey results, interview transcripts, text files, spreadsheets, websites, digital information artifacts, or databases compiled from documentary sources. Questions were asked about the nature, types, forms, and formats of data, as well as estimated size of the data. It is notable that, regardless of discipline studied or methods used, the highest percentage of the researchers used some sort of spreadsheet application to analyze, manipulate, or share research data (see the blue bar in figure 2). As a common data form, spreadsheet could be easy to exchange with other researchers; but, when sharing and preservation are not adequately considered, there can be difficulties when using and interpreting other researchers' spreadsheet data.[21] Lab and field notes (see the green bar in figure 2) are another form of data that often get lost in transition and encounter major preservation and sharing problems. Data management and curation services should be highly attentive to these issues when providing training and developing resources and support.

The survey also asked faculty researchers to indicate whether their research data are transdisciplinary, multidisciplinary, or interdisciplinary in nature. Table 2 shows the definition and response rate for each of these categories. The results show that a

FIGURE 2
Forms of Data (n = 544)



fair number of faculty researchers (a total of 27%) considered their data to have these boundary-crossing natures. This is considerable, given the institution's strategic vision and priorities placed on developing a multidisciplinary workforce and building pathways for interdisciplinary success.[22] Within such context, it is particularly meaningful to further explore how emerging cross-boundary work negotiates data management, preservation, and sharing processes and how data are being defined and redefined in the process of producing multidisciplinary, interdisciplinary, and transdisciplinary knowledge.

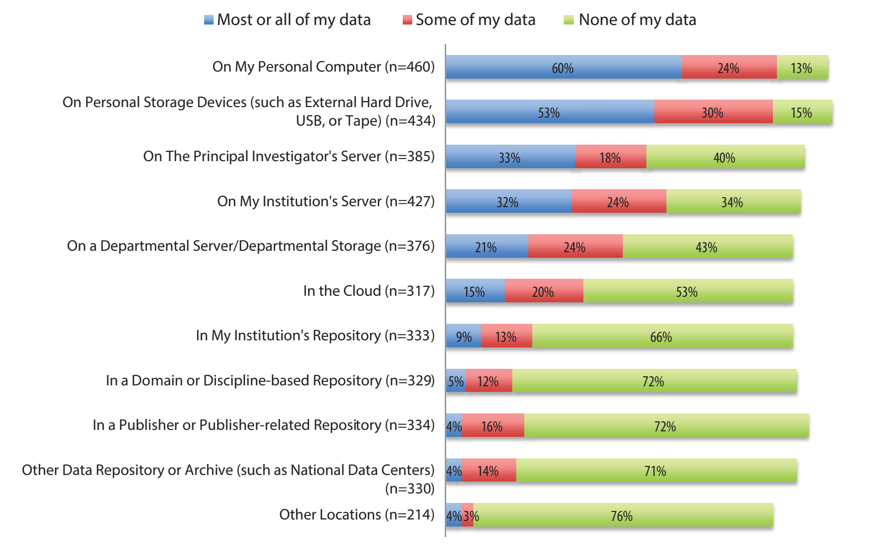
Research Data Storage

The survey asked faculty researchers a set of questions focused on their data storage and backup practices. Figure 3 demonstrates how much of the data that the faculty researchers plan to store in the various locations during or after the project(s) is/are completed. The results show that their data storage mainly stays at a personal level, either on personal computers or personal storage devices. Formal repository systems such as institutional, domain or disciplinary-specific, publisher or publisher-related,

TABLE 2
Nature of Research Data (n = 584)

Nature of Research Data	Response	%
Transdisciplinary (that is, to transcend two or more discipline perspectives and traditional boundaries to form a new holistic approach)	36	6%
Multidisciplinary (that is, to combine or contrast multiple discipline perspectives in an additive manner)	35	6%
Interdisciplinary (that is, to synthesize or harmonize two or more discipline perspectives to form an integrated and coherent level of understanding)	85	15%

