Virginia Tech Data Landscape and Environmental Assessment:
Technical Briefing on Data Curation

Yi Shen

A *Virginia Tech Research Data Assessment and Landscape Study* was conducted in 2015 to take stock of the data assets being created and held within the institution and to examine data sharing practices and expectations of VT faculty researchers. Targeted at a multifaceted and multilevel assessment, this research assesses current repository strategies and user needs, and identifies changing modes of scholarly communication.

The survey asked faculty researchers a set of questions regarding the basic characteristics of their digital research data. These included nature, types, forms and formats of data, as well as estimated size of the data. Below are selected findings and conclusions.

As to the different forms of data, it is most likely that researchers use some sort of spreadsheet application to investigate, manipulate, or share research data regardless of discipline studied or methods used. Lab and field notes are another form of data that often get lost in transition and encounter major preservation and sharing problems.

There is a need for careful selection and application of metadata standards (in regard to data types and relevant disciplinary) to enhance or supplement the informal documentation provided by researchers themselves to ensure broader access and long-term use. For those that use standard metadata and documentation schemes, they provided a few examples as shown in Appendix 1.

The most common data management issues are poor naming and filing systems, migration to new formats, platforms, or storage media, and obsolete hardware and software environment. These are often encountered during the active use of data when conducting research.

As to the current status of data management planning, a majority of faculty researchers do not have a DMP. The researchers either have a personal, informal plan that may not be closely followed, or are in the transition of having DMPs for new projects that still need to be implemented, or have no formalized plan or policy across projects.

For those who do have a data management plan, most commonly, the principle investigator, graduate research assistants, or research project manager are responsible for carrying out and complying with the plan. The attached file (named “Funding Sources”) specifies their primary funding agencies, which could inform the

library services to monitor and anticipate possible movements of the relevant agencies that still don't have data management requirements.

Among the reasons why faculty researchers do not make their data openly available to others after project completion, the results indicate a majority of confidentiality or data protection issues. In this space, libraries could support putting data in "dark archive" and prepare for possible future release according to policies. The second and third major reasons for not sharing are the time and effort required and sharing being not required, as anticipated.

As to openness of data, the results suggest the larger percentage of pockets of activity and moderate activity, with over 50%. In these cases, data are shared within limited scope or under limited conditions, for example, data are described in literature but not made available, or data are available on request, after embargo or with other conditions. There are smaller fractions of widespread activity and complete engagement in openly sharing or making efforts in sharing data (24%). There are 12% of respondents indicated "no sharing and no details released for data" or in other words, nominal activity. Chart 6 shows the differences among the colleges in their community engagement in openness of data.

There are great interests in and educational needs for data collection, processing and analysis techniques, including cloud computing, visualizations, statistical analysis, simulations, and modeling. Working with increasing amount of diverse data, faculty researchers need a good understanding of how to organize data and create unique identifiers to make research data discoverable and reusable. It also requires the ability to find, retrieve and repurpose existing data sets. They certainly indicate educational needs and interests in archiving and curation techniques.

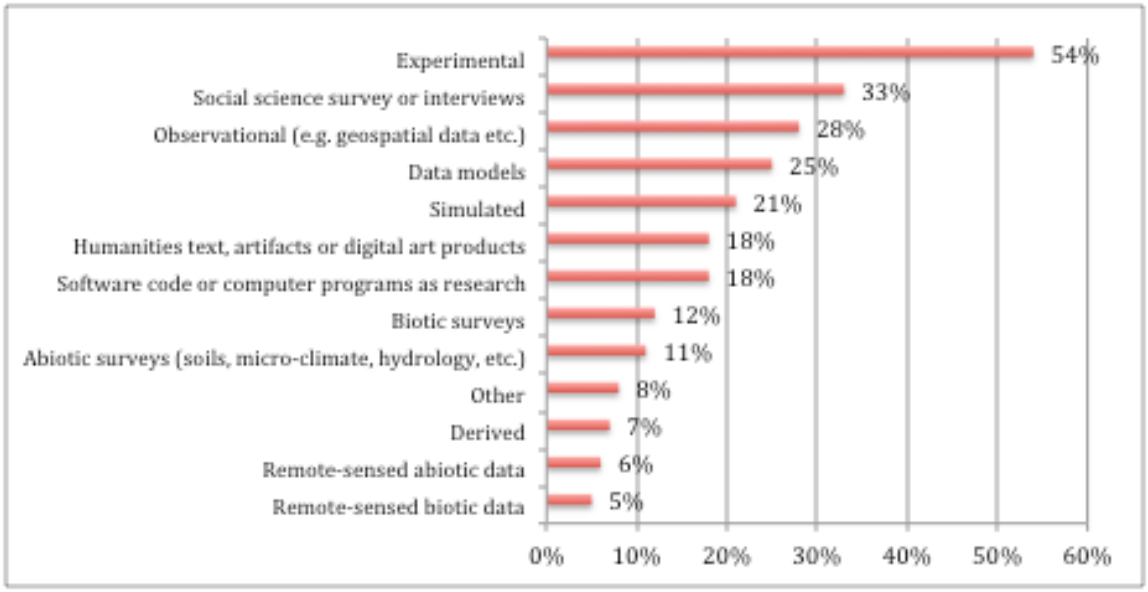
Other participants-specified needs include: education in "combining parts of different existing datasets from repositories" and "courses on mid- to long-term data storage, meta analysis, use of public data, etc."

