VITAL SIGNS

Vital Signs Information System Workshop

Bill & Melinda Gates Foundation
Seattle, Washington
October 18, 2013

Contents

The mission of Vital Signs (VS) is to create an integrated system for monitoring, analysis and decision making to address key agricultural development, human well-being and ecosystem service challenges in Africa.

To further this mission, Vital Signs is undertaking intensive and extensive data collection and needs to efficiently manage the data in a way that feeds into advanced models, analytical outputs and index computations. The latter, in turn, need to be presented in a way that communicates complex information effectively to decision-makers, e.g., as visualizations, through online interfaces (dashboards) and through cell phones and tablets. Vital Signs fills a critical unmet need for integrative, holistic measurements of agriculture, ecosystem service and human well-being. The system quantifies sustainability and provides tools to evaluate risks and trade-offs by pooling multi-scale data into an open access online dashboard for policy makers, the private sector and the scientific community. Vital Signs provides indices of sustainable agricultural intensification, resilience, sustainability, water security, climate forcing, biodiversity, nutrition, among others.

The Vital Signs Information System Workshop brought together a set of IT experts with the Vital Signs Technical to identify key design considerations and to discuss staging the development of the information system. The Workshop hosted by the Bill and Melinda Gates Foundation in Seattle, October 18th, 2013.

The objectives of the workshop were to:

1. Review a preliminary infrastructure blueprint for the Vital Signs Information System prepared by Community Commons, provide feedback and provide input on future development needs;

2. Share relevant experience constructing and integrating building blocks of related observatory and information management systems;

3. Review existing observatory systems and software that Vital Signs can leverage;

4. Discuss a transition from infrastructure blueprint to construction and operation phases, including identification of potential partners, and risk mitigation strategies;

5. Discuss costs and benefits of Vital Signs cloud deployment.

The workshop had 17 participants (Appendix 1) with expertise in information system development, software infrastructure for environmental observatory systems, communication networks, ecosystem analysis and modeling, agriculture development, and environmental modeling. The meeting was led by Dr. Sandy

Andelman, Executive Director, Vital Signs (Conservation International), Dr. Cheryl Palm, Deputy Director, Vital Signs (Earth Institute, Columbia University), Dr. Bob Scholes, Deputy Director, Vital Signs (CSIR, South Africa), and Dr. Ilya Zaslavsky (University of California San Diego, San Diego Supercomputer Center). The Gates Foundation was represented by Dr. Stanley Wood and Kate Schneider.

To support communication with participants before, during and after the workshop, organizer developed a web site with the agenda (Appendix 2), reading materials, workshop minutes, presentations, etc. (https://sites.google.com/site/vsainfosystem/). Subsequent to the workshop, an initial issues backlog was added to the web site, to aid during the construction phase of the project.

The morning discussion focused on the Vital Signs activities and its initial information system design. An information system blueprint (Figure 1) was presented, and its outline is described below. The afternoon discussion focused on selected Vital Signs information management issues, and deployment strategies and options.

1. Overview of the Vital Signs information system design

As an information system, Vital Signs is designed to enable:

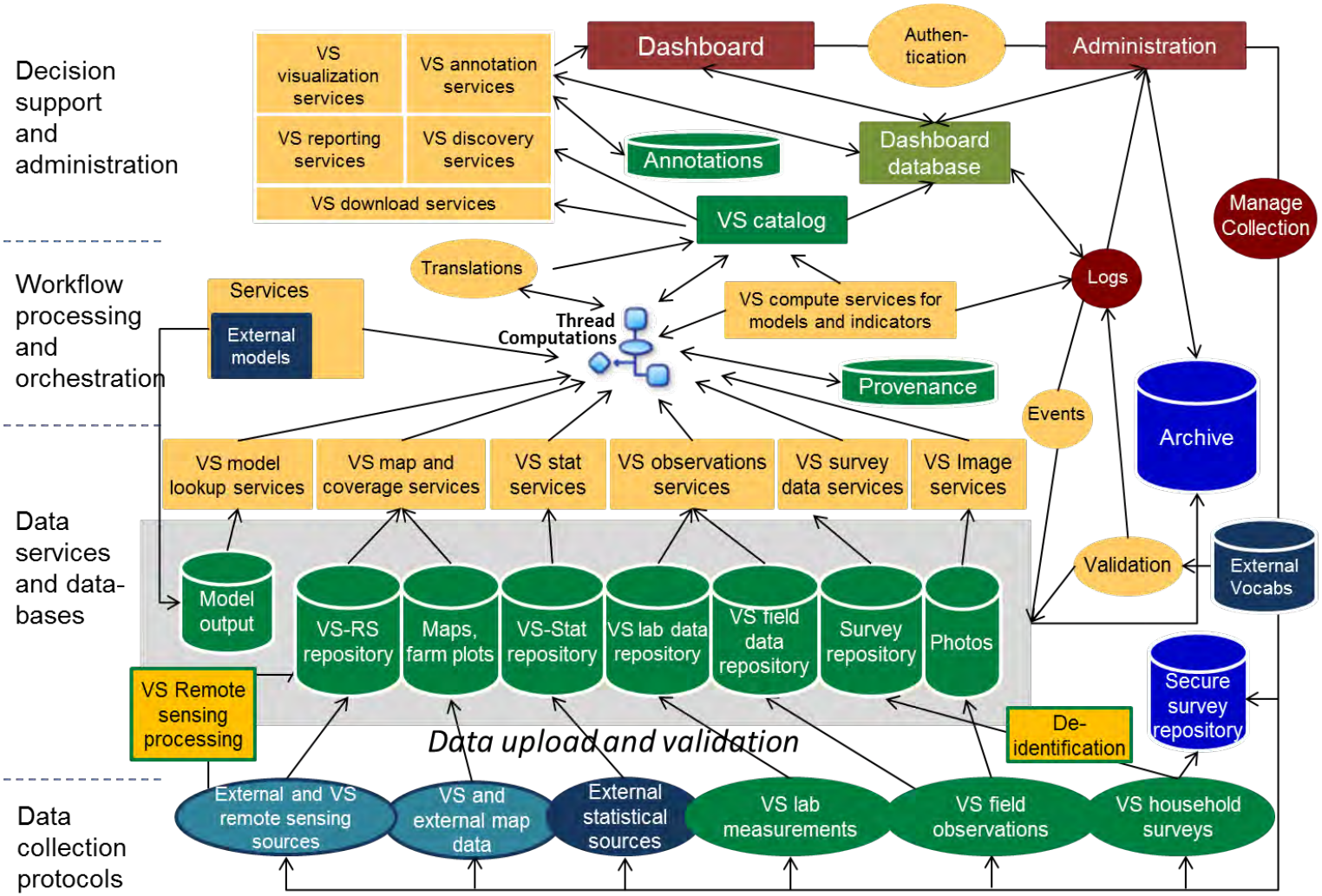
- Comprehensive data capture, data publication, discovery, interpretation, access and integration across internal data gathering protocols and multiple heterogeneous external data sources;