FIGURE 3
Summary of Data Storage Choices



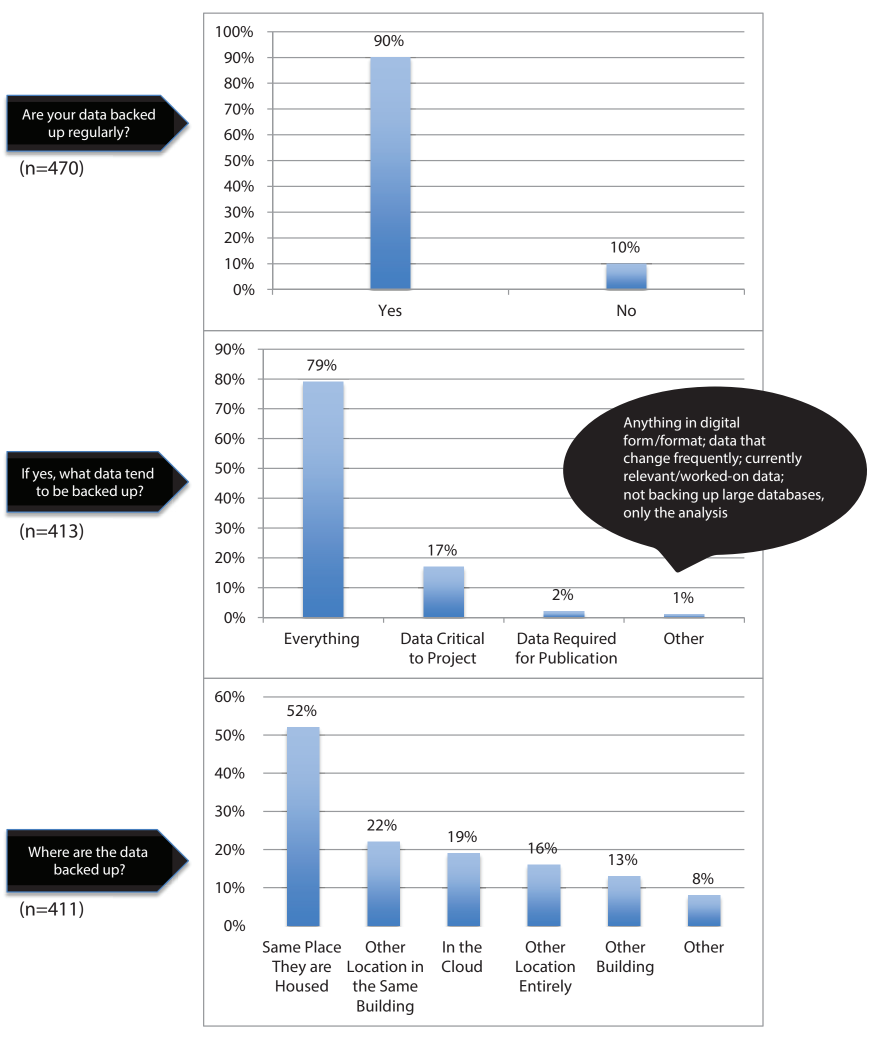
and other types of repositories or archives (such as national data centers) are rarely chosen for data storage. Figure 4 indicates that data backup is commonly practiced, but mostly on the same device where the original data are housed. However, the best practices for data backup involve using mixed media for storage and maintaining three copies: the original files, a local external copy, and a remote external copy.[23] This survey demonstrates that faculty researchers need guidance and strategies for developing a plan of data backups, security, and preservation. As one participant commented:

> "Our department has no plan for backing up data, and there is no support (i.e. no server, no statement or policy on backup strategies). Every dept. should have a plan for backing up computers in offices and in labs, and should not rely on the individual PI to purchase an external hard drive or cloud space. There should be some type of centralized server at the College level on which each PI has space available to store or archive important data. This server should be housed in another building and backed up daily without intervention by PI's."

Research Data Documentation and Management Practices

Data description and documentation are central to managing, preserving, and sharing research data. Metadata standards and documentation schemes play critical roles here. Many current approaches are exploring how various kinds of lightweight metadata can be used to better facilitate data exploration.[24] For more complex search systems, there have been focused efforts in the development of more domain-specific metadata. For example, Research Data Alliance (RDA), as an international coordinating entity, has created a number of interest groups exploring the development of metadata standards for specific scientific fields and a general working group building a catalog of metadata standards for better search and exploration. Then the question is, in reality, how do researchers describe or document research data? As shown in figure 5, a total of 457 faculty responded to this question and about half (49%) reported no

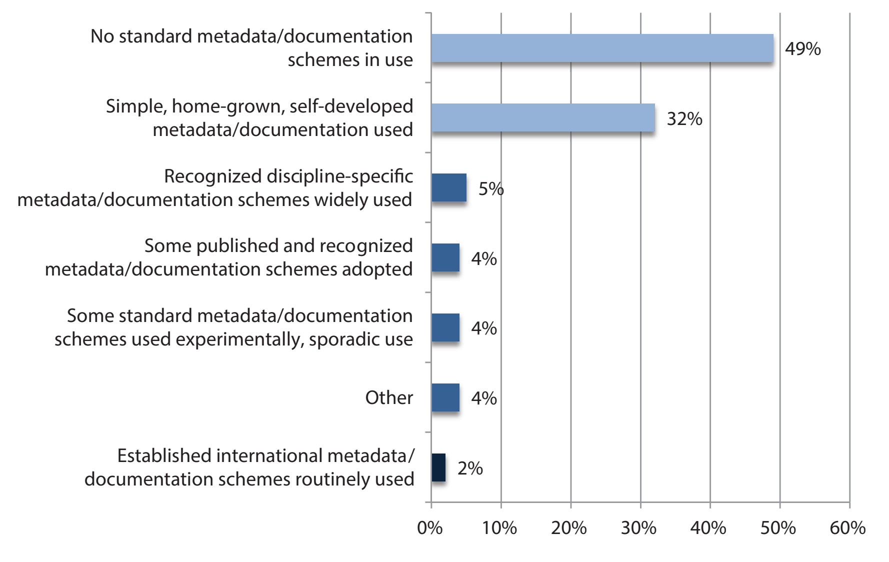
FIGURE 4
Data Backup Practice



standard metadata and documentation schemes in use, while 32 percent indicated the use of simple, home-grown, self-developed metadata and documentation. Only a small fraction, ranging from 4 to 5 percent, are using some form of published or recognized standards, while only 2 percent are using established international metadata standards or schemes. It seems that RDA needs to increase its impact on the adoption of standards.