Long-term data storage and archiving come as the top services needed. Next, support and services involving data preservation were highly required among the faculty researchers. These include preparing and archiving data for long-term preservation, technical support on format migration and long-term data integrity, as well as guidance on creating data and metadata documentations. The researchers have certainly shown interest in active data storage as well.

In brief, 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 could be emphasized.

The following charts and table provide more information.

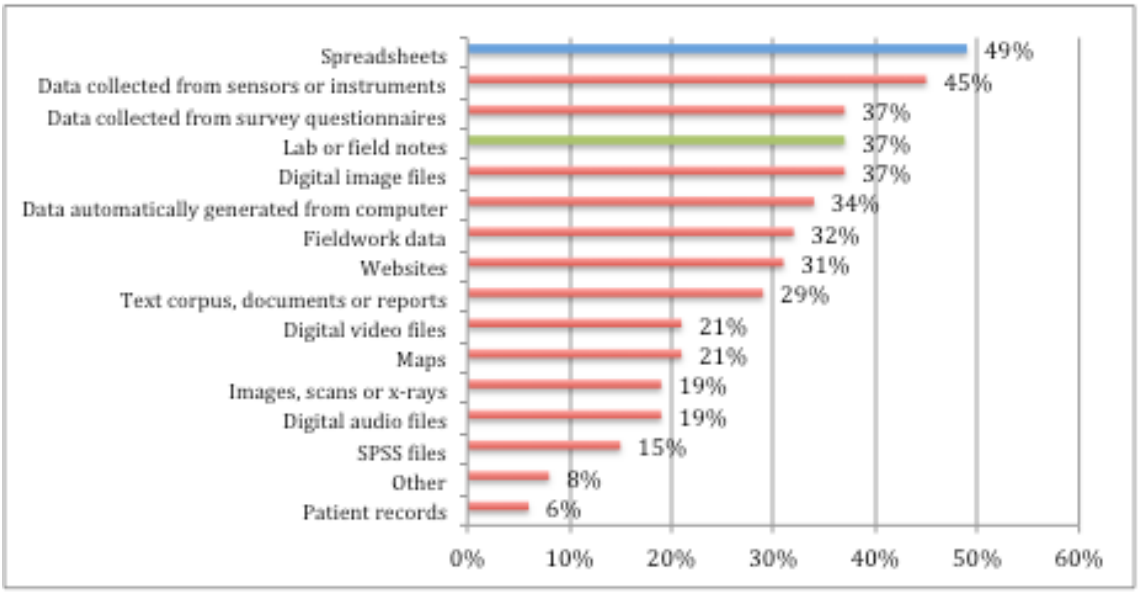
Chart 1. Types of data



Total Responses (n): 542

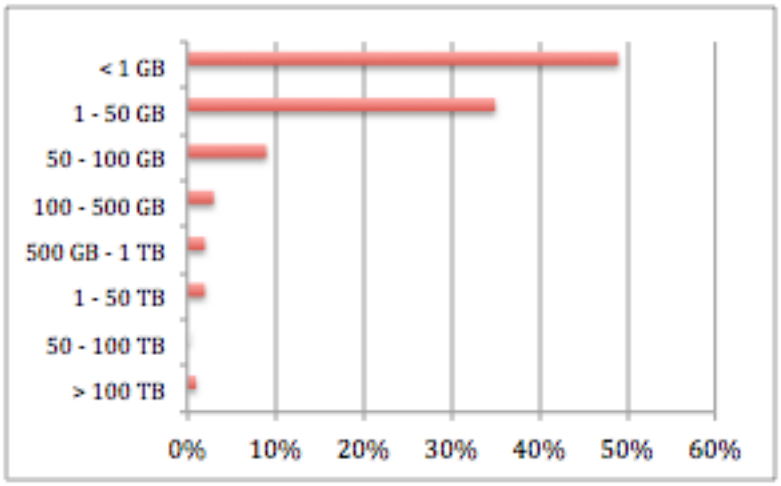
Chart 2.

