# Virginia Tech University Libraries' Data Service Pilot with the College of Natural Resources and Environment (CNRE)*

## Data Profile and Needs Assessment Project Report

**July 2015**

**Natsuko Nicholls, PhD (PI)**
**Andi Ogier (Co-PI)**
**Kyrille DeBose (Co-PI)**

* This study is IRB approved: *IRB # 15-325, Research Data Profile and Needs Assessment Study, approved March 2015.*

# Table of Contents

# Executive Summary

## Introduction

This summary will provide the background, objectives, activities, and outcomes of the CNRE data service pilot project. Although the project has different components (training, collaborative research support, and needs assessment interviews), this report will primarily highlight the research questions, methods, and findings from interviews conducted in April 2015. A set of recommendations will be developed as we expand our capacity for data services to support the diverse research environments at Virginia Tech.

## Project Description

We partnered with CNRE to explore the diverse range of research data generated in the College and assess data-related needs. We focused on three areas: 1) data management training through guest lecturing in CNRE undergraduate and graduate courses, 2) data profiling and needs assessment interviews, and 3) support of collaborative projects designed to strengthen partnerships and support curation infrastructures on campus (e.g. GeoBlacklight Project Team, Geospatial Metadata Working Group, Purchased Data Working Group). This report discusses our findings related to the data profiles and needs assessment (area 2 above). In support of this analysis, we conducted in-person interviews with faculty to gather more meaningful information than a general survey could provide. Our goal was to gain a better understanding of their research data, particularly the associated value of it, and data management activities surrounding it, with an eye toward gaining insights on how the Library can provide further support in this area.

## Project Outcome: Interview Results Summary

We recruited 15 CNRE faculty members for a 50-minute interview in order to better understand their research data, particularly the associated value of it, and data management activities surrounding it. We were ultimately able to schedule interviews with six faculty in three of the four CNRE departments: Forest Resources & Environmental Conservation (FREC), Geography (GEOG), and Sustainable Biomaterials (SBIO). We developed five questions around the following areas of interest: 1) data profiles, 2) data management workflows, 3) data challenges, 4) data value-add, and 5) data management planning. What we learned from interviews are:

1. There is a significant data diversity (in types, formats, size, software/tool, environments) within one single research project—accordingly, research management workflows and tools are needed at a project level, not institutional/department level.
2. Researchers are cognizant of different states of data (raw, analyzed, curated, finalized) and associated needs and challenges—accordingly, we should provide data stewardship to reflect the needs and requirements specific to different states of data even within the same type of data.

3. Researchers find data management time-consuming—accordingly, data stewards should focus on helping them develop high-quality workflows that increase efficiency while allowing for creativity and spontaneity.
4. Other data challenges are related to data ownership, version control, lack of (format and metadata) standards, fragmented workflows, complicated lifecycles, difficulty of tracking data within silos, storage space, data deluge—accordingly, for instance, new systems are needed to allow researchers to bring algorithms and processes to data, instead of the other way around.
5. Researchers commonly understand that good data documentation and use of metadata significantly helps increase data discovery and reusability, and that such practices can be more efficiently done with high-quality workflows and tools in the research process—accordingly, we should continue to assist researchers with data documentation best practices as well as explore opportunities to help them select a tool (e.g. electronic lab notebooks) that is better suited for their discipline-specific data annotation and documentation practices.
6. Researchers tend to interpret/define the value of data from data producers' perspectives, rather than considering historical, public interest, or interdisciplinary value; there exists a significant concern that important data will be misinterpreted, and this concern may outweigh the wider value of data sharing—accordingly, we should emphasize the importance of well documented data for data re-use, and attempt to harmonize the needs of data producers and data consumers.
7. A Data Management Plan (DMP) (due to 2-page limit, advance-planning, a gap between simplified documentation and actual data management practices over multiple-year project) is still considered by researchers as a hindrance—accordingly, we should provide additional training as to why data management plans are important to research, and practical examples of how data management can increase research effectiveness.

## Recommendations

Interview results from this study have helped us gain insights on how we should/could provide further support in this area. We recommend the University Libraries to:

1. Stay engaged with data conversations and further promote partnerships with VT researchers.
2. Discover similarities and dissimilarities across disciplines while not assuming heterogeneity in data management practices and needs at the departmental level.
    a. focus instead on data management and workflow analysis at the project level, paying special attention to interdisciplinary project teams.
3. Emphasize the importance of standardization in data management planning and practices while keeping in mind that some researchers have reservations for doing this as they think standardization may hinder innovation.
4. Provide information about a variety of tools that are available and let project teams make the choice for data management tools that best meet their needs.

5. Continue to offer Data Management Plan Consulting, but also pilot the creation of practical and immediately useful data management workflows for an interdisciplinary research team in order to demonstrate the benefits of data management planning.
6. Extend this project to the College of Agriculture and Life Sciences for extending partnerships and enabling a comparative analysis of the interview data from this study.

## Interview Questions

### Q1: Data Profiles

We are interested in learning about the data you create in the course of your research. Please tell us about the data that you work with on a regular basis, including how they are generated and analyzed. Do you use particular software or environments for analysis? Are there specific file types or formats that you use? Do you distinguish between raw, analyzed, curated, and finalized datasets in the course of your research?