Quality control and validation are closely connected with reproducibility and support experimental transparency.[25] The survey asked faculty about their modes of quality assurance with regard to data. Respondents were asked to select all approaches that apply to their work. As illustrated in figure 6, the results show that self-review of data is most common among 81 percent of 456 respondents, followed by the check of data by colleagues or team members among 58 percent of the re-

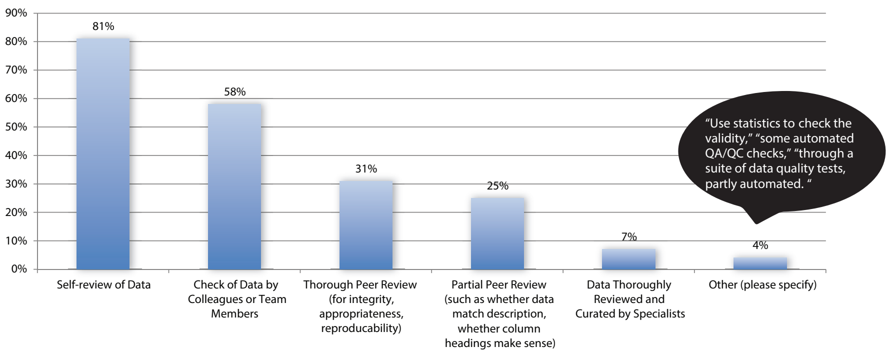
FIGURE 5
Data Description Practice (n = 457)



spondents. Nearly one-third (31%) reported thorough peer review, and one-quarter (25%) indicated partial peer review of the data. The use of peer review is expected to grow, given the rising publishers' requirements to deposit and share research data associated with publications. Only a low percentage (7%) of the researchers indicated that their data are thoroughly reviewed and curated by specialists, and this is the type of exercise that academic libraries should foster and support in current and future services.

Next, the respondents were asked to check all data management issues they have encountered. The most common issues identified are poor naming and filing systems, migration to new formats, platforms, or storage media, as well as obsolete hardware and software environments (see figure 7). These issues are often encountered during the active use of data when conducting research. One respondent commented:

FIGURE 6
Quality Control and Validation Approaches (n = 456)



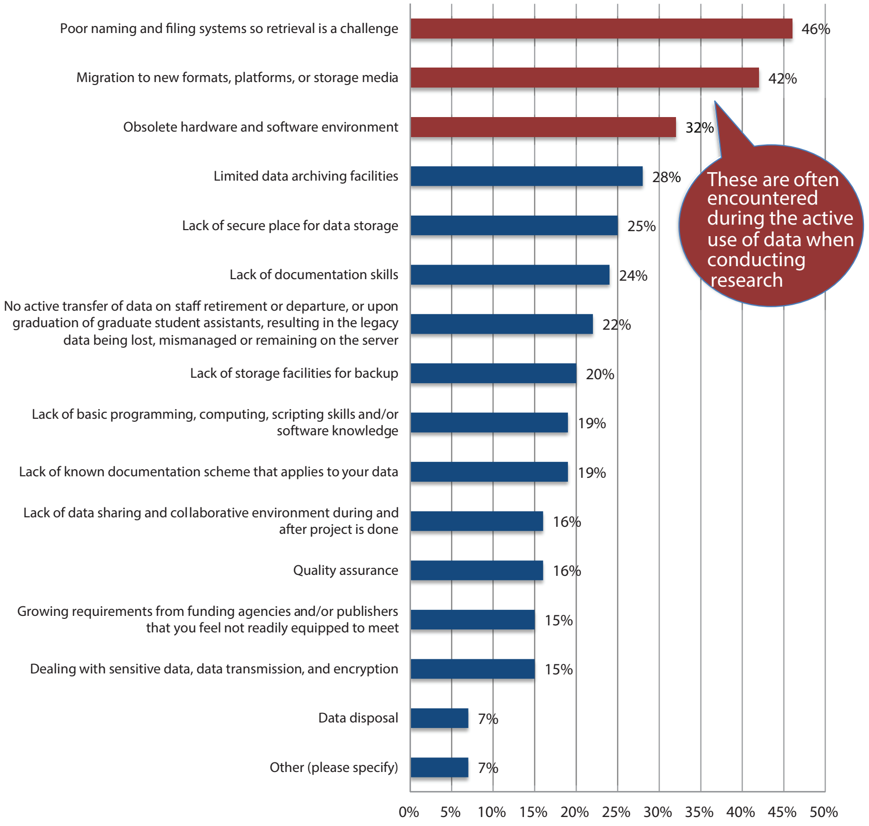
"Use statistics to check the validity," "some automated QA/QC checks," "through a suite of data quality tests, partly automated. "

> "I am space physics research[er]. There are more and more observations based from ground and satellite instruments, but most of them are in different formats. So I have to learn how to convert them into the format my code could process. If there is any solution for better accessing different format of data, it would be great help."

The next set of issues relates to data storage and archiving, including documentation, followed by sharing and collaboration, which also concern quality assurance. Meeting funders' and publishers' requirements in data management is a concern among a comparatively small pool of 15 percent of the respondents. The ranking of data management issues encountered by the faculty indicates that data services and solution providers should prioritize attention and responses accordingly.

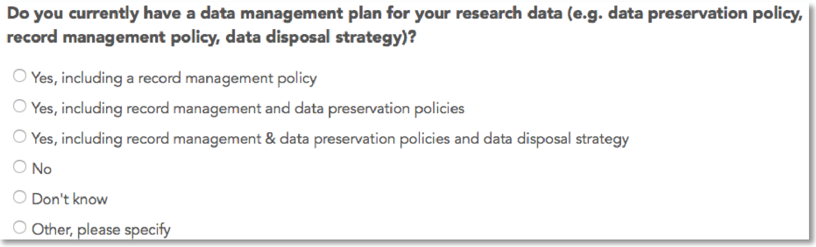
The participants also described other data management issues. These included the "difficulty anticipating future needs that would allow them to set up the datasets better initially," a "lack of dedicated person to oversee data," "too much data to easily manage," "server downtime and maintenance," the "inability to acquire data permanently," and the lack of "means by which to share data confidentially." All these issues demand solutions in data management, archiving, preservation, and curation. Additionally, time is a major concern, especially there are "limited time and personnel to devote to extensive documentation" and archiving. Location is another major