- Analysis and modeling to generate actionable information for decision-makers based on key indices of ecosystem services, agriculture and human well-being;

- Comprehensive interdisciplinary communication, data dissemination, knowledge sharing and collaboration among Vital Signs stakeholders and users.

To be economical and efficient, Vital Signs should leverage and integrate information systems, analysis and modeling modules developed in several disciplines. Designing for the necessary flexibility and scalability dictates a loosely-coupled, service oriented and standards-based architecture approach. Therefore, an important part of the system will be modules for managing standards-compliance of Vital Signs services and applications.

Figure 1. Generalized Vital Signs architecture diagram / blueprint.



In addition, the system outline will include an initial assessment of relevant existing tools and interfaces. Compliance with adopted standards such as those developed through ISO, OASIS and OGC processes will ensure that the Vital Signs information system can re-use standards-based third-party software, in particular free and open source software, and make it easier to evolve the system as modules with new capabilities become available.

As our understanding of dependencies and feedbacks in the system of ecosystem services, agricultural production and human well-being evolves rapidly, the Vital Signs information system shall also evolve. In particular, this means that the system shall be able to incorporate new types of measurements, new methods and models as they become available. While it is impossible to design against yet unknown future system requirements, the flexibility and configurability of the system can be improved if the Vital Signs information system is designed to anticipate such changes.

This can be done in several ways: (1) designing the system in a modular and layered manner, where capabilities can be refined and modules extended or replaced as new functionality is added; (2) identifying building blocks and layers of the overall Vital Signs information system that are responsible for a general class of capabilities (e.g. data discovery, data interpretation, subscription to information feeds, high performance computations) and ensuring that the implemented blocks are scalable, configurable, extendable, accessible via documented APIs – and can be replaced with other implementations as needed; (3) following standards and best practices for information exchange, in particular standards that have shown their longevity and are managed by respective standards development organizations (SDOs); (4) ensuring that feedback about system limitations, data quality, indicator interpretation, or new requirements and use cases is easy to provide via annotations and related building blocks.

The Vital Signs information system blueprint is being developed to address the following issues:

- Description of users and use cases to be supported by the Vital Signs system, and mechanisms for engaging them in providing requirements and feedback

- Initial scope of Vital Signs requirements and needs, with respect to Vital Signs sampling frame, analysis threads, protocols and data collection methods

- Review of available systems and building blocks previously developed for neighboring domains or related projects and addressing similar challenges

- Review of available data sources, data collection protocols, and process models specified as part of analytical threads, with specific attention to source metadata, available formats, units, data volumes and data quality

- Support for data collection workflows, to ensure that the collected data are verified and become immediately available for model and indicator computations, as well as disseminated to interested stakeholders

- Robust decision-making support that includes effective visualizations, and indicators that are easy to interpret and trace to data sources and models

- Cost-effective cloud-based implementation strategies and implementation risks.

Preliminary VS documentation identified fourteen analytical threads (Table 1) that represent "thematic clusters of information relating to a topic". Each thread includes measurement, analysis, indicator and index layers.

Table 1. Vital Signs Analysis Threads

1) Climate forcing
2) Poverty
3) Wood Fuel
4) Soil Health
5) Livestock
6) Sustainable Agriculture Intensification
7) Water
8) Nutrition
9) Resilience
10) Health
11) Biodiversity
12) Disability Adjusted Life Years
13) Food Security
14) Sustainability

To support each thread, Vital Signs is conducting extensive data collection. Detailed descriptions of the following data collection protocols (Table 2) are available on the VS web site. In addition, remote sensing data acquisition and processing protocols for Vital Signs is being developed by the University of Maryland.

Table 2. Vital Signs protocols (available from vitalsigns.org)

- E-Plot Biomass Measurements
- E-Plot Soil Sampling and Processing
- Rapid Roadside Assessment
- Water Availability and Quality
- Weather Stations
- Household Survey
- Agricultural Management Intensity Survey
- Farm Field Soil Sampling and Processing

Discussion of the requirements and characteristics of selected threads was followed by a more detailed examination of selected threads and associated implementation issues - presented in the next section.