### Q2: Data Workflows

We'd like to learn about how you store, access, archive, and publish (or cite) the data you work with. We are particularly interested about where these tasks occur within the the processes or methodologies that guide you through the project.

### Q3: Data Challenges

What data-related challenges have you experienced in the course of your research? How did you solve the problems? Who helped you solve the problems?

### Q4: Data Value-Add

As funding agencies begin to place increased emphasis on the value of data beyond the project that creates it, please think about the value your research data might have to other researchers. Are there other disciplines that might be interested in your research? How might other scholars use the data that you create?

### Q5: Data Management Planning

Have you submitted a grant application that required a Data Management or Data Sharing Plan? If so, how did you create it? What did it say? If not, do you have a record management policy or a data disposal strategy for your current research project?

# Quantifying Interview Data

## Table 1: Profiles

| | ***Type*** | ***Type*** | ***Type*** | ***Type*** | ***Size*** | ***Format*** | ***Software*** |
|---|---|---|---|---|---|---|---|
| **ID** | **observational** | **experimental** | **derived** | **remotely sensed** | **TB level?** | **# of formats** | **# of software used** |
| ID_02_06 | Yes | No | No | No | No | 3-4 | more than 5 |
| ID_02_07 | Yes | No | Yes | Yes | Yes | more than 5 | more than 5 |
| ID_02_08 | Yes | No | Yes | Yes | Yes | more than 5 | more than 5 |
| ID_03_09 | Yes | No | No | No | No | more than 5 | more than 5 |
| ID_03_12 | Yes | No | No | No | Yes | more than 5 | more than 5 |
| ID_04_14 | Yes | Yes | No | No | No | 3-4 | more than 5 |

## Table 2: Workflows

| | ***Storage*** | ***Storage*** | ***Access*** |
|---|---|---|---|
| **ID** | **# of storage options** | **storage location** | **metadata standards: awareness and practice** |
| ID_02_06 | 3-4 | local PC + optical media + VT server(s) + off-campus storage | aware |
| ID_02_07 | more than 5 | local PC + optical media + VT server(s) + off-campus storage | aware |
| ID_02_08 | more than 5 | local PC + optical media + VT server(s) + off-campus storage | aware |
| ID_03_09 | 3-4 | local PC + optical media + VT server(s) + off-campus storage | aware |
| ID_03_12 | more than 5 | local PC + optical media + VT server(s) + off-campus storage | aware<br>applied to practice |
| ID_04_14 | 3-4 | local PC + optical media + VT server(s) + off-campus storage | aware |

## Table 3: Challenges

| ID | # of data issues addressed as challenges | What are these challenges? |
|---|---|---|
| ID_02_06 | More than 5 | media; format; copyright of original data; ownership |
| ID_02_07 | More than 5 | formats; lack of standards; storage space; data ownership |
| ID_02_08 | More than 5 | space; formats; workflows; different lifecycles (e.g. student, grant, research, and data lifecycles) |
| ID_03_09 | 1-2 | fieldwork-related challenges |
| ID_03_12 | More than 5 | space; formats; workflows; sustainability;different lifecycles (e.g. student, grant, research, and data lifecycles) |
| ID_04_14 | 1-2 | cultural resistance towards data sharing |

## Table 4: Value-Add

| ID | awareness | applied to practice |
|---|---|---|
| ID_02_06 | Yes | Yes |
| ID_02_07 | Yes | No |
| ID_02_08 | Yes | No |
| ID_03_09 | Yes | Yes |
| ID_03_12 | Yes | Yes |
| ID_04_14 | Yes | No |

## Table 5: Data Management Plan

| ID | Opinion about DMP mandates |
|---|---|
| ID_02_06 | ambivalent |
| ID_02_07 | ambivalent |
| ID_02_08 | ambivalent |
| ID_03_09 | positive |
| ID_03_12 | positive |
| ID_04_14 | negative |

Table 6: Library’s Role and Involvement in RDM

| ID | Opinion about Library’s role in RDM | Having Concrete Ideas about What the Library Should/Could Do |
|---|---|---|
| ID_02_06 | positive | Yes |
| ID_02_07 | ambivalent | No |
| ID_02_08 | positive | Yes |
| ID_03_09 | ambivalent | Yes |
| ID_03_12 | positive | Yes |
| ID_04_14 | ambivalent | No |

# Full Interview Notes

## Interview Subject ID_02_06

**Overview of his major project:** Treedata - the legacy tree data that involves detailed measurements of individual trees, US and North America, Forest Service sponsored project.