FIGURE 7
Summary of Common Data Management Issues (n = 416)



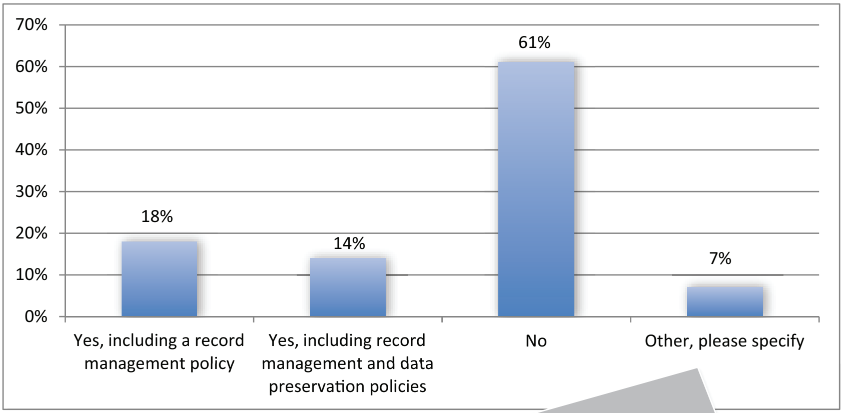
concern; for example, there are problems associated with "moving [data] from one machine to another with a different path" or not "knowing how to access available storage on different computers and devices," as well as limited " data transfer speed." There is an expressed desire of "one [centralized] location where the team spread across institutions can store or access data easily." As such, the major challenges are how to facilitate seamless access across distributed data collections or sources and "streamline data for faster access and transfer."

FIGURE 8
Current State of Data Management Planning (n = 374)



Do you currently have a data management plan for your research data (e.g. data preservation policy, record management policy, data disposal strategy)?

○ Yes, including a record management policy
○ Yes, including record management and data preservation policies
○ Yes, including record management & data preservation policies and data disposal strategy
○ No
○ Don't know
○ Other, please specify



"We have a plan, but we don't really follow it"
"Yes, but it is not formalized"
"Yes, but a self-maintained record management"
"A basic plan but no record management policy"
"I have a personal policy/habit that includes very limited record mgmt, complete digital data preservation, and paper/filed note preservation for a limited time (followed by disposal)."
"Data are retained in perpetuity. Most sensitive data (MRI) have home-grown record management policy."
"Wrote DMP for recent grant, but haven't implemented it yet"
"Have a plan...have not really implemented it yet."
"My 'current' research is in a stage so preliminary that data management has not yet been addressed."
"We have developed DMP for new proposals but existing data are managed in old fashion"
"Depends on the project – NSF requires"
"It depends on the research project. Some projects require a careful data management plan. I do not have a master plan for all of my data though, which is much needed"
"No universal policy"
"Yes, but I don't recall the details; And, it tends to be project specific (i.e., based on requirements of the grant)"
"No 'policy,' since my 'data' is not the kind that requires a 'record management policy' or 'data disposal strategy.' Guided by ethical questions in the discipline."

Data Management Planning

As we know, government funding agencies increasingly require the submission of Data Management Plans (DMPs) together with grant applications. Most important, it constitutes the first step of data stewardship for preservation, sharing, and reuse. As such, a set of questions focused on the current status of planning. As shown in figure 8, the results suggest a majority of the faculty respondents (61%) do not have a DMP. Of those with DMPs, 18 percent only have a record management policy and 14 percent have both record management and data preservation policies. None of the faculty have a whole set of record management, data preservation, and data disposal policies in place. Comments of the respondents centered around three major themes: the respondents either have a personal, informal plan that may not be closely followed, are in transition to DMPs for new projects that still need to be implemented, or have no formalized plan or policy across projects. In all three circumstances, libraries could step in offering support or partnering with researchers on active management or curation of data to prepare for future archiving. This is further discussed in the last section of this paper.

Education Needs and Service Scoping to Improve Efficiency of Faculty Working with Data

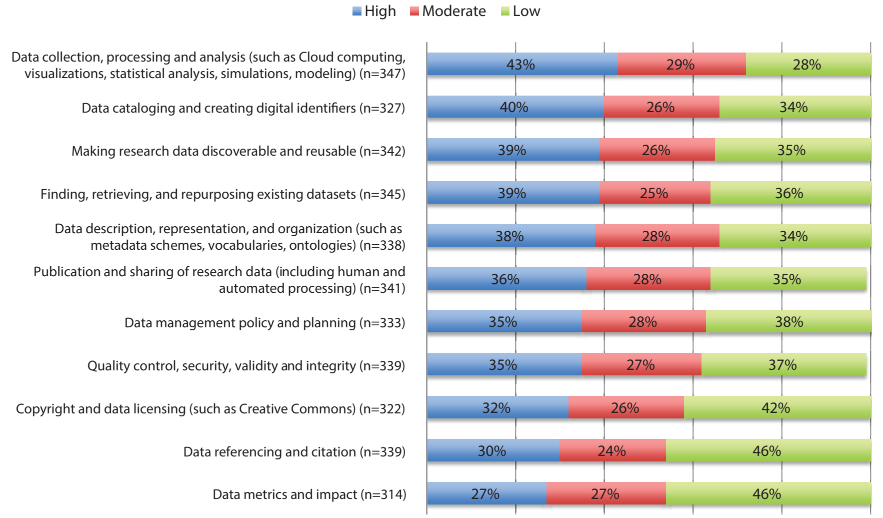
In the final section, the survey asked about the level of data-related educational needs and support services desired by faculty researchers. As big data and data sciences have been gaining significance over the recent years, there has also been a surge of interest in data collection, processing, and analysis techniques, including cloud computing, visualizations, statistical analysis, simulations, and modeling. Working with a growing amount of diverse data from internal and external sources or collaborating on large projects requires a better understanding of how to organize data and create unique identifiers to make data discoverable and reusable. It also requires the ability to find, retrieve, and repurpose existing data sets. Consequently, the faculty respondents indicated educational needs in a wide range of data and information science topics.