Forms of Data



Total responses (n): 544

Chart 3. Typical data volume that faculty actively work with on a weekly basis



Total Responses (n): 431

Chart 4. Primary Roles Responsible for Data Management Plan and Implementation

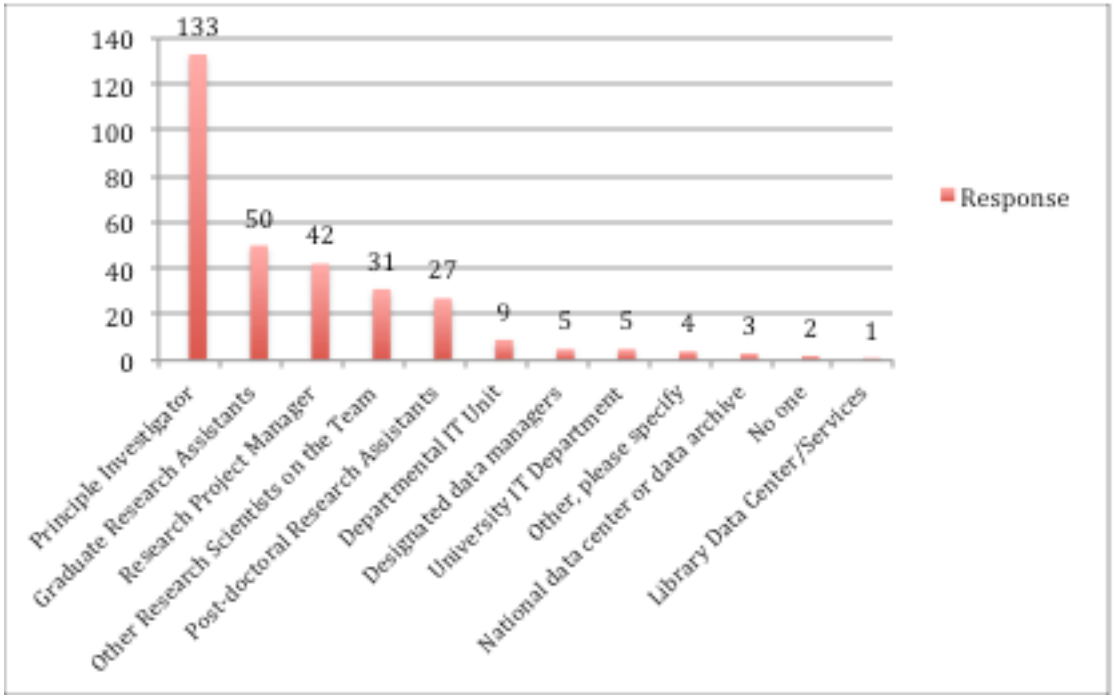
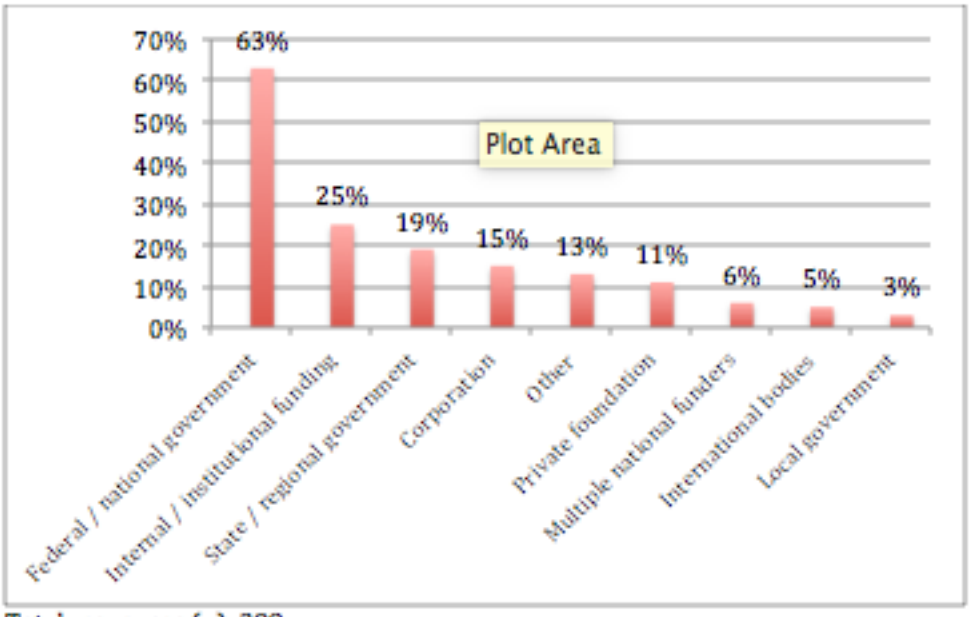
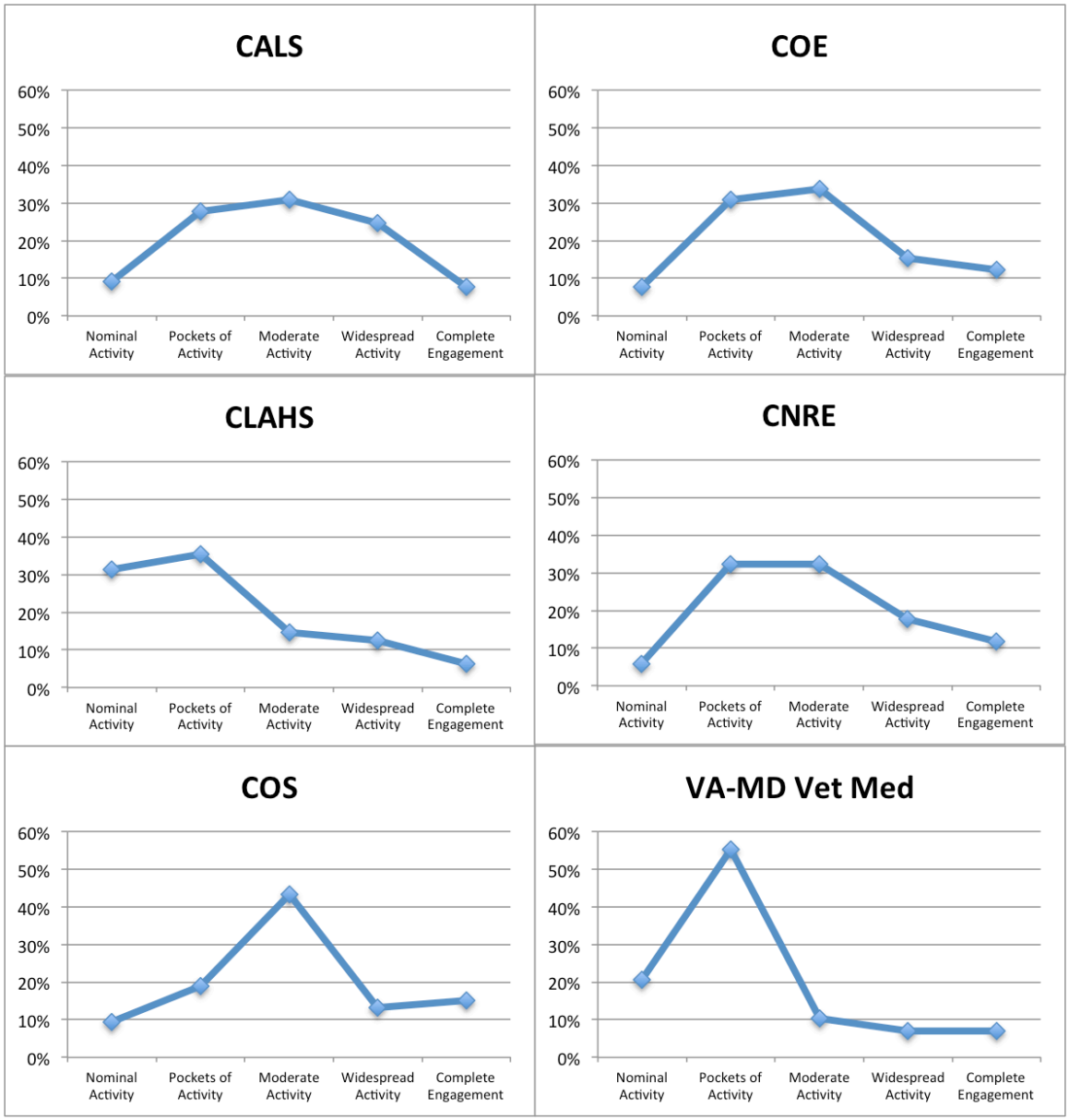