- **The value of (legacy) data (Historical value)**: Tree measurements has been done over 100 years. The tree measurement data from the past for different research questions, different value out of tree measurement data - e.g. When people cut trees for timber, they take detailed measurements; the measurements (from decades) are still valuable today, shapes of trees how it relates to the carbon that is stored in the forests; e.g. Wood pellets are shipped to Europe for heating — they like to know how much bark is in the pellets because it affects the ash.  All these measurements can help researchers answer new questions — it's expensive and hard to collect the data; they would prefer to use the data that was already collected.
- **Data formats:** paper (OCR), digital, media (floppy, magnetic tape, modern electronic formats) — hard to get into a format that they can use. Media issues, transcription issues for handwritten records. Small army of students that are entering data. Much more cost effective to enter these records than to send people out in the field. They have records from all over the country.
- **Types of data:**
  - Complicated datasets: different studies have different measurements, they have a database with different tables (location, individual trees, stem data, branch data, specimens/subsamples/discs) —> all useful measurements. Complex data structure, 11 different tables linked via primary and secondary keys. Images? Digitized paper records are linked to the measurements as well.
  - There is data on the paper records that they can't use or don't have time to work with — they preserve them in case someone from the future can use it. They always link a copy of the raw data file to the database.  They also track any reports, technical reports, journal articles — having those reports linked to the datasets is difficult b/c of paywalls.
  - Includes images
- **Challenges and Opportunities** - Researcher has solved his challenges himself, sort of...not all researchers can do or cannot have resources to do so.

- **Challenges with copyrighted materials**: linking to journal articles — having those reports linked to the datasets is difficult b/c of paywalls.
- Some are government documents which are public domain. The problem is copyrighted material that is hard to track down — they get it through ILL but can't make it available.
- **Dynamic data:** Treedata is dynamic, it's growing, people are adding data to it.
- **Campus Partnership:** Working with CGIT and CMI for technology. CGIT is helping with database structure and modern database architecture and build utilities, including adding new data to the database. CMI is working on a web portal for the database, simple queries. [CMI— Researcher contacted Lola at CMI, specialist in websites that use geospatial tools. Lola contacted CGIT, Brian. Tabular data type problem. Not really a geospatial problem, more of a database and data management problem — more of a programming and web design problem];
- **Future plan**: Soon, they want a DOI assigned to datasets, and versioning as the database changes (by updating DOI for versioning) - Library!
- **Benefit of data sharing**: he is aware of the literature about the positive impact of data sharing — people who share their data are cited more frequently. Data sharing enables collaboration. Treedata is going to be very unique — there is another dataset that is global in scope but with a more limited perspective. They created it as a separate publication.
- **Challenges with storage:** Researcher doesn't take personal possession of the original records (that go back to Forest Services) — he doesn't want to deal with long-term curation of paper records. So he scans/digitize, the originals are scanned/duplicated for future use, or as a backup. There are some that have historical value (1902. that map to published documents from the 1920s. The individual value them enough to keep them for many years — if Researcher could find a stable home the keepers would be willing to donate them.

**Other research project:**

- **Profile:** Forest Inventory and Analysis (FIA) is the funder; Researcher works with large datasets that the forest service puts together. He puts other tree data that are not going into TreeData, but related; they make a large amount of their data available to the public. FIA data mart, download tabular data by state. text-based tabular data state by state, also possibly access database. VERY important, land-based, forest-based measurements across the united states. Conditions of forest — mandated by congress in the 1930's recognized that timber resources are important.  Treedata has more details on the trees but less detail on the landscape. The data is disappearing.

  - **Challenges associated with data re-use**: This FIA database is complicated, thus users has a steep learning curve. A few in CNRE have spent a long time working with the data, a lot of knowledge, they can help students and research associates get up to speed, but still, "it is easy to mis-analyze these datasets if you don't understand the way the database is structured. You have to filter them in a certain way to get the collection of data that makes sense (29:45)."
  - **Challenges with data disclosure and data privacy:** Congress passed a law in the 80's that the forest service (FIA) could not release the exact locations of the plants on the ground—to protect corporations and private landowners.  They have to hide the true locations of where the plants are located. Privacy issues. With Treedata, they have to hide the location — corporations will give Researcher the data as long as they aren't identified. Privacy of Location (it is part of TreeData!). 'Sensitive' tree data — make the location data private in the database (but majority of Researcher's data is not confidential). Location of Researcher's TreeData: The website is not a forest service website, it is a VT website — Forest Service has high hurdles to meet for putting something on the internet on a government website.  Researcher works with the IT people in CNRE — the hurdles are more reasonable. Keep software updated, keep personally identifiable information private. (to avoid any liability issue as well)
  - **Public Interest in the data (= value of data)**: Outreach - Researcher was in Idaho last week — mid-level executives can use the tree data into their enterprise (potlatch corporation, boise cascade corporation). Researcher can calculate the value in terms of what it would cost to collect the data. There's other value to consulting forestry companies — Not directly related to the taxpayers, but the corporations or consultants that employ the taxpayers.

- **Data Management Planning/Policy**: NSF DMPs are too rigid, kind of limited — they know that a lot of detailed measurements were made and summarized, but it's the individual measurements that are important. "Different researchers are interested in different parts, different details in the data (48:35)." They distill the data into a few coarse records and publish those. There is a lot of data loss. "What we are striving for is to preserve more information (49:19)." Metadata is vitally important. Students all the way up through Bureaucrats — at either end of the spectrum people don't know — it's the researchers and experts in the middle who really understand it. “I wouldn’t say data management plans are hindrance, but what they are is NSF DMPs are too rigid, kind of limited (47:38)...Standardization is not always good because different people measure different things (50:25)..it seems to me this standard should have something that allows you find your data and tell us what it is.”