As funding agencies' and journal publishers' data management and sharing requirements become mandates, the faculty researchers have demonstrated their desire to better understand the range of data management issues, such as description and documentation, sharing and publication, policy and planning, quality control and security, as well as copyright and licensing. Other topics such as data referencing and citation as well as data metrics and impact at the other end of research lifecycle are certainly gaining some interest as well, although at a relatively lower rate. This may be due to the fact that data sharing and reuse cultures have not been developed, and the full effects of data literacy required in these other areas at the early stages of the lifecycle have not yet propagated to actual data use and reach. Figure 9 illustrates the educational needs across the range of data science, information science, archiving, and curation topics. Note that the percentages shown are rounded numbers.

Other participants' comments indicated the need to learn how to "combine parts of different existing datasets from repositories" and the demand for "courses on mid- to long-term data storage, meta analysis, and use of public data, etc." One participant's comment summarized the need to "create a culture of using, documenting, and analyzing all types of data."

Aside from educational needs, the survey also asked about the level of data-related support and services desired by faculty. Figure 10 shows the results. Long-term data storage and archiving was ranked as the top service needed. Next, support and services involving data preservation were desired. These included preparing and archiving data for long-term preservation, technical support on format migration and long-term data integrity, as well as guidance on documenting data and metadata. The respondents also demonstrated interest in active data storage. Guidance and support for data analysis was

FIGURE 9
Level of Educational Needs to Work Efficiently with Data



ranked at the lower end of the scale compared to its top ranking in educational needs. This may be due to the fact that researchers are more inclined to learn how to perform data analytics themselves to organically integrate analytical techniques into their theoretical reasoning and research programming. In theory, cross-disciplinary data sharing, integration, and collaboration are projected as a major advancement in the future of research inquiry and scientific discovery. In reality, these are largely in the early stages of development. This may explain why the needs for cross-disciplinary data-sharing repositories and collaboration platforms are not as immediate as in these other areas.

Finally, one participant articulated other desired support:

> "I can use help in development of programs to extract and merge data properly and efficiently from large commercial datasets (e.g., Compustat, CRSP, IBES etc.). This is an important task for my research projects, but one that is only needed occasionally. It would be very helpful to have a College wide or university wide resource who could help with these needs on a fairly efficient basis."

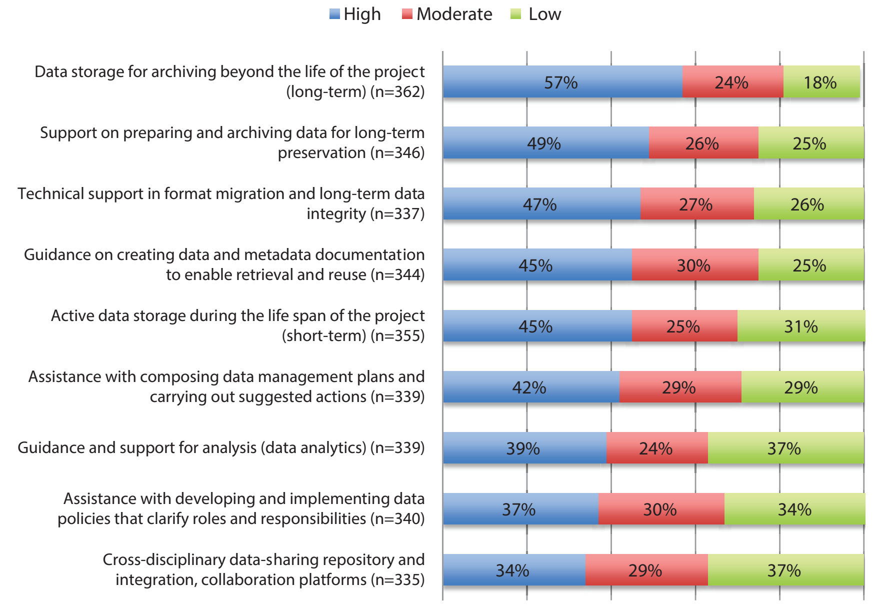
Discussion and Conclusions

This section draws on the study results and divides the discussion into three parts. First, the socio-technical dimensions of the research data landscape and the associated challenges are discussed. Then a changing data culture and its practical implications are described. Finally, the strategies for developing a coherent data management, research, and education system are highlighted. The paper concludes by outlining opportunities for further research.

Mapping the Data Landscape

This research identifies the lack of systematic planning and preservation activities, limited backup and storage options, as well as sporadic and informal documentation practices among the faculty researchers when working with data.

FIGURE 10
Level of Data-Related Support and Services Needed



There is a clear need for long-term data-archiving support. One respondent pointed out that departmental data-serving resources are "ephemeral." Another described the holding of multiple data sets "in the cloud that need a more formal home." And one other researcher indicated the need for "initiative to provide data archiving." There are certainly limitations in the current institutional repository system to accommodate data deposit and archiving requirements, and the researchers expressed the desire for a more integrated system between institutional repository and disciplinary-specific or community-driven data repositories.

> "There is no facility to store large datasets at the University that does not have a fee attached to it. The question is really who is responsible for archiving, the researcher that generated the data or the institution where the work was performed. Seems like funds from overhead dollars should support data storage and management. The University will archive a thesis but not the data." "Data storage capabilities and maintaining data confidentiality are non-existent at VT."