Chart 5. Geographic Scale of Funding for Research



Total responses (n): 388

Chart 6. College level engagement in openness of data



Appendix 1: Examples of data documentation methods used

Some standard metadata/documentation schemes used experimentally. Sporadic use. Please specify	Some published and recognized metadata/documentation schemes adopted. Please specify	Recognized discipline-specific metadata/documentation schemes widely used. Please specify	Established international metadata/documentation schemes routinely used. Please specify	Other, please specify
Picas	XML, SQL	Git	Metadata for RAW images	LitLink
Equipment proprietary formats	Asprs standards for some geospatial data	In compliance with university/journal policies for paleontological data	Proteomic and metabolomic conventions	All of the above, in different instances
Geospatial metadata following national standards	Standard terminology in the field	GFF etc. common file types for DNA/RNA sequence and annotation information	ASPRS LAS file definitions	Typically commercial packages (e.g. Compustat) have metadata docs which are mixed with other data I collect
We are working with CUAHSI to share much of our monitoring data, they have in house formats	Dublin Core	Description in experimental sections of papers	International Tree-Ring Database	Varies with project type and data type
http://dataprotocols.org/	netcdf	MLA Format		
VTTI [Virginia Tech Transportation Institute] established its own data storage standards and schemes for our unique data	Several data sources are used including public use survey data and medical claims data for which the format and contents are well documented	network analysis, ethnography		Not sure. Our data is uploaded to servers through an automatic process setup by our IT group, and we access the data through a proprietary program developed in-house.
NIST standards		MLA; APA		Humanities--use MLA format to document sources
netcdf, IDL save file		I have a wide variety of data from soil surveys to GIS data and generally follow the metadata and documentation schemes required for formatting the documents for publication and general organization.		Ad hoc. We document, but methods vary as appropriate to datasets.
FGDC		TEI XML		
PDF, LaTeX, JSON				
Depends on the project, many different forms				

Answer		Response	%
Comma-separated values (CSV) file (.csv)		224	47%
MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) or OpenDocument Spreadsheet (.ods)		225	47%
JPEG (.jpeg, .jpg)		210	44%
Widely-used proprietary formats, e.g. MS Word (.doc/.docx)		184	39%
Some widely-used proprietary formats, e.g. MS Word (.doc/.docx) or MS Excel (.xls/.xlsx)		172	36%
Adobe Portable Document Format (PDF/A, PDF) (.pdf)		168	35%
Plain text (.txt)		141	30%
Rich Text Format (.rtf), PDF/A or PDF (.pdf), HTML (.htm), OpenDocument Text (.odt)		137	29%
Tab-delimited file (.tab)		127	27%
Plain text data, ASCII (.txt)		116	24%
TIFF (other versions) (.tif, .tiff)		115	24%
Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information		106	22%
TIFF version 6 uncompressed (.tif)		104	22%
MPEG-4 (.mp4)		93	20%
Rich Text Format (.rtf)		85	18%
Photoshop files (.psd)		78	16%
Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta)		73	15%
ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn)		63	13%
Hypertext Mark-up Language (HTML) (.html)		56	12%
MPEG-1 Audio Layer 3 (.mp3)		59	12%
Waveform Audio Format (WAV) (.wav)		54	11%
Delimited text of given character set with SQL data definition statements where appropriate		49	10%
Geo-referenced TIFF (.tif, .tfw)		45	9%
Some proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti		43	9%
SPSS portable format (.por)		42	9%
MS Access (.mdb/.accdb)		40	8%
Standard applicable RAW image format (.raw)		31	7%
eXtensible Mark-up Language text according to an appropriate Document Type Definition or schema (.xml)		30	6%
XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0		30	6%
Adobe Illustrator (.ai), CAD data (.dxf or .svg)		23	5%
ESRI Geodatabase format (.mdb)		26	5%
Some structured text or mark-up file containing metadata information, e.g. DDI XML file		26	5%
Tabular GIS attribute data		24	5%
Audio Interchange File Format (AIFF) (.aif)		21	4%
CAD data (.dwg)		20	4%
Binary formats of GIS and CAD packages		13	3%
Keyhole Mark-up Language (KML) (.kml)		15	3%
Motion JPEG 2000 (.mj2)		11	2%
Free Lossless Audio Codec (FLAC) (.flac)		2	0%
MapInfo Interchange Format (.mif) for vector data		2	0%

Table 1 Format of Data (note the percentages are rounded numbers, n=475)