## Interview Subject ID_02_07

- **Types of data:**
  - Gather and heavily use remote sensing data
  - Data collection methods: Mainly airborne (LIDAR hyperspectral & gathered by plane fly-overs) and satellite (LandSat); LandSat release in 2008 has caused inundation of data, volume; becoming the hardest thing to manage (thus, a major challenge).
  - Future: Landsat 10 is going to be hyper spectral — data volume will expand even further.
  - Change in data collection method brought a paradigm shift with this explosion of satellite data availability - research has completely changed.
  - Landsat is a raster image, LIDAR is point data, a coordinate cloud, attributes for every point.
- **What changed:** Creating new algorithms to handle bigger sets of data; before could only run an algorithm over hand-selected images, now it’s over large sets.
- **Data standards:**
  - The satellite data is standardized (same specifications), which is a plus, but becoming an issue to provide intermediate products, where the algorithms can be used over time to form processed data sets.
  - Airborne data, i.e. LIDAR and hyperspectral data, have no standards, so while not high in terms of volume, are expensive to gather; They are useful for filling in gaps for time series data, where cloud cover might have made changes less apparent to see, algorithms designed to address those issues as well.
  - There are 21 specific formats that data could be stored in.
  - The community is working towards standardization for LIDAR and hyperspectral data although vendors (such as ESRI) are developing their own proprietary data.
  - The data are large and not the same type, so data becomes a challenge, but they want to be able to pull out ALL the data for a particular location (polygon).
- **Data storage:**
  - Raw data are archived elsewhere — not a big push to archive.

  - Data are usually not stored in the same place. CNRE’s departmental server Minnow & CDs (raw, GB’s not TBs. Airborne data — once it’s collected and processed, put it on the library.they want to move it to the library); some of it has been posted elsewhere because of DMPs.
  - VGIN (VITA, state agency) is coordinating the collection effort, right now data storage is hodgepodge.
  - Data deluge: With Satellite imagery the issue is volume (Google also collects airborne data, but they only store the algorithm, they recreate each time from the raw data).
- **Data services needs during the workflow; potential collaboration with the Library**
  - Different kinds of digital objects to be published — algorithms, finalized datasets for laypeople, parameters for algorithms that fill in the gaps; the library published the algorithm with a DOI for a CNRE researcher.
  - To fill the gap, what is the gap? There are times where clouds will impact a location, the location is missing, evan has been writing algorithms that could fill in the gap and give a continuous curve over time. There are parameters that describe those gaps—could be published.  The parameter set would make the data more usable for everyone.
  - Looking at NEON (National Ecological Observatory Network) to create a data repository in the future, also looking at how have structured own servers to migrate data to the library (for permanent archive) once have a collection that can be utilized (would like a dynamic space, but will most likely be static – find set, download, work on own machine).
  - GeoBacklight (ongoing project between CNRE/EGIS people with the Library) to run instance and see if it can help with indexing data in Minnow
  - Work with 2 students (one incoming in fall, the other midstream in data collection/analysis) and look to train on best practices for data management/creating metadata schemas from the beginning and one pathway into a project.
- **Data Challenges:**
  - Data deluge: With Satellite imagery the issue is volume (Google also collects airborne data, but they only store the algorithm, they recreate each time from the raw data).
  - Data collection methods: Mainly airborne (LIDAR hyperspectral & gathered by plane fly-overs) and satellite (LandSat); LandSat release in 2008 has caused inundation of data, volume; becoming the hardest thing to manage (thus, a major challenge).
  - Lack of standard for LIDAR and hyperspectral data
- **Data Value-Add:**
  - Discipline allows for easy collaboration/sharing of data for research, but not necessarily of raw data to be shared if you’re not the owner.

    - There is a growing push to share land cover change with the public, but still lots of concerns and questions about data ownership. Analysis could be shared, but hardest part there again is with lack of standards, organizing it and making it accessible in an understandable format is a challenge.
- **Data Management Planning:**
    - **Awareness and experience**: Yes – have for NASA/NSF, and includes short section for USDA grants (not yet required). If raw data is proprietary, will share derived data instead (a lot of what they use is hosted by USGS, refer back to that set).
    - **DMP-related issues**: For creating plans, use a template got from a training some time ago and update with pertinent information for current grant proposal.

## Interview Subject ID_02_08

- **Specifics/Types of data**
    - Remote sensing and ground-based data collection
    - Data collection instruments: ground, satellite, and aircraft based sensors (those data types are similar)
    - Rather different type: Soil data, trying to understand what would allow scaling of point-based data in space in time; Point-based data writ large.
- **Formats of data**
    - "Data formats, this is a nightmare for us" — e.g. in class, we juggled between 2 or 3 different formats to open in one software.
    - The Hadoop people have been recommending ASCII; they are tripling or quadrupling the data storage (one integer = two ascii bytes). They are thinking about Migration issues — in curation mode, how best to store. Binary flat file ASCII is quite good (file with a header metadata). Landsat Science team is migrating to jpeg2000. With LIDAR came as ASCII mass point file;
    - new format called las binary file format put together by professional society, companies are making their own "flavors" —> moving away from open source, they have to be able to open all.
    - Question is data migration 20 years from today - how best to store these data in terms of curation
    - Different communities are moving into different directions
- **Software used when working with data:**
    - **GIS-side:** two major ones - 1) ESRI suite; 2) QGIS (slow, gedol).

