**Discussion of EDAPT limitations for the DSP Hub Project**

*Bottom-line is that the DSP Hub is an advanced data science workbench project specifically for soil data scientists to develop authoritative soil data from existing geospatial data. If EDAPT will work for DSP Hub, it will need to be outside of the official workbench pilot, support heavy geospatial analysis, and feasibility of using EDAPT and tools evaluated during iterations that are currently under contract with TSPi.*

Using EDAPT can work for DSP Hub to test feasibility under the following conditions:

* A separate workspace is set up in EDAPT that TSPi could develop out the DSP Hub environment and tools outside of the “workbench pilot”
* NASIS and SSURGO soil data needed for the MVP can be added to the EDAPT by January 15.
* A quarantined workspace that has a separate data management and data governance structure maintained by the DSP Hub owners for the libraries of activity data and stored procedures
* Adding data into the DSP Hub is streamlined so that the DSP Hub owners can manage data
* If the EDAPT is determined to not be feasible for the DSP Hub, the decision can be reversed with minimal technical debt to the budget of the DSP Hub project

**Potential timeline limitations with EDAPT that are blockers for the DSP Hub**

* DSP Hub is under contract and in development - EDAPT is not currently ready for the requirements of the DSP Hub as EDAPT is still initiating a pilot on how to do a “Workbench” for data scientists which is what the soil scientists already do and what the DSP Hub is for.
* DSP Hub has urgent deadlines for delivering products for Farm Bill and cannot wait for the USDA pilot to be completed
* DSP Hub short term priority for Program Deputy Area is using CART and CPDES reference data, which is currently not in the EDAPT environment
* The current TSPi contractors on the DSP Hub project do not have access to EDAPT and will need additional licensing and access (if this is even permitted with the Accenture contract)

**Potential technology limitations with EDAPT that are blockers for the DSP Hub**

* Processing requirements (bandwidth/performance) for nationwide analysis of geospatial soil data is extremely heavy and DSP Hub data scientists will need server-level applications for heavy lifting of soil data, not desktop applications.
* DSP Hub must have its own rules and processing engine customized for creating authoritative soil data

**Potential process limitations on EDAPT that are blockers for the DSP Hub:**

* Specific questions on the tools and capabilities of EDAPT have not been answered with enough specificity to make an informed decision
* Speed and flexibility – DSP Hub will produce data products rapidly for urgent initiatives, which means that soil data scientists need to be able to pull in various geospatial and other data sources quickly and perform analysis and testing without the limitation of processing approvals
* DSP Hub products will focus on external process models and various reference data for those models which do not run in EDAPT.
* The DSP Hub will provide data to live to CART/CD processes and in the future other complex models – can this be supported by EDAPT?

**Dynamic Soil Property (DSP) Hub Technology Needs**

* PostgreSQL + PostGIS
* R Studio Server
* Rules engine might have to be a google earth engine or something similar or more flexible.
* Platform to tie everything together (most important but still TBD)
* SQL Server

**The DSP Hub platform (stack/workbench) has to allow for the data stewards to do the following:**

1. Platform must support robust geospatial analysis
2. Data visualization needs to be included in every part of the process (loading, analysis, testing statistical models and algorithms)
3. Platform has to be able to ingest/import raw data and allow for tagging and processing of data, adding additional data such as chain of custody, data source information, and other metadata. Some of this should be automated but need flexibility to do manually. Chain of custody should include a quality check/approvals for upload.
4. Platform has to support developing statistics/algorithms/rules to perform analysis on existing reference data and raw data
5. Tool needs to support manual querying and filtering to curate and aggregate data
6. Tool needs to support automated data mining and machine learning
7. Ability to aggregate and analyze data for some simple quick summaries (counts, averages, quick stats) from the database
8. Data unification and joining from multiple data sets to create new datasets and performing spatial and tabular joins
9. Ability to create and attribute data products (outputs) sets that will be used in various models
10. Analysis tools that allows for complex statistical framework (building and developing new models)
11. The tool includes a “test mode” with data visualization (charts, graphs, maps) to develop and test statistical models and algorithms
12. Tool needs to support generating and providing different output types, including customization of outputs
13. Data curation for external users allows for automated decision-making on systems that an organization uses to manage its interactions with customers, employees and suppliers.
14. The tool should allow for users to perform the above without doing scripting
15. Options to use R Studio server and publish to shiny app
16. Options for Python and SQL
17. API services
18. Predictive modeling