> "VTechWorks is a great idea, but there is very little support, at least that I am aware of, in terms of helping PIs and grad students get data into the facility. I would have used it long ago, but this barrier makes it very time consuming and therefore difficult to use it. This needs to change if VT PIs are going to avoid major time burdens in complying with current federal grant requirements." "My understanding is the library repository is not curated and regularly updated to comply with changing formats. There are repositories such as DataDryad that do this. Are we partnering with any of these services?"

Among the faculty participants, concerns over technical operationalization, trust, security, and control, as well as the practicability of data access and discovery are still persistent. To address these concerns, we must work on user adoption, facilitate communication, and demonstrate measurable academic, social, and economic values that result from appropriate data stewardship. As information technology communities are striving to optimize platforms and integrate applications, a common culture of collaborative scholarly communication needs to be forged. This should rigorously include data.

A data-driven shared-access research enterprise has great implications for unlocking the possibilities of interdisciplinary engagement and international expertise. When supporting interdisciplinary applications and the reuse of data for new purposes, extra care is needed for data curation. Particularly, efforts are needed to systematically communicate the various dimensions of research data, to fully describe the characteristics and attributes of data, and to carefully develop shared terminology to connect different domain cultures and research communities.[26] Throughout the lifecycle of data, even more complex curatorial activities are needed to accommodate the fluidity and dynamics of interdisciplinary data interactions. This entails developing necessary metadata and indexing around data objects in support of cross-disciplinary networking and integration.

A Changing Data Culture

The current research scenario shows a gradual transition in data culture and the changing perception of faculty researchers, demonstrated by the high level of demands in data-related support and services for long-term preservation and access. These include highly ranked areas such as data storage for archiving beyond the life of project, support on preparing and archiving data for long-term preservation, technical support for format migration, as well as guidance on creating data and metadata documentation to enable retrieval and reuse. These are followed by support and services needed in active data storage during the lifespan of projects. Notably, the levels of demand in the above areas were all ranked higher than the level of demand in assistance with data management planning and implementation. Such trends indicate the increased awareness and broader interest of the academic community in deeper data-related issues and preservation values than the simple concern of fulfilling funder requirements and government mandates.

This changing culture is also demonstrated in the examples given by the faculty on their current data practices, the issues and challenges they have encountered, and their ongoing sharing efforts.

> "1. I lead an effort to collect continuous data related to Stroubles Creek (weather, hydrology, water quality) and share it with many faculty for classes and research. It takes incredible time and manpower that is hard to fund through normal grants.
>
> 2. I am working on a project where I am trying to find, collect, and make data available from studies done in VA (mostly at VT) from the 1930's–1990's. These are watershed studies and the data are invaluable, but hidden in closets. I also have an archive of photos of many of these research efforts that I am working with Imaging and Repository Initiatives, some of which are hosted already at Discovery Commons.
>
> 3. I am working with [another researcher] on two projects related to data sharing and dark data (collected data that are no longer available). One is working to

> post data in CUAHSI's Water Data Center which is funded mostly by NSF, and the other is related to Earth Cube."
>
> "My long-term plans involve translating XML documents into an RDF repository and exploring the possibilities using NoSQL databases for purposes of sharing and graphing."
>
> "Most biological scientists are not very familiar with 'big data' formats and use. We are just starting to expand our capability to develop, prepare data in the format that is user friendly."
>
> "We are making an effort thru CUAHSI [Consortium of Universities for the Advancement of Hydrologic Science]."

Strategizing a Coherent Data Management, Research, and Education System

There is the need to develop a pipeline of services, technical support, and educational activities to enhance the overall capacities of the institution in data management, documentation, archiving, and preservation.

As the current data-documenting practices of faculty researchers are rather localized and informal, there is a great need for careful selection and application of metadata standards (in regard to data types, relevant disciplinary standards, and repository standards) to enhance or supplement the simple, home-grown, self-developed metadata and documentation provided by researchers themselves to ensure broader access and long-term use. Extra curatorial efforts are needed to capture and refine contextual representation information of data and to track complex relationships among data components and types. Guidance and assistance are certainly needed to help researchers transit from localized micropractices to more standardized, community-sensitive approaches.

To deal with the scaling issues of today, computational methods are being developed to support data publishing, including automated quality control and validation of domain-specific data (such as biodiversity data), and machine-aided reviewing. To support automated handling, standardized methods and procedures need to be developed. Particularly, data structures and formats for specific data types need to be registered and formalized (like RDA's Data Type Registries).

To further support data integration and knowledge representation, research data services need to develop logic-based methods for aligning or merging immediately related disciplinary taxonomies and conceptual frameworks. Such efforts will bridge the collaborative gap and support the necessary coordination of different data sets (for example: economic datasets, geographic data sets, and census data) to make strategic predictions and solve complex societal problems.[27] Research data management (RDM) not only deals with large-scale data collections, but also handles numerous small data collections that are sporadic, complex, heterogeneous, and widely dispersed.[28] To streamline the various services and support the flow of information across data lifecycle, RDM also needs to understand the gaps and bridge the continuum between curation and analytics.

This study also identifies important educational opportunities. Some faculty researchers equated data management issues (including record management, data preservation, and data disposal) with IRB policies. IRB addresses confidentiality and sensitivity issues related to human subjects that need to be taken into consideration while conducting data management and sharing activities, but it is not intended to deliberately and thoroughly address data management. In the field of data management, there is existing guidance for de-identifying human subjects data for sharing. Techniques are also developed for assessing disclosure risk and removing personal identifiers from research data that make the data sharable while following IRB and HIPAA guidance. These distinct topics need to be introduced to clarify different concepts, improve understanding, and develop practical skills among faculty researchers.