2. Information Support for data collection protocols

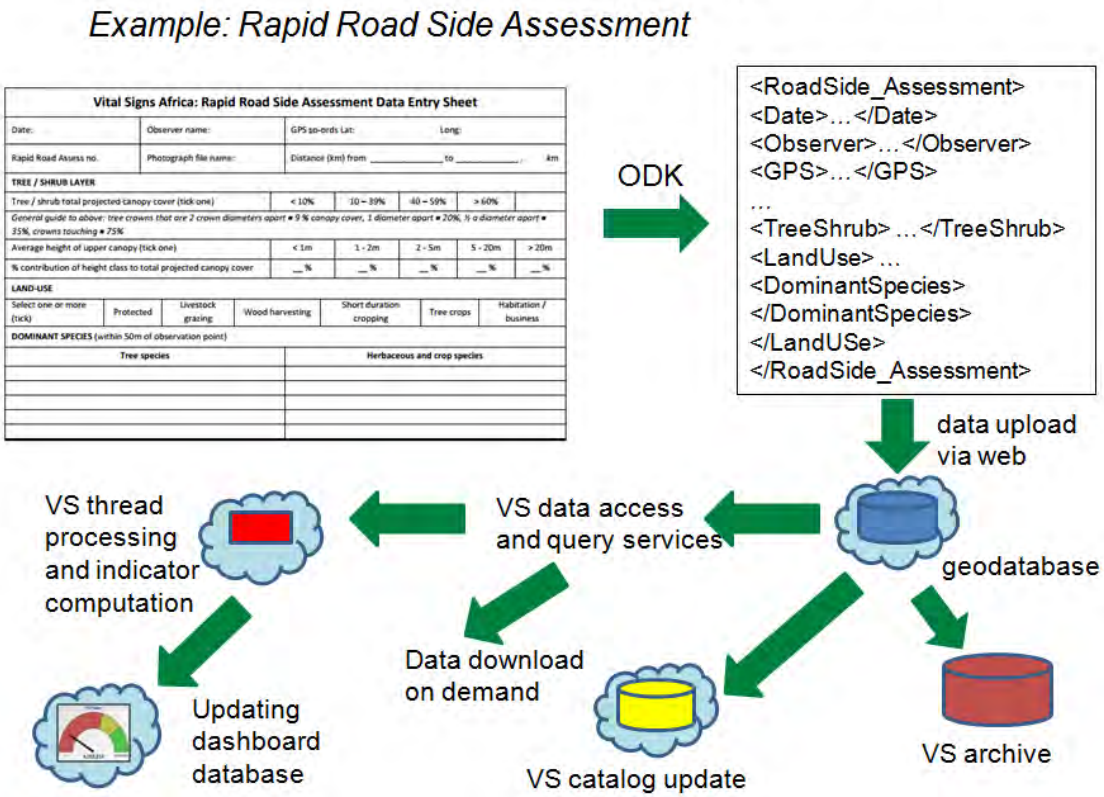
The Rapid Road Side Assessment protocol was discussed as an example of data collection protocol implementation within the Vital Signs Information System (Figure 2). Key elements of the data workflow include:

- The Rapid Road Side Assessment Data Entry Sheet is currently used to collect data in the field. Currently the data from the paper form is subsequently entered into a tablet, using forms developed with Open Data Kit (ODK) software. This process is described in the Rapid Roadside Assessment protocol document available from the Vital Signs web site. Vital Signs will be phasing out the paper data collection forms and will transition to tablet-based data collection once the tablet template stabilizes.

- When the data collection device is connected to the cell phone network or to the internet, or the data are otherwise transmitted to a connected device (e.g., through a laptop), the gathered information (typically, for multiple road side assessment events) is loaded in database on a Vital Signs server, which resides in the cloud. The database may implement a schema common for databases designed for managing observations and objects (photos), and include standard metadata at the assessment event level (ID, date, time, location, observer name) plus metadata for additional characteristics or lists (i.e. characteristics of vegetation, land use, dominant species). Such a schema can be implemented in any relational database management system that supports spatial indexing (PostgreSQL, MySQL, MS SQL Server, etc.), imported into SOLR, or managed in ESRI's geodatabase.

- The raw data will then be made available via data access and query services – ideally using standards-compliant encodings and service interfaces. For this type of data it may be a standard GML-encoded data (GML is Geography Markup Language, adopted as an international standard via ISO (ISO 19136) and OGC specification processes made available via WFS (Web Feature Service interface specification), and/or GeoRSS feeds, and/or as GeoJSON objects. Having the data exposed via multiple standard mechanisms is preferred as it will support a more flexible design at the analytical/modeling and the indicator levels.

Figure 2. Information system components supporting a data collection protocol (with Rapid Road Side Assessment as an example)



- In addition, the newly submitted rapid roadside assessment records will be added to a VS archive. Respective catalog records in a VS metadata catalog will be updated as new data become available. This will support data availability requests as well as metadata browsing, sub-setting and data download.

- The data available via standard data exchange mechanisms (e.g., WFS/GML, GeoRSS, GeoJSON) will be then used as input in processing routines and models specified for different threads, and used to generate indicators made available via a VS dashboard.

- Additional components of the data processing workflow (not shown on the diagram below) concern data validation and quality assurance (e.g. validating the entered data against controlled vocabularies or allowed value ranges).

