Virginia Tech Data Landscape and Environmental Assessment:
Technical Briefing on Data Repository & Information System

Yi Shen

A faculty-wide data environmental scan and landscape study at Virginia Tech was conducted in 2015 and concluded with 652 responses received from Teaching & Research Faculty and Research Faculty in 8 different colleges. The survey asked basic characteristics and special features of digital research data that faculty currently create and hold in the course of their research. This study also explored faculty researchers' data storage and backup options, their data handling challenges and reuse concerns as well as needs and requirements for technical support and services. Below are selected findings and conclusions from this study.

Among the special data features identified, a few notable ones include the data form part of a larger data set, they are complex and have inter-relationships with other data sets, and the data have the potential to be integrated with other data sets from different disciplines or domains to answer large-scale, complex question. The complexity of scale and relationships between data sets demands effort to support more granular levels of data description and access that can enable elastic discoverability into either larger or smaller or specific segments of the data.

The researchers' data storage mainly stays at personal level, either on personal computers or personal storage devices. Data backup practice is commonly exercised among the faculty researchers, but mostly in the same place where the original data are housed. It is important to provide them with guidance and strategies for developing a plan for backups, security, and preservation for research data.

Most faculty researchers reported no standard metadata and documentation schemes in use, or only simple, home-grown, self-developed metadata and documentation being used. Only a small fraction of the faculty researchers are using some form of published or recognized standards.

In addition to common data issues such as poor naming and filing systems, challenging format or platform or storage media migrations, and obsolete hardware and software environments, other issues were also specified by the participants. These include: "difficulty anticipating future needs that would allow us to set up the datasets better initially," "lack of dedicated person to oversee data," " too much data to easily manage," "server downtime and maintenance," and so on.

Besides time restraint, location is another major factor of concern, for example, there are problems "moving [data] from one machine to another with a different path," or not "knowing how to access available storage on different computers and devices," and " data transfer speed." One major challenge is how to facilitate

seamless access across distributed data collections or sources and “streamline data for faster access and transfer.”

The respondents described their top needs in long-term data storage and archiving services and also expressed their interest in active data storage. In data preservation, they asked for technical support in format migration and long-term data integrity as well as support in preparing and archiving data for long-term preservation.

The participants also specified other required supports. For example, one stated, “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.”

Finally, one important requirement expressed is to support automated handling, such as extracting and merging data from large datasets. As a result, standardized methods and procedures need to be developed, in particular, data structures and formats for specific data types need to be registered and formalized.