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

Yi Shen
Virginia Tech
Blacksburg, VA
yishen18@vt.edu

ABSTRACT
The data landscape study at Virginia Tech addresses the changing modes of faculty scholarship increasingly driven by data and supports the development of a user-centric data infrastructure, management, and curation system. The study investigates faculty researchers’ current practices in organizing, describing, sharing, reusing, and preserving data and the emerging needs for data services and education. The results determine the different dimensions of data-related activities, challenges, and opportunities and the changing nature of faculty demands regarding data documentation, storage, and archiving. Asking faculty researchers to self-reflect data sharing and reuse from both data producers’ and data users’ perspectives, the study reveals a significant gap between the rather localized and sporadic data management and sharing activities and the highly perceived reuse values of data that often get lost right after the original work is done. With expertise in data, information, and archive fields, libraries have great values to offer in bridging the gap and building a data-driven, shared-access research enterprise.

Keywords
Data science, information behavior, digital libraries.

INTRODUCTION
Against the backdrop of global efforts in building scientific data infrastructure (e.g. Research Data Alliance) and national initiative in developing Shared Access Research Ecosystem (SHARE), data sharing and open access movements are quickly gaining momentum in academic communities. Guedon (2015) has pointed out, “across the centuries, researchers have learned to share their papers, now they must learn to share their data.” With such understanding, “the leaders of the scientific community are recalibrating their requirements, pushing for the sharing of data and greater experimental transparency” (Achenbach, 2015).

Among the many literature and press releases, reproducibility is one core argument for data sharing to avoid the so-called “data dredging” maneuver (Achenbach, 2015) or “over-fitting of data” (Provost & Fawcett, 2013) in which researchers go on a deep dive for something publishable that may turn out to be a “statistical fluke” (Achenbach, 2015). Another core argument for data sharing is the tremendous value of reusing or repurposing data. Especially as new analytical techniques become usable, academics may want to explore their data in ways that were not planned for in the original design of their data collections. Scientists and scholars may also be increasingly looking at how to integrate their structured data holdings with those of others and to explore links to both internal and external unstructured data sources (Hendler, 2014). As a result, discovery of and access to data outside their own control will become even more crucial.

In order to support these efforts in an academic institution, we need to ask the important questions: what are faculty researchers’ behaviors and attitudes towards data sharing and reuse? How accessible and discoverable are their data? What are the major concerns regarding reusing and repurposing data among faculty researchers?

PROBLEM STATEMENT
To address the changing landscape and new directions in faculty scholarship, especially the fast-expanding data-driven scholarship, data landscape research is needed to investigate the changing nature of faculty needs regarding the spectrum of support and services in sharing, preserving, and organizing 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 applies a newly engineered research data assessment tool in the institutional context of Virginia...
Figure 1. Reuse or repurpose values of faculty researchers’ own data (n=368)

Tech (VT) to investigate how data are being stored, managed, shared, and reused by VT faculty and researchers.

RESEARCH METHODOLOGY
Targeted at a multifaceted and multilevel assessment, a data collection survey instrument was developed with the incorporation of multiple frameworks. These include the Data Asset Framework (the UK Digital Curation Center, 2009) and its pilot studies (Jones et al., 2008), the Community Capability Model Framework (CCMF) (UKOLN of the University of Bath & Microsoft Research Connections, 2013), DataOne’s scientists and research data survey (DataOne, 2013), as well as other institutional data management surveys (e.g. Emory University Research Data Management, 2012) and data management planning questionnaires (e.g. JHU Data Management Services, 2013). The survey questionnaire was pretested by cross-campus faculty representatives serving on the University Library Committee.

The formal data collection using Qualtrics web survey took place in November 2014 and targeted at Teaching and Research faculty (T&R) and Research faculty at Virginia Tech. They are from eight colleges ranging from agriculture and life sciences, architecture and urban studies, business, engineering, liberal arts and human sciences, natural resources and environment, science, to veterinary medicine. A total number of 2532 email invitations were sent and 652 responses were received, among which are 423 completed entries. The respondents can skip any questions while taking the survey.

Statistical analysis was performed to test hypothesis and recognize patterns. Content analysis was conducted on the qualitative responses of the participants to discern contexts and gather insights on the data-related practice, concerns, and needs of faculty researchers.

FINDINGS
Mapping the Data Landscape
A high percentage of the faculty researchers considered their own data to have reuse or repurpose values (see Figure 1). However, the openness of data, methodologies and workflows (see Figure 2&3), as well as the discoverability and accessibility of data continue to remain low. In practical applications, there has been a lack of regular or formalized data documentation practice. Often there are no standard metadata or documentation schemes being applied by the researchers, or only simple, home-grown, self-

Figure 2. Openness of data (n=426)

developed metadata and documentation schemes are used (see Figure 4). When in turn asked about their own data reuse practice and concerns, relatively fewer faculty researchers indicated reuse of existing data collected by others or from other disciplines. Their major concerns about reusing others’ data concentrate on the difficulty of finding or accessing reusable data, difficulty integrating data, and
possible misinterpretation of data (see Figure 5-7). The rather localized and sporadic data management and documentation practices of researchers as data producers are actually causing the obstacles they themselves (from data users’ perspective) often encounter when reusing others’ data. Consequently, the highly perceived potential values of data for future research are often lost right after the original work is done.