**Example Workflow of a single product needed for the MVP (currently in development)**

DSP Hub process needs for MVP

1. Ingest quality-reviewed activity data (such as results of laboratory analysis soil lab test results from the location (tabular and not standardized data - likely .csv).
2. Ingest geospatial data on the location of the activity data
3. Ingest covariate from various data sources
   * National Planning and Agreements/CART data
   * Individual PRISM datasets (raster datasets, a series of layers with a single variable)
   * Model outputs (final or interim, such as COMET-model emissions per scenario)
   * Geospatial soil data and attributes selected (vector/SSURGO/NASIS)

* NASIS Lab Data Mart – curate data from NASIS for the analysis
* CIG On-farm Soil Health Demonstration data (Excel spreadsheets?)
* Other geospatial data for covariates at the location of the project site

1. Data processing to get into the same data structure/framework (spatial and temporal) for the analysis
2. Load, extract/transform model reference data for identified analyses
3. Develop derivative data from the analysis (could be an interim data product for testing or further analysis)
4. Allow selection from a library of rulesets/functions/codes to:
   1. Existing stored coded processing step/function/ruleset that can be used for a variety of datasets (library of custom rules and functions for SHAPE scoring, COMET, aggregation of property by conservation practices etc.)
      1. Identify necessary rules, crosswalks and data needs
         1. INPUT data
         2. Covariate data
         3. Etc.
      2. Define AOI or other method to define scope
      3. Format INPUT data for analysis (cross-walk inputs into common fields, DSPs and mgmt. info)
      4. Retrieve and process covariate spatial Information (external (PRISM) and internal (SSURGO fields)
      5. Apply previously developed statistical analysis (R SHAPE function)
      6. Output result (SHAPE score with mgmt. info)
      7. Visualize results (maps, graphs by mgmt. info etc.)
      8. Use result as new input for further analysis (Advanced Map or Analyze scores by new covariates)
   2. Development of new output/products- iterative
      1. Identify necessary data, rules and crosswalks
      2. Organize possible covariate info
      3. Develop cross-walks for input and covariate info
         1. Use stored ‘roestta stone’
         2. Develop new cross-walks between datasets
      4. Develop statistical model for analysis
      5. Visualize and test interim results
      6. Define AOI or other method to define inference (where product is viable)
      7. Iterate
      8. Finalize as a known process /ruleset with all data, cross-walk and stats needed for future use
      9. QA/QC of new process/ruleset
      10. Make available for users to re-use as ‘known process’

So specifically for SHAPE

1. SHAPE - soil health scoring
   1. On a quarterly basis, ingest quality-reviewed results of laboratory analysis from project sites
   * Core dataset
     + CIG OF-SHDT
   * Additional data
     + CAP soil health testing
     + SPSD - Lab Data Mart
     + Other cooperating labs?
   1. Identify necessary rules and crosswalks
      1. Updated shape data inputs
      2. Updated crosswalks
      3. Updated stats models
   2. Format INPUT data for analysis
      * 1. join OFSHDT, CAP and LDM data into common fields - DSPs and mgmt. categories and info
   3. Retrieve and process covariate spatial Information
      * 1. PRISM - MAT and MAP
        2. SSURGO - surface texture and suborder
           1. Apply rule to reclassify for statistical analysis
   4. Apply Previously developed statistical analysis
      * 1. R function - developed by researchers and peer reviewed
   5. Output result (SHAPE score with all other information attached)
      1. Visualize results (maps, graphs by mgmt. info etc.)
   6. Perform some sort of QA/QC process
   7. Publish individual results and visualization back to users
   8. Use results as new input for further analysis (Advanced Map or Analyze scores by new covariates)
      1. Aggregate and analyze by categories (mgmt. categories)
      2. Aggregate and analyze by spatial areas (states)
      3. Aggregate and analyze by other covariates (SSURGO, MLRAs, PRISM derivaties)
      4. Create visualizations
      5. QA/QC
   9. Publish aggregation and maps
      1. SHD
      2. SPSD
      3. Users
      4. Other Stakeholders