    - **Remote sensing-side**: 1) ArcMap, 2) ENVI, and, 3) ERDOS Imagine; They buy big licenses with lots of seats for teaching and research; Imagine has a valuable data type conversion suite (100+ data types supported for input and output);
    - **Scripting-side**: They write a lot of their own scripts, 1) R, 2) Python, 3) C/C++, 4) Fortran, 5) Extant Libraries, 6) Lastools, 7) Matlab
- **Identification/distinction of stages/status of data (raw, analyzed/processed, curated, finalized, etc.):**
    - “We try things that don’t work — generates a lot of data that they don’t actually need.”
    - **Pre-processing stage** — “the question is whether to retain the original data or not. The question is about whether the data is retained elsewhere (14:02).” “relying upon the fact that we can always go back to the original data, which is not quite true. This is a problematic because the data volumes are unreasonable.” e.g. LandSat refines the production algorithms as they find something new — every time they download data, it is different. They can’t always go back and get the original data. The data volumes are unreasonable.
    - Preprocessed —> Analysis — not final, but worth keeping it to be able to reproduce process. They would really like to the point where replication could be completed. “Data is somewhere, bring algorithms (process) to data” “This is where we want to go.” “that’s the direction many researchers are moving to and this is an elegant solution.”
    - e.g. Google Earth Engine — Hadoop Writ Large— Google keeps it in storage (storage is cheap) it’s easier for them to recreate a dataset than it is to create and store it. “They ONLY store the raw data, they are assuming that you will bring the process to the data”
- **Challenge particular to his project:** compute space and storage. Hadoop model (the interviewee uses campus one) has led them to develop their own data structures - traditional data structure no longer works!
- **Other challenges associated with storing, tracking, organizing**
    - “Keeping track of own data frankly is important and a challenge.”
    - Two decades worth of stuff
    - “New faculty and students come and go - It’s a hodgepodge mess.”
    - They need a way to catalog and organize the data they have. There are interesting research projects, but he can’t find the data. Tracking, organizing is a problem. It’s expensive. “Need to invest in a system for ongoing and legacy data.”
- **Electronic Lab/Research Notebook** - they have been reticent because their stuff is a little different from the “normal” lab notebook — the workflow has a maintenance element anyway via versioning (of scripts). “They need some additional movement but my caution is two-fold: 1) we want to be sure the caliber and quality is high without stifling the creativity of the process - if you force them into a workflow, they lose edge .2) They would like to be able to know if someone is making stuff up. They encourage students to keep files in their main spaces, but they don’t know if things have been massaged.” “We

want to value the creativity, spontaneity, group innovation." They've done Microsoft Project.

- **Tools:** Online Plotting/graphing Services: Plotly — directly tie into python; publishing data and plots at the same time. Linking every figure in a paper back to a dynamic data + visualization.

## Interview Subject ID_03_09

- **Types of research data:**
  - Mix of qualitative and quantitative data
  - mid-size
  - primarily household surveys from Africa/Tanzania (100+ datapoints, survey to a couple of hundreds of households)
  - longitudinal study, taking place over 10 years (he's been on for the past 5 since grad school) so far.
  - uses some geospatial data (GPS coordinates/points); but not in a great deal; more in terms of information about locations and infrastructures (e.g. water development, schools, churches, roads); his new project looking at tracking cell phone coverage, needs GPS datapoints to create mapping of area.
  - Also has pictures and videos as raw data, but only uses these in classes, not publications.
- **Data collection/organization and analysis; formats**
  - fieldwork in Africa; interviews and surveys; via enumerators and translators
  - In study design, first spends weeks using an interview template; Then moves the survey into field testing, training of survey administrators, then into the field to gather data. Training is an issue – some don't fill all of the questions, others have asked questions to groups rather than individuals which skews the data.
  - Challenges in data collection: logistic issues (e.g. a enumerator getting sick), missing data, etc. (-->this applies to Q3: Data Challenges)
  - **Excel** for initial data entry and **Stata** for most of the data analysis
  - **dedoose** - his students use dedoose for qualitative data analysis
  - **ATLAS.ti** for analyzing/coding and qualitative analysis
  - **Microsoft Office products**
  - **ArcMap** if using georeferenced points (some special measures)
- **Data storage; formats**
  - Hard-copies in his office lockers - an original copy + a few (hard) copies
  - no raw data in recording (due to cultural reasons; no pic-taking)
  - All original hard copies scanned (multiple digital copies stored)
- **(Qualitative) Data format (with 'sharing' in mind)**