Other topics such as how to prepare research spreadsheets for sharing, how to find, retrieve, and repurpose existing data, and how to prepare for data archiving should also be introduced. "Given the nuanced nature of each researcher's specific data needs," the faculty respondents expressed the demand for project-specific, deep-dive data curation support. This should be done "by technical staff that have the bandwidth and purview" to "talk in-depth with" and "work with individual researcher to understand her/his needs" and to "set-up a data management and archiving plan."

Furthermore, the faculty participants expressed broad educational needs ranging from data science techniques (such as data aggregation and analytics), library and information science subjects (such as organization, search, and retrieval), archival topics (such as preservation and metadata), to data curation strategies (such as value-added data sub-setting, documentation, or cross-disciplinary functionalities). Accounting for the evolving and encompassing needs of the academic community as related to data, there are significant opportunities for libraries to build cohesion in these different areas of specialization to align with the dynamic, intersecting scholarly endeavors and research expertise.

Many academic fields (like humanities and natural sciences) are transforming, often adapting to new data-driven methods and technologies. The libraries at Virginia Tech are recruiting and developing a new breed of data librarians to support a wide range of topics in data and information sciences and are striving to integrate information technologies, data curation, and data analytics within the knowledge production process of faculty and researchers. As we are building a nexus for informatics research—to support health informatics, environmental informatics, decision-support and business informatics, and other ever-emerging informatics fields at Virginia Tech—there are significant opportunities to provide a common ground for different informatics researchers to jointly tackle a wide variety of problems and to facilitate new types of use, research, and analysis of the valuable research records created by the community. In addition to bridging knowledge pools, deep scholarship from inside the libraries will play critical roles in informatics research. As we regularly work to solve complex information, communication, and service problems, we need to engage in the analysis of actual research processes, working situations, and specific data practices that requires in-depth evaluation of contextual variables and nuanced factors. Uniquely positioned in knowledge representation and information management, the libraries at Virginia Tech can become a locus for informatics research in science and scholarship.[29]

Future Research

This study provides an overall understanding of the current research data landscape and supports the scenario mapping and strategic planning for a data-driven, shared-access research ecosystem at the institution. While valuable in planning system and developing services, the study also has its limitations. In particular, the current lack of common knowledge and varying levels of understanding among the faculty researchers with regard to data management, preservation, sharing, discovery, and reuse might potentially lead to unreliable responses to the survey. On the one hand, it reinforces the necessity of identifying educational opportunities through assessing the current level of understanding and faculty demands in data management. On the other hand, it is also necessary to reconduct the survey at different key stages when broader understanding is developed and when widespread care and management of data is adopted. Conducting the survey later at major milestones of development could test the reach and impact of data stewardship activities and educational efforts at the institutional level. It can also serve as a measure of success for data services and support provided by the libraries and other partners, such as central IT and the Office of the Vice President for Research.

Further research could look at how different disciplines organize and govern the management of shared data resources and how existing and emerging multidisciplinary, interdisciplinary, and transdisciplinary fields mediate, steer, and transform this process. As the social and academic value of various types of data could change over time, it is useful to look at how the value changes and what new "markets" for data and data repositories could be created. As we continue to support community building and knowledge networking, it is important to understand the ongoing handling and maintenance of data, as well as how data become defined and redefined in the production of knowledge.

Notes

1. Research at Virginia Tech, "Areas of Research" (2015), available online at http://www.research.vt.edu/areas-research [accessed 27 February 2015].

2. The Virginia Tech Office of University Relations, "A Plan for New Horizon: Envisioning Virginia Tech 2012-2018" (2012), available online at http://www.president.vt.edu/strategic-plan/2012-plan/2012-strategic-plan.pdf [accessed 27 February 2015].

3. SHARE (SHared Access Research Ecosystem) News, "Major Repository Networks Agree to Collaborate on Data Exchange, Technological Development, and Metadata" (2015), available online at http://www.share-research.org/2015/07/major-repository-networks-agree-to-collaborate-on-data-exchange-technological-development-and-metadata/ [accessed 27 July 2015].

4. The World Data System (WDS), "Trusted Data Services for Global Science" (2015), available online at https://www.icsu-wds.org/ [accessed 27 July 2015].

5. McKinsey Global Institute, "Open Data: Unlocking Innovation and Performance with Liquid Information" (2013), available online at http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information [accessed 9 March 2015].

6. The Office of Science and Technology Policy (OSTP) and Networking and Information Technology R&D (NITRD), "Data to Knowledge to Action: Building New Partnerships" (2013), available online at http://www.nitrd.gov/nitrdgroups/index.php?title=Data_to_Knowledge_to_Action [accessed 9 March 2015].

7. The UK Department for Business, Innovation & Skills, "UK Data Capability Strategy: Seizing the Data Opportunity" (2013), available online at https://www.gov.uk/government/publications/uk-data-capability-strategy [accessed 9 March 2015].

8. Thomson Reuters, "Thomson Reuters Collaborates with Australian National Data Service to Raise the Profile of Research Data" (2013), available online at http://thomsonreuters.com/press-releases/112013/Thomson-Reuters-ANDS [accessed 9 March 2015].

9. SURF Collaborative Organization for ICT in Dutch Higher Education and Research, "European Landscape Study of Research Data Management" (2013), available online at http://www.surf.nl/en/actueel/Pages/EuropeanLandscapeStudyofResearchDataManagement.aspx [accessed 9 March 2015].