The discussion during the workshop confirmed that this overall design matches expectations and typical workflows implemented by both ODK and Ushahidi teams. Information support for other VS data collection protocols will follow the same overall pattern – but details may differ significantly.

One obvious difference for other protocols will be reliance on other types of standard data encodings and services, appropriate for the data. For example, water gauge time series information would be transmitted using the Water Markup Language standard; remote sensing data may be made available via WCS (Web Coverage Service), household survey information and codebooks may be encoded using DDI (see Table 3).

Table 3. Possible standards-based information encoding and service interfaces for selected VS data types (*)

Type of Data	Common formats used for data exchange	Standard information encoding	Standard service interface
Household surveys	SAS, SPSS	DDI	REST
Map data	ArcGIS, shapefiles	GML, KML	WMS, WFS
Climate series (grids)	ASCII, NetCDF	ASCII, NetCDF, GRIB	WCS, OPeNDAP
In situ time series	ASCII/CSV	WaterML, O&M, GML	SOS
Statistical data	Excel, ASCII/CSV	Possibly GML	oData, WFS
E-plot observations	Excel, ASCII	WQX, O&M	SOS
Remote sensing images	Difer by product	TIFF; GML and ISO 19115-2 for metadata	WMS, WCS
Local imagery/photos	JPG, ad hoc	JPG, PNG	
Agricultural intensity	CSV, Excel	ICASA standards	ICASA
Annotations	-	W3C open annotation, RDF/OWL/SKOS	REST SPARQL
Provenance	-	Open provenance model (OPM), W3C PROV	

Additional differences would derive from specific de-identification and privacy requirements of household survey information, or the need to manage and re-publish large volumes of remote sensing imagery collected by the Vital Signs project. Discussion at the workshop pointed to key features of different types of data collected by the project, which need to be taken into account in information system design.

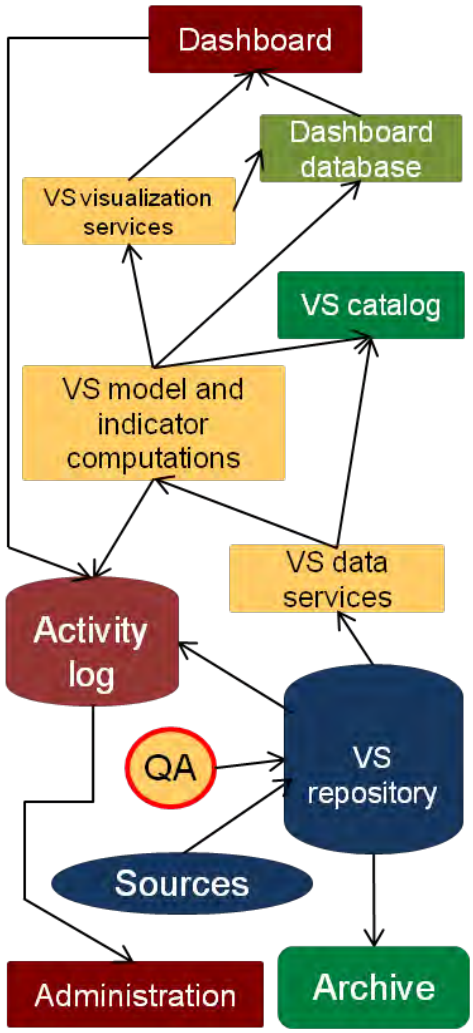
(*) The acronyms in the table refer to:

STATA: Data analysis and statistical software (www.stata.com)
SPSS: Statistical Product and Survey Solutions (http://www-01.ibm.com/software/analytics/spss/)
DDI: Data Documentation Initiative (http://www.ddialliance.org/)
REST: Representational State Transfer (http://en.wikipedia.org/wiki/Representational_state_transfer)
GML: Geography Markup Language (http://www.opengeospatial.org/standards/gml)
KML: Keyhole Markup Language (http://www.opengeospatial.org/standards/kml)
WMS: Web Map Service (http://www.opengeospatial.org/standards/wms)
WFS: Web Feature Service (http://www.opengeospatial.org/standards/wfs)
ASCII: American Standard Code for Information Interchange (http://en.wikipedia.org/wiki/ASCII)
NetCDF: Network Common Data Form (http://en.wikipedia.org/wiki/NetCDF)
GRIB: GRIdded Binary (http://en.wikipedia.org/wiki/GRIB)
WCS: Web Coverage Service (http://www.opengeospatial.org/standards/wcs)
OPeNDAP: Open-source Project for a Network Data Access Protocol (http://opendap.org/)
CSV: Comma-separated values (http://en.wikipedia.org/wiki/Comma-separated_values)
WaterML: Water Markup Language (http://www.opengeospatial.org/standards/waterml)
O&M: Observations and Measurements (http://www.opengeospatial.org/standards/om)
SOS: Sensor Observation Service (http://www.opengeospatial.org/standards/sos)
oData: Open Data Protocol (http://www.odata.org/)
WQX: Water Quality Exchange (http://www.epa.gov/storet/wqx/)
ICASA: International Consortium for Agricultural Systems Applications (http://www.icasa.net/)
W3C: World Wide Web Consortium (http://www.w3.org/)
RDF: Resource Description Framework (http://www.rdfabout.com/intro/)
SPARQL: Simple Protocol and RDF Query Language (www.w3.org/2009/sparql/)
SKOS: Simple Knowledge Organization System (http://www.w3.org/2004/02/skos/)
OWL: Web Ontology Language (http://www.w3.org/TR/owl-features/)

1. Mapping data flows in selected threads to information system components

Additional VS information system components supporting data flows for a typical thread are shown in Figure 3. Once the collected data are uploaded to a database in the VS data repository, the data are quality controlled, archived offline, and exposed via standard data access services. QA/QC is done at several levels: initially within data collection forms (functionality available in ODK), followed by quality control at the repository level (validation against controlled vocabularies and database constraints). Eventually, quality control annotations can be added by users for other phases of model and indicator computations, via a user dashboard.