    - Survey-based qualitative data - how did that work? question about how to share qualitative data
    - What are the quotes related to research questions?
- **Data Workflows:**
- **Basic/general workflow regarding data collection, process, storage, access, etc.**
    - Collection, data entry, then analysis are the most important aspects.
    - Some unique characteristic of his research due to the field work in xxxx: cultural aspects of group he's studying very suspicious of technology, combined with remoteness (and lack of infrastructure) leads to most data being collected in paper format,
    - Print surveys are kept in file cabinets in locked office
    - Then, scanned and stored in **DropBox** (most reliable way to access files, particularly when overseas).
    - it's always the question of 'data safety vs. access convenience'
- **Working with sensitive data (storing sensitive data)**
    - Yes; IRB on all surveys.
    - Print surveys are kept in file cabinets in locked office
    - The interviewee's research does not ask 'high risk' questions (e.g. health) but rather more about household purchases and uses of items; 'opinions' at most.
    - May have some policy tie-ins (e.g. infrastructure), but for the most part survey questions have low benefit to participants.
    - Many people he interviews are illiterate, but "that's not the justification for not taking care of your data (20:10)."
- **Data Challenges:**
- **Challenges in general:** For the interviewee, "data challenges are not about data storage or use; rather, it's about data collection (22:59)"
- **Challenges associated with data collection** (given the challenging infrastructure)
    - Paper and pencils vs. Tablet data collection; The interviewee prefers a paper-n-pencil method; the issue of power (charging iPad, tablets, etc); data collection vs. issue of data collection speed (it gets to more of administration!)
    - Enumerators that conduct the surveys vary – some very good, others very poor even after extensive training.  Poor infrastructure (getting from village to village takes a lot of time) and cultural aspects (e.g. punctuality).
    - Getting the surveys to be filled out consistently is a huge challenge – has lost data points because of how little (or incorrectly) the survey has been filled out. Going back to ask people to repeat a survey is very difficult.
    - Translations between xxxx and xxxx and xxxxx (using the method of back-translation): they have a very good process, but the intentions of what the questions are asking may not be understood by the enumerators (may explain why not all surveys get filled out, or why some are answered in a way that makes it obvious the question was not interpreted correctly).

    - Participation and survey fatigue are always challenges.
    - Incentives: Can’t give money (no compensation), so offer small commodities/gifts of tea and sugar (goal is to benefit the family as a whole, not just one family member).
    - Repetition issue: having to take the same survey again because of poor data collection the first time; some understand there is not an immediate benefit to answering the questions, others want to see change happening quickly based on some of their responses (e.g. new school). Building rapport with the survey participants is also a challenge – they answer the questions, but wonder what’s in it for them. Have given/fundraised some small amounts of money to help with community projects (e.g. new room in health center to have dividers) but the projects come only after community as a whole has agreed what will be worked on.
- **Data Value-Add:** As funding agencies begin to place increased emphasis on the value of data beyond the project that creates it, please think about the value your research data might have to other researchers. Are there other disciplines that might be interested in your research? How might other scholars use the data that you create?
- **Who:**
    - Anthropologists and other human geographers would get the most from this data.
    - Santa Fe Institute has researchers looking at origins of inequality; this study could lend insights (along with many others).
    - Studying a particular ethnic group, but can be a piece of other studies that looks at human behavior more broadly.
    - Currently there is a model being built (check voice transcript for name) that will allow for both current and new variables to be added and analyzed.
- **Other comments:**
    - #1 priority is to have his questions addressed, the use by others is secondary.
- **Data Management Planning:**
- **Experience with DMP writing**
    - Yes: Three grant proposal are under review; the interviewee was responsible for writing a DMP for the one (study about Masai) of these three grant proposals; For his other projects, he is a co-PI on two projects, the PI’s have put the DMP together; the other is part of a multi-institutional proposal, he’s just looking at the social science aspect; the PI is from Colorado (who constructed the DMP) and studying soils and climate change) - PI at other university will be responsible for data management
    - DMP content
        - data retention, in his DMP he noted data retention period as 10 years
        - data sharing: both quantitative and qualitative data will be shared
        - VTechWorks as data deposit destination

    - He used the DMPTool and looked at examples from others to construct his, had also talked with Andi when putting this one together (I remembered looking at it as well).
    - He also has another project, examining responses to his “pink slip” project (classroom-based activity) but no DMP for that project at this time, just a lot of data entry at this point.
- **Quotes:**
  “I feel like I probably am a good person to talk to, as I have the data that are different from a lot of people in CNRE -- Mix of qualitative and quantitative data collections, mid-size datasets” (2:04)
  “I noted in a DMP to make my qualitative data publicly available through Virginia Tech Library VTechWorks”
  “I don’t know it’s novel that qualitative data will be made publicly available… how would that work (to share raw data, i.e. interview transcripts, as in many cases, there are incomplete sentences and fragments..unclear in some cases (to other researchers).”