10. The Board on Research Data and Information (BRDI), "Symposium on Global Scientific Data Infrastructures" (2012), available online at http://sites.nationalacademies.org/pga/brdi/PGA_070715 [accessed 9 March 2015].

11. Karen Antell, Jody Bales Foote, Jaymie Turner, and Brian Shults, "Dealing with Data: Science Librarians' Participation in Data Management at Association of Research Libraries Institutions," *College & Research Libraries* 75, no. 4 (2014): 557-74; David Fearon, Jr., Betsy Gunia, Barbara E. Pralle, Sherry Lake, and Andrew L. Sallans, *Research Data Management Services* (SPEC Kit 334) (Washington, D.C.: Association of Research Libraries, 2013); Michelle Kahn, Richard Higgs, Joy Davidson, and Sarah Jones, "Research Data Management in South Africa: How We Shape Up," *Australian Academic & Research Libraries* 45, no. 4 (2014): 296-308.

12. The Office of Science and Technology Policy (OSTP), "Memorandum for the Heads of

Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research" (2013), available online at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [accessed 9 March 2015].

13. The Association of American Universities (AAU), Association of Public and Land-grant Universities (APLU), and Association of Research Libraries (ARL), "Shared Access Research Ecosystem (SHARE) Draft Proposal" (2013), available online at http://www.arl.org/storage/documents/publications/share-proposal-07june13.pdf [accessed 9 March 2015].

14. Tyler Walters and Katherine Skinner, *New Roles for New Times: Digital Curation for Preservation* (Washington, D.C.: Association of Research Libraries, 2011).

15. McKinsey Global Institute, "Open Data: Unlocking Innovation and Performance with Liquid Information," 12.

16. Digital Curation Center (DCC), "Data Asset Framework Implementation Guide" (2009), available online at http://www.dcc.ac.uk/resources/repository-audit-and-assessment/data-asset-framework [accessed 7 November 2013]; Sarah Jones, Alexander Ball, and Çuna Ekmekcioglu, "The Data Audit Framework: A First Step in the Data Management Challenge," *The International Journal of Digital Curation* 3, no. 2 (2008): 112-20.

17. UKOLN of the University of Bath and Microsoft Research Connections, "Community Capability Model for Data-intensive Research" (2013), available online at https://communitymodel.sharepoint.com/Documents/CCMF-Profile.xlsx [accessed 7 January 2014].

18. DataOne, "Scientists and Research Data: Continuing to Build an Understanding of Your Data Needs" (2013), available online at http://www.dataone.org/news/help-us-understand-how-scientists-work-data [accessed 7 January 2014]; Emory University Research Data Management, "Faculty Practices and Perspectives on Research Data Management" (2012), available online at http://guides.main.library.emory.edu/content_mobile.php?pid=333927&sid=3327853#box_3327853 [accessed 7 January 2014]; Johns Hopkins University Data Management Services, "JHU DMS Data Management Planning Questionnaire" (2013), available online at http://dmp.data.jhu.edu/assistance/nsf-data-management-plans/#Questionnaire [accessed 7 January 2014].

19. The Virginia Tech Office of the Senior Vice President and Provost, "Faculty Handbook Chapter 02: Policies and Procedures for All Faculty" (2014), available online at http://www.provost.vt.edu/faculty_handbook/chapter02/chapter02.html [accessed 11 November 2014].

20. Philip Young, "The Research Data Assessment Survey at Virginia Tech" (2014), available online at https://blogs.lt.vt.edu/openvt/2014/11/20/the-research-data-assessment-survey-at-virginia-tech/ [accessed 21 November 2014].

21. Johns Hopkins University Data Management Services, "Guidance on Aspects of Data Management" (2015), available online at http://dmp.data.jhu.edu [accessed 9 March 2015].

22. The Virginia Tech Office of University Relations, "A Plan for New Horizon: Envisioning Virginia Tech 2012-2018," 4-10.

23. Johns Hopkins University Data Management Services, "Backup Strategies for Research Data" (2013), Internal Training Sessions Handout.

24. James Hendler, "Data Integration for Heterogeneous Datasets," *Big Data* 2, no. 4 (2014): 205-15.

25. Joel Achenbach, "The New Scientific Revolution: Reproducibility at Last," *Washington Post*, January 27, 2015, available online at http://www.washingtonpost.com/national/health-science/the-new-scientific-revolution-reproducibility-at-last/2015/01/27/ed5f2076-9546-11e4-927a-4fa2638cd1b0_story.html [accessed 9 March 2015].

26. Tiffany C. Chao, Melissa H. Cragin, and Carole L. Palmer, "Data Practices and Curation Vocabulary (DPCVocab): An Empirically Derived Framework of Scientific Data Practices and Curatorial Processes," *Journal of the Association for Information Science and Technology* 66, no. 3 (2015): 616–33.

27. National Science Foundation, "NSF-supported Research Data Alliance/U.S. Collaborates with International Partners to Accelerate Data Sharing" (2012), available online at http://www.nsf.gov/news/news_summ.jsp?cntn_id=126010 [accessed 9 March 2015].

28. Yi Shen and Virgil E. Varvel Jr., "Developing Data Management Services at the Johns Hopkins University," *The Journal of Academic Librarianship* 39, no. 6 (2013): 552–57.

29. The author thanks Julie Speer and Tyler Walters for providing administrative support throughout the project, the VT Libraries' Research & Informatics faculty and Liaison Librarians for providing useful comments on the questionnaire, as well as the anonymous reviewers, Gail McMillan, and Philip Young for providing thoughtful feedback during the final production of this paper.