To the researchers, benefits of sharing information, exchanging knowledge, and extending digital platforms are apparent when it comes to research collaboration and scholarly publication, but not quite so when it comes to data. Without recognizing the importance of accessibility and discoverability, one respondent questioned the necessity for openly sharing data: “If someone wants to see the data from someone else’s research, don’t they just ask?” Similarly, others do not think of sharing data the same way as they’ve learned to share papers. Guedon (2015) pointed out that researchers should reflect on their role in the “greater scheme of scientific work” and “be socialized into the network vision of research activities,” which is sharing and collaborating, now at the level of data, just as they’ve learned to share their papers. Libraries can play significant roles in making data retrievable and discoverable by curating the data, providing the metadata, and within existing policies, exposing the data.

**Figure 5-7. Data reuse practice and concerns**

**Figure 8. Level of data-related support and services needed**

**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 in long-term preservation and access. These include highly ranked areas such as data storage for archiving beyond the life of the project, support on preparing and archiving data for long-term preservation, technical support in format migration and long-term preservation, 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 life span of projects. Notably, the levels of demand in the above areas were all ranked higher than the demand in assistance with data management planning and implementation (see Figure 8). 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 funding requirements and government mandates.

**Strategizing a Coherent Data Management, Research, and Education System**

As the current data documenting practices of faculty researchers are rather localized and informal, guidance and assistance are certainly needed to help researchers with the transition from a localized micro-practice to more standardized, community-sensitive approaches. Extra curatorial efforts are needed to capture and refine contextual representation information of data and to track complex relationships among data components and types.

Furthermore, the faculty participants expressed educational needs in a wide range of areas 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 issues (such as value-added data sub-setting, documentation, or cross-disciplinary
functionalities). Figure 9 shows the statistical results. Accounting for the evolving and encompassing needs of the academic community as related to data, there are significant opportunities for us to build cohesion in these different areas of specialization to align with the dynamic, intersecting scholarly endeavors and research innovations.

In addition to bridging knowledge pools, deep scholarship from inside the libraries will play critical roles in data and informatics research. As we regularly work to solve complex information, communication, and service problems, we need to engage in 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 data and informatics research in science and scholarship.

**SIGNIFICANCE AND IMPLICATIONS**

The study of data sharing and access activities helps determine the unique set of obstacles related to data production, use, and reuse. It reveals the potential future values of research data from both data producers’ and data users’ perspectives. It also identifies a major gap between the localized and limited data management and sharing activities and the highly perceived reuse or repurpose values of data that often get lost in the transition of research practitioners and communities of practice. By mapping the level of user engagement in “openness of data,” “openness of methodologies and workflows,” and “reuse of existing data,” this research contributes to the current state of knowledge in data sharing and open access and supports future development of data-related curation services. A further investigation of college-level engagement in “openness of data” also reveals the different patterns and concentration of activities in individual colleges and indicates the need to develop college-oriented strategies and approaches to a changing data culture.

**KEY REFERENCES**


---

**Figure 9. Level of educational needs to work efficiently with data**

<table>
<thead>
<tr>
<th></th>
<th>High</th>
<th>Moderate</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection, processing and analysis (e.g. Cloud computing, visualizations, statistical analysis, simulations, modeling) (n=347)</td>
<td>41%</td>
<td>28%</td>
<td>28%</td>
</tr>
<tr>
<td>Data cataloging and creating digital identifiers (n=327)</td>
<td>43%</td>
<td>26%</td>
<td>34%</td>
</tr>
<tr>
<td>Making research data discoverable &amp; traversable (n=342)</td>
<td>39%</td>
<td>20%</td>
<td>35%</td>
</tr>
<tr>
<td>Finding, retrieving and re-purposing existing data sets (n=345)</td>
<td>35%</td>
<td>25%</td>
<td>30%</td>
</tr>
<tr>
<td>Data description, representation and organization (e.g. metadata schemes, vocabularies, ontologies) (n=338)</td>
<td>38%</td>
<td>28%</td>
<td>34%</td>
</tr>
<tr>
<td>Publication and sharing of research data (including human and automated processing) (n=343)</td>
<td>26%</td>
<td>28%</td>
<td>35%</td>
</tr>
<tr>
<td>Data management policy and planning (n=335)</td>
<td>35%</td>
<td>28%</td>
<td>38%</td>
</tr>
<tr>
<td>Quality control, security, validity and integrity (n=335)</td>
<td>35%</td>
<td>27%</td>
<td>37%</td>
</tr>
<tr>
<td>Copyright and data licensing (e.g. Creative Commons) (n=322)</td>
<td>32%</td>
<td>26%</td>
<td>42%</td>
</tr>
<tr>
<td>Data referencing and citation (n=309)</td>
<td>20%</td>
<td>24%</td>
<td>46%</td>
</tr>
<tr>
<td>Data metrics and impact (n=314)</td>
<td>27%</td>
<td>27%</td>
<td>46%</td>
</tr>
</tbody>
</table>