Figure 3. Key information system components supporting VS threads



VS data sources are characterized using the following metadata descriptions:

- Collection protocol or URL for data retrieval;
- Data format;
- Data volumes and update frequencies;
- Data access API if available, or procedure for loading data into repository;
- Data schema and encoding, including metadata fields, units, coordinate system;
- Spatial granularity (i.e. at which spatial scales the data are accessible);
- Data quality information;
- In which thread and to which model the data source provides input;

- If it is an external data source: its stability, reliability, any data access agreement in place;

- If it is an external data source: known costs associated with obtaining, curating, transforming data from that source;

- For a VS data collection protocol: operational arrangements and people involved.

To ensure that the collected data can be used as input into model and indicator computation, similar characteristics need to be developed for VS models, including:

- List of input variables;

- For each input variable: descriptive name, units, format, spatial and temporal granularity, gridding scheme if applicable, time-stepping scheme if applicable; coordinate system if applicable, approximate volume of data, plausible value ranges;

- Availability of external model implementation or pseudo code that can be implemented by VS;

- Key model assumptions, a reference to publications if applicable;

- A list of outputs;

- For every output that VS will use in indicator workflows: data format and schema, units, spatial and temporal granularity, data volumes;

- Data quality requirements for inputs and uncertainty estimates for outputs, if available;
- Model calibration information;

- What other VS components the model connects with;

- Expected frequency of model runs.

Based on these characteristics, one can detect potential mismatches – in data structures, formats, units, access protocols, coordinate systems, resolutions, variable semantics – between available data and expected model inputs, and develop translation services where necessary.

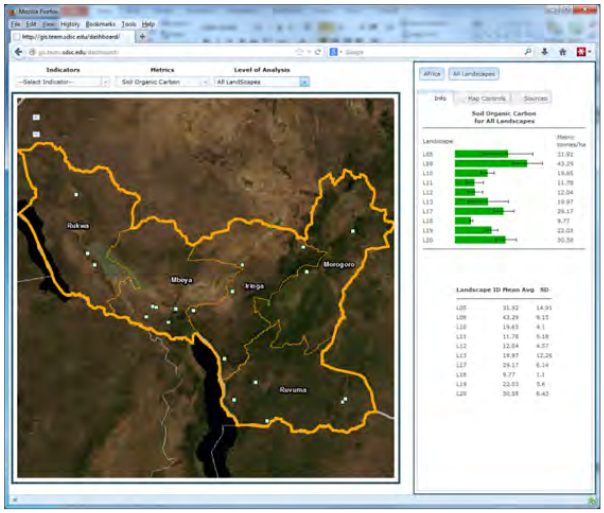
We expect that indicator computations, translation services, and models (in particular those relatively simple models that the VS team is planning to implement internally) will also adhere to a standard service interface specification, such as the OGC Web Processing Service (WPS) standard. This will ensure that the models and other processing routines will publish their functionality, inputs and outputs in a standard and unambiguous manner, and can be consistently catalogued, discovered, invoked, and chained in processing workflows. This will also support VS information system evolution towards newer and better model implementations as they will seamlessly replace earlier versions as long as model interfaces remain the same.

As part of the proposed information system, all system components, including datasets and their versions, models and other services and their versions, will be registered in a standards-based catalog. Virtually all of the data in the VS system are geo-referenced, hence it would be appropriate to use a CSW-compatible catalog (OGC Catalogue Services for the Web), accessible online via a Geoportal or CKAN portal interface, and programmatically. The catalog will let VS system users discover, interpret and download VS resources, including raw and derived datasets, data services, models, indicator computations and other services. It will also provide information for developers working on additional models and indicators based on the same VS data foundation.

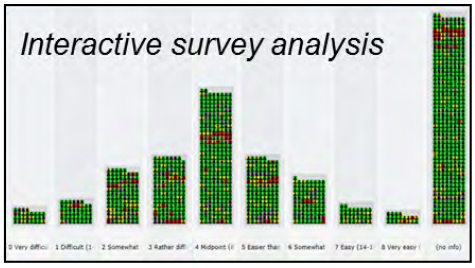
Indicators will be presented via a dashboard or on cell phones, which will expose different sets of indicators for different types of users. The dashboard should be flexible, customizable for different types of users (decision-makers, policy analysts, NGO and general public, or users interested in specific threads), and present a range of indicator tracking, visualization (as tables, maps, charts), interpretation, annotation, discovery, reporting and data download capabilities. There was a discussion among participants about whether dashboards are obsolete. In particular, concern was expressed that if a user goes to a dashboard and then subsequently – whether it is an hour later or a month later – nothing has changed, research has shown that the user will never go back. So a dashboard is an appropriate tool for visualization only if the visualization changes more frequently than the users consult it.

Examples of dashboard components are shown in Figure 3. In addition, the same dashboard front-end would be used for operations management and tracking of VS team data collection efforts and issues. To support these capabilities, building a dashboard over a common portal framework, such as Drupal or CKAN, will have advantages over a standalone dashboard module: in particular, the ability to incorporate user management and many freely available modules. Pre-computed indicators and model results will be loaded into the database underlying the user portal (dashboard database), to make the data readily available to users.

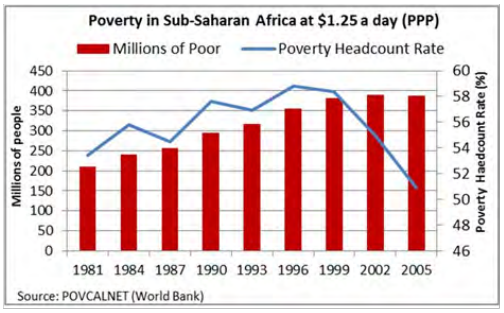
Figure 4. Examples of dashboard components for VS dashboard (The initial dashboard snapshot is from a prototype ecosystem services project in the southern highlands of Tanzania).



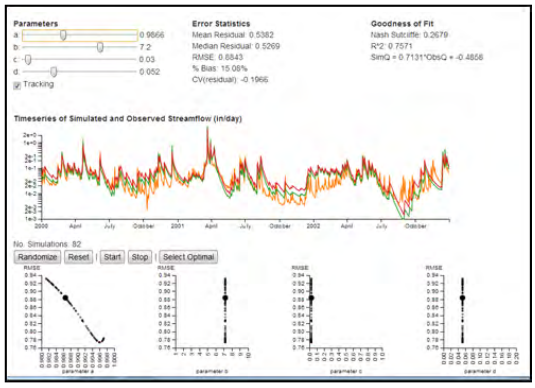
Initial dashboard interface based on ESRI ArcGIS Server



Dial charts…



Tracing key indicators over time



Simulation modeling interface

In addition, all changes in the system, and user actions, will be recorded in the activity log. The activity log will be monitored for specific types of events that require intervention of VS administrators, for example: detected anomalies in data, user annotations referring to data quality, missed operational deadlines, non-responsive servers or services. Alerts about such events will be made available through the administration part of the dashboard, and possibly sent to responsible parties via email or SMS.

An architecture diagram for the entire system is presented in Figure 4. It is necessarily generalized to fit on one page. Details of the system architecture, including components not shown in the diagram, are described below, by Vital Signs tiers from data collection to data services and databases, to analysis, modeling and indicator computation, to dashboard services and online interfaces for system users and administrator

Data collection tier. This is a generalized view of data collection procedures envisioned in the VS project. We organize elements in this layer by types of data collected rather than by the specific protocols listed in Table 2. Data collection protocols can be grouped into: remote sensing imagery, collected by the project or obtained from other sources; map data and digitized farm plot geometries; statistical and other data obtained from external sources; measurements obtained through lab processing of samples collected in the field; regular or irregular observations recorded in the field (including stream gauge and weather station data); household and other surveys (shown as ovals at the bottom of the diagram).

Some VS data collection protocols generate data of similar types (e.g. field observations), suggesting that they can be managed in similarly structured databases or repositories using metadata schemas common for that type, and then published using standard service interfaces and data encodings as listed in Table 3. In addition, several protocols generate data of more than one type (e.g. farm plot geometries and agricultural intensity measurements), suggesting that these data will be managed in distinct (though linked) parts of the system.

Data services and databases tier. The purpose of this tier is to establish a solid and consistently managed data platform for the different types of data in the VS information system, and enable construction of multiple analytical and modeling workflows and computation of different indicators. The collected information is initially validated on input into data forms (e.g. using Open Data Kit), and then ingested in a Vital Signs repository, following the process depicted in Figure 1. The repository includes a number of databases for different types of data collected by Vital Signs (green shapes within the gray rectangle in the diagram).

Where possible, the database schemas will follow common specifications for these data types (e.g. the Observations Data Model (ODM) – for field observations at point locations.) Besides the initial on-entry validation, loading data into these databases will include additional integrity checking and validation, e.g. against externally managed controlled vocabularies (for vegetation species, crop types, etc.) Regular synchronization of external vocabularies with local data entry devices would ensure that local validation rules are current and consistent across data collectors. Several types of data will require additional data ingestion steps, such as removing personally identifiable info and location obfuscation (for household surveys and similar data) and processing of raw remote sensing imagery into remote sensing products.

Information in VS repositories would be accessed via documented service interfaces (a row of yellow rectangles above the repositories), ideally via standard

services as per Table 3. An additional step specific for household survey data is placing unaltered survey responses in a secure survey data repository, to ensure that personally-identifiable information is not accidentally shared with unauthorized users. Data loads into VS repositories will be analyzed for anomalies, with alerts generated on any detected events that require attention of VS administrators. Besides information obtained through VS data collection protocols, the repositories will include output of VS and external models, exposed through model lookup services.

In addition, all data will be archived in a separately managed offline data store, to ensure that the data are preserved for long-term use.

Workflow processing and orchestration tier. This tier of the VS information system will enable organization of available data services and processing services into executable workflows specific to each VS thread. A key requirement for this tier is to be easily configurable and extendible, so as to incorporate new workflows and indicator computations as they are requested by decision-makers or analysts. A workflow engine at the center of the system can be implemented using free workflow systems such as Kepler, Taverna, Ptolemy, or Trident; for managing simpler processing workflows chains of standard WPS services can be used as well. The workflows will organize data services and processing services into generation of indicators. The processing services would represent process model wrappers and translation services (coordinate transformations, unit conversions, etc.)