## Interview Subject ID_03_12

**Workflow/infrastructure building process to date:**

- Lessons learned / experience gained: from 17 yrs of research and teaching: he learned a lot about lifecycles of students and research projects within the university, and the typical problem is each student and each project too often has too narrow purpose for data - though we all collect/create, curate data, but once the project is over, data is shelved; thus data is not reused though it would have benefits to many other communities.
- Problem identified: This kind of fragmented workflow is particularly problematic given his research data characteristics, i.e. geospatial data and spatial data infrastructure - given that the purpose is to build hierarchical scales of nested geography that relate to each other (e.g. campus data, Blacksburg’s geospatial data, and then county’s, state’s, federal/national, and international)
- Issue addressed: for CGIT, to address these issues is important and over the last 3 years they are conscientiously building workflows to support computational infrastructure — it adds lots of overhead to each project to do it right; it takes longer in the beginning; it would be easier if we continue to treat each project as a ‘blackbox’.
- Curve: Once it has the infrastructure (once it gets rolling), it work, it makes things easier in a long-run; “there is a curve that we have to get over (4:48-)”
- Cost of data management practice in terms of time;

  - Before: estimated 10-15% of their time and effort
  - Today: estimated 20-30% of their time and effort
- Benefit of having built workflows: It is more integrated; the right infrastructure to collect and support research

**Today's system:**

- Now he has a whole system—including source data, separation of concerns re: creating scripts for different pieces such as downloading, parsing, qa/qc, translation, new format (pre-processing), and the actual model that they run on data
- Hardest part: training researchers and students to follow this and enforcing them; addressing and training versioning practices (recent focus today via brown bag seminar topics, etc)
- Built a web of infrastructure between VT storage (Minnow w/ CNRE) and Enterprise GIS for production GIS (many users, no control over space allocated to us/other departments)  (only 5G left out of 2T departmental space amongst 20 departments - there was no good plan to manage amongst departments when it started.

**Status of data and location/resources**
**VT vs. Non-VT resources:**

- **NAS Minnow (VT)** is used for intermediate and project work files (processing in Minnow)
- **Enterprise GIS System (VT)** is only for production-level GIS services (e.g. project with the outcome like web portals) - once it's ready to go, copy them over and "publish"
- Thus, testing and production lines in parallel.
- Non-VT resources like **Rackspace/Amazon/DigitalOcean** (SSD cloud server) because of limitations on Enterprise services — unlike average consumers (e.g. teaching purposes), usual researchers who need root-level access to machines to install custom libraries and software
- **Hosting@VT** (VT)for small projects
- He and his team built **computational Infrastructure plan** that specifies all the storage types
- Their own **project website** (php)
- **University CMS** for public webpage
- **Blue Ridge HPC** — 30TB — they need to figure out how to move it.
- Influencing factors on what to use for what: "a lot of it is driven by technical considerations - what we can do, what kind of permissions do we have, bandwidth, fee structure) (10:50)"

- Many are facing a similar issue: Enterprise-level something vs. Cloud (cost comparison; capacity; technical considerations):

**Challenges in Access Control:**

- College-level access control: challenge is enforcing as bringing on new students—College IT manages permissions (manual/paperwork process to have a new student added) - "less control but getting done correctly, though - this challenge is limited to 'internal' systems and for external systems, they have control."

**Project Management System/Tool:**

- **Redmine (built on Rackspace)**: not only data tasks but all sorts of project management tasks (the system that enables team members to issue tasks via CAS); e.g. 20+ team members on his Eastern Vineyards project; tracking different branches of code (when and what is shared/updated/deleted as to code) - "this is how people manage code now (17:35)"
- **GitHub** to store their code—they pay for private repository; separate buckets organized by:
    - Source Data vs. Derived Data (each bucket to separate these two)
    - Organization/Separation of Concerns:
        - Things with general purposes
        - Downloaders, Converters, Processors
        - Sand-alone things
    - Interesting Distinction between data concerns and data process concerns (but they are all integrated! - Andi) - "we train people, and this is how you should do it. And when they don't do it, we point to the documentation (19:42)."
- **GitHub** also holds models and documentation via wiki:
    - Using wiki within management tool (redmine/GitHub) - write formulas, how we build models, writing definitions, writing equations, etc.
    - Also to track what is located where, it is also documented!

**The ending of the workflow (What happens at the end?)**

- Retention: if continued funding, they can sustain—if not, they have to wind it down.
- Sustainability concerns with regard to hosting, storage and maintenance cost— web services that depend on the tools; browser technologies change — developer time is needed to maintain; maintenance needs contingent upon funding; otherwise, need to zip it up, put it on archival file on NAS store, and that's it. "I address the question to all of our stakeholders on board of sustainability of awesome things that we just built - they can be used for many years, but we realize that we need to keep investing in maintenance, at

least hosting cost. One way to look at program (project) management is sort of communicating with stakeholders long-term plans (24:45)."

**Values of what they build:**

- This is one way to justify we keep investing - to keep the value of tools, e.g. Web Portals that many users use!
- Data themselves: keeping some legacy projects
- Scripts-built for download/update/refresh (to maintain the currency of data) is important - those from 3 years ago no longer works (e.g. anything to do with GoogleMap v2. no longer works)
- Metadata (they also write metadata for products they create): Standard ISO, FGDC (Federal Geographic Data Committee standard)
- Re. data value: "Sometimes they don't know the value of data (30:51)" -- e.g. Eastern US project -- before overserved and done by each state, each state has its own protocol; Another example is 3DBlacksburg (created in 2008) to solve the problem of all sorts of different standards with different users and producers for city model and campus model. Different people created things in different ways and stored in different standards - we saw that as a problem. Though everyone has different use cases, in the end, there is no meaning collections or standards. How do we harmonize all these users and producers (32:43)."