As mentioned earlier, VS will maintain a range of relatively simple models, which could be invoked via the WPS service interface. In addition, the VS system will make use of externally run models, either accessing them via an agreed upon interface (such as WPS), or incorporating model computations in the VS repository and updating it on a regular basis. Workflow steps will update provenance information so that the computed indicator values can be traced back to input data and processing algorithms. This information should be available to users of the system on demand. All data services, processing services and workflows will be registered in VS catalog.

Decision support and administration tier. This tier will manage user interactions with the VS system, and provide services for information visualization, reporting, annotation, data discovery and data download. These services will be available to users and VS system administrators accessing the VS dashboard and the VS administration interface.

As mentioned earlier, implementing the dashboard and the administration interface within a common portal framework would be advantageous because of the many built-in functions, availability of a large number of third-party data management,

cataloguing, visualization and other modules, easier integration with other systems, and a large development community. The database supporting the portal (referred to as the dashboard database on the diagram) will be populated with pre-built indicators and other data that should be readily available to users. The administration interface will have additional capabilities of tracking system events and responding to alerts generated within the system or annotations posted by users, and adjusting data collection management as needed.

4. Implementation options and cost considerations

There is no single platform that would cover the entire set of capabilities outlined above. However, a combination of software components from several vendors and community projects would address most of the needs. The sources of software to be considered for the system include:

- ESRI ArcGIS Server, GeoPortal, GeoEvent Server, GeoAPI, and Model Builder. These components can be used to implement the user dashboard (custom GeoPortal or ArcGIS Server), databases and services for geo-referenced data (ESRI geodatabases; WMS, WFS, WCS services), geoprocessing and spatial data translation services (ArcGIS Server geoprocessing services can expose WPS capabilities), VS catalog (using Geoportal Server which exposes CSW catalog interface). ArcGIS Model Builder can be used to implement threads workflows, in a system where all workflow processing components are available as ArcGIS geoprocessing tools or can be scripted with Python.

- Boundless (boundlessgeo.com) provides a free alternative to the ESRI set of tools. The OpenGeo Suite includes PostgresSQL/PostGIS, GeoServer and GeoWebCache for serving map and other data, and OpenLayers user interface combined with Boundless SDK for rich web map applications.

In addition, both ArcGIS Online and Boundless provide a managed cloud platform that can be used to host VS data. An alternative hosting mechanism would rely on local servers or virtual servers hosted in the Amazon cloud or a similar option (a self-hosting solution). A brief comparison of pros and cons of the two approaches is given in Table 4.

Several packages were suggested by workshop attendees: ODK for managing forms-based data collection on mobile devices, and Ushahidi CrowdMap, which can be deployed as a crowdsourced solution for mapping events reported by users. ODK and the Crowdmap were discussed at the workshop as indicated in the notes below.

Table 4. Hosting alternatives: managed (e.g. ArcGIS Online, Boundless) vs. self-hosted (local servers or amazon cloud)

Managed	Self Hosted
Higher cost for servers	Lower costs for servers
No cost for licenses (built into total cost)	Lower costs for software licenses
Lower costs for day to day management and maintenance	Higher costs for day to day management and maintenance
Same cost to customize or build custom solutions	Same cost to customize or build custom solutions
Doesn't offer all the components; hence would need to build	Can select or build components

- There are many other popular standards-based open source community projects that can be leveraged for various VS components. See:
    - http://OSGeo.org for systems dealing with spatial data;
    - http://www.opengeospatial.org/resource/products/compliant for a list of standards-compliant software products;
    - ESPER/NESPER (http://esper.codehaus.org/) for complex events processing;
    - Workflow systems (https://kepler-project.org/, http://ptolemy.eecs.berkeley.edu/, http://www.taverna.org.uk/);
    - Portal frameworks (https://drupal.org/, http://ckan.org/, http://geoportal.sourceforge.net/);
    - Relational databases with Spatial data support (http://postgis.refractions.net/, http://www.mysql.com/);
    - A recent review featured 30 free tools for data visualization that can be used for dashboard visualizations (http://www.computerworld.com/s/article/9214755/Chart_and_image_gallery_30_free_tools_for_data_visualization_and_analysis);
    - Free survey analysis online can be done with a system such as the one implemented for NSF EarthCube surveys (http://connections.earthcube.org/ecsurvey/).

5. Key workshop recommendations

a) Refining the scope and requirements of the Vital Signs Information System

a. The project will target decision-makers, donors, international organizations, NGOs, possibly large cooperative farmers, but not individual farmers (at least in Phase 1).

b. The initial focus of use cases will be supporting needs of policy analysts as proxies for decision-makers.

c. The key indicators to explore and use as the model for the entire VS development should include agricultural sustainability and human well-being, with environmental response and ecosystem services being the key feedback mechanisms to be considered. Sustainable increase in productivity and improvement of human well-being are the main criteria.

d. The system should be modular, support standards-based interoperability, follow loosely coupled, service-oriented design, and rely on free open source components where feasible.

e. System components should have well defined inputs and outputs, so that standardized procedures can be implemented in various ways and plugged in existing threads.

f. The system should be resilient to varying internet availability and electrical outages.

g. The system should be relatively simple to operate and maintain. This means that the number of technologies and platforms should be limited. A limit is defined by one FTE for operation, maintenance, troubleshooting and customization.

h. In the current phase of the project, the system won't focus on crowd sourced data.

b) High-level design recommendations for VS Information System

a. The design should initially focus on data collection protocols already being implemented by the Vital Signs team. This will ensure that the collected data are well managed and provide a good foundation for the modeling and indicator levels (Vital Signs Data Platform).