**Other comments:**

- Metadata: GeoNode
- LIDAR (large laser scanning data, 19TB raw) processing into surface models — How can a Hadoop cluster speed up access time?

**The potential collaboration area with the Library:**

- DOIs for data sources
- Publish data before publishing a whole research article
- Citation for data
- Storage — need PBs of storage
- GeoSpatial Help Desk — figuring out which questions Ed can answer, which questions need to go to CGIT (Ed knows this)
- Data Management Plan — DMP examples

**Sponsors:**

USDA (Vineyards)

CIT--State Budget

VA911 Services through Montgomery County + Maintenance Plan (5 years)

## Interview Subject ID_04_14

- **Data Characteristics**
    - Gathers a lot of observational and instrumental data, project design and parameters are very important.
    - Data are complicated, but not large amounts --- biggest issue is connecting data with experimental parameters and good notes on how it was gathered.

- **Data Workflows (and challenges researchers likely encounter in workflows)**
  ***(Note: Data Challenges are noted when discussing workflows)***
    - Record Keeping/Data Documentation
        - Making sure data is recorded, recorded properly, witnessed properly, tying digital to physical recordings properly (sample numbers relate to sample wherever it goes throughout testing/analysis process)
    - Store/Archive
        - Currently relying on print lab books, printouts, emails from internal and external analysis groups.
        - Started crossing the digital/physical divide, but it is still a challenge
        - Notebooks/printouts vs Electronically-generated and saved data
        - Higher fidelity when they do it manually on paper.
        - Software may go out of date making data inaccessible (print while cumbersome has been highly reliable for later retrieval of information)
        - Carrying sample numbers alongside the data for tracking. It would be great for archiving.
        - Electronic lab books (ELNs) would be a huge help:
            - streamline the process of input by the student with instrumental analyses/samples prepped and observations.
            - Not sure if e-lab books could be provided through the library, but this is an area where the research system could use improvement
            - Maybe pilot the use of e-lab books with groups on campus, share results and best practices (some systems better applications in one discipline over another).
    - Student lifecycle
        - Students come and go;  keeping track of students (has 10 now, prefers 6-7 as time demands make it more difficult to ensure everything is being

captured properly for data recordings); managing data after students have gone is an issue.

- Training
    - Training students on record keeping and data documentation: While teaches how to maintain a lab notebook properly, there is a lot of variation based on students' abilities. Would like a less random system, but go with what funding allows for. Telling students what to do/how to do it and following through to ensure it's being done has high variability… some students are great and get it right away, others less organized and takes more training.
- Data Sharing
    - Cultural practice and resistance to data sharing; It would be rare for them to share data outside of their lab — it's a question of trust, interpretation of quality; You have to have control over who sees it.
    - Timing of data sharing: Sharing during the project and after project is complete
- Data Publishing
    - Disciplinary gap: He edits two journals, doesn't hear anything about publishing data outside the journal article. He thinks this is because he is Chemistry, not Biology; His data is harder to manipulate. Biological data is more complex because it's based on living systems. Harder science so data is stronger, more repeatable, less opportunity to fake data.
- Data analytics:
    - They would like to be able to link experimental parameters input with the student's output from instruments and analytical data after the fact. They are using a paper system.
    - Bringing tools to data: Data are no huge volumes, though it is complicated and needs to be connected to the experimental parameters.
- Other issues:
    - Problem with workflows in general: the broken part in workflow is communication and collaboration, not necessarily about data or data management.
    - Staying current: Keeping abreast of what systems work best for their area of research (e.g. e-lab books, bib managers, research databases, other digital tools) is a concern.

- **Data Value-Add**
    - Data value reflected into student's learning outcome: Most students go into industry, so wants them to have understanding of why data is important (patent info key) and give them a competitive edge over others; There is a trust issue with how others would understand how to use the raw data and use it

  - appropriately (takes a bit of knowledge in the field, general public would not have an idea how to interpret it).
  - Data value potentially increased by more tools: The more he has tools that will help him hold the students to standards, the easier things will be.
  - Data Value-add through more training: Students need to know about patenting, maintaining information, public access; he relies on written and verbal communication with students on these matters; some sustainable training material would be efficient.
  - Other: Use of analysis tools and providing 3-4 spectra to indicate how the polymer was created is critical, this data is included in articles (schemas/figures/tables/footnotes) and all articles have chemical drawings (uses an old program, but has been tried and true over the years).

- **Data Management Planning**
  - His take is that it has 0% influence on whether or not a grant is funded, and that's appropriate.
  - Data points in this field are "hard points" and he doesn't see the need for data management plans as part of the grant application. Chemistry is straightforward, analytical tools provide data points that are very difficult to manipulate, and already have means in the discipline to share the types of data that would be of use to others (e.g. how polymer was created).
  - Notebooks are broken. Should be fixed. Need to be archival.
  - Managing data after grad students are gone — need to make it easy to archive and share data within the group, potentially ELNs might be solutions.