b. On the contrary, the key requirement for the modeling and indicator layers of the system is flexibility: we don't want to prescribe specific models or indicators to be computed since use cases and decision-making needs will vary. Rather, we need to demonstrate how models, indicators and visualizations can be computed from the collected data following standardized data access interfaces, and how they can be compared across countries and jurisdictions. Others should be allowed to design indicators and visualizations over the same Vital Signs data platform, as needed for specific use cases. While a design consideration for VS information system, support for incorporating user models in VS is not part of this phase of the project.

c. All data should be georeferenced at the appropriate granularity level.

d. Data, models and indicators used in VS should be catalogued, and changes should be available via an RSS feed. All resources should be described consistently; besides common standard metadata should have update frequencies, data volumes, all inputs and outputs; for models: version, platform, compiler, calibration procedure, and uncertainty analysis (the latter mostly for scientists, not for policy makers).

e. It would be important to consult similar projects that integrate environmental, social, and economic data in an observatory setting (e.g. NSF Water Sustainability and Climate, EarthCube, Kilimo Salama (http://www.syngentafoundation.org/index.cfm?pageID=562), FEWS (fews.net)).

f. Find out how people work with data for decision-making, and what visualizations or aids they use. For this, it would make sense to have a series of small pilot projects/demos that would present data management and visualization options to decision makers, leveraging existing software with minimal customization – accompanied by ongoing evaluation.

g. Shadowing decision-makers would be excellent, though not likely to be feasible. Having specific use cases – especially a policy use case to demonstrate the policy impact pathway – is key. One such case would focus on a decision-maker who needs repeatable and valuable indicators that should be tracked over time at the top level. Another use case would be a policy analyst, with more sophisticated data and indicator needs. Another case would be people who just need the data.

h. Ultimately, VS is seeking to transform decision-making behavior, make it more data-driven. This is a huge challenge, especially considering the question of local ownership. Need to educate and nurture an "ecosystem" of local users, invest in community building, and take into account the lead time needed for decision making.

i. Once an initial set of indicators is constructed, VS could run a competition for the "best dashboard" (eg via Kaggle).

c) High level deployment recommendations

a. Need a beta version of the system – even with mock data – ASAP to demonstrate to potential users. Get something that works, then popularize, demonstrate to decision-makers, build on it, cultivate local ownership, and create a community of users. Having

b. Cloud deployment is a logical cost effective choice these days. However, it is unclear whether remote sensing data and processing should be cloud-based, without more precise estimates of data volumes, processing load and frequency, and I/O.

c. Leverage existing software as much as possible.

d) Specific design recommendations: Privacy

a. Individually identifiable information should be stripped from datasets before they are made available for external use. For georeferenced information at the household level, aggregation and/or location obfuscation shall be used for privacy protection.

b. Two versions of datasets should be generated: with complete de-identification, and potentially identifiable.

c. To ensure that personal identity is not traceable, VS may also limit the types or number of queries that users are allowed to make over an online system (assuming that an unlimited number of queries would let one slice and dice the system such that to extract identities.)

d. Mappings between record IDs and personal information shall be kept secure.

e) Specific design recommendations: Annotations

a. Any part of the system shall be annotatable.

b. At this point, we won't impose strict annotation templates, apart from requesting an annotation "category" (data quality annotation, analysis annotation, risk annotation etc.) Annotations would be full-text indexed for searching.

f) Specific design recommendations: Validation and provenance

a. We should not assume data as valid a priori. Data can be corrupted, intentionally or unintentionally, and can be incomplete.

b. Data validity and transparency need to be emphasized at different levels in the system: through visualization at the dashboard (e.g. show sites with worst data quality), through keeping track of data provenance, through validation against constraints within entry forms, local database, and a central VS data repository.

c. Data validation should be conducted in all phases of processing threads. During upload, data will be validated against locally stored range constraints and controlled vocabularies. Local users may override the constraints (e.g. selecting "Other" if an appropriate vocabulary term is not available). The constraints and vocabularies should be periodically synchronized with a central server, where all new entries will be validated by a curator. ODK may be able to detect various data input errors and biases. In addition, looking at outliers and trends would be useful for asserting data quality.

d. Ideally, possible input values would be constrained to a subset based on location of data entry (e.g. only local species will be shown).

e. Cross-validation (e.g. across different collected datasets) won't be used in this phase of the project – apart from photos and ground observations that would be used for reality checks and image interpretation.

f. Instrument calibration values shall be part of common instrument metadata, and include calibration procedure and the last time calibration was performed.

g. Fitness for use, rather than abstract data quality measures, should be a key assessment factor.

h. Reputation model can be used to characterize quality of data coming from different enumerators; wrong data should be annotated as such.

i. Automatic validation should be combined with QA/QC training and spot-checking.

j. There is a potential to use crowd sourcing outlets for data quality (e.g. crowdflower.com, https://www.mturk.com/mturk/).

g) Specific design recommendations: Models

a. VS will be responsible for documenting and maintaining selected model implementations that are components of defined threads.

b. Simple models will be incorporated in VS based on pseudo code. More complex models may be run by VS partners on a specified schedule.

c. Must maintain and report model statistics.

d. Take advantage of external engines, e.g. Google's EarthEngine.

h) Specific design recommendations: Dashboards and Indicators

a. A VS dashboard should not be fixated on a limited set of indicators, but allow other indicators to be plugged in (depending on decision-making needs and use cases).

b. Indicators are interconnected, and should change simultaneously if dashboard allows for some elementary re-computation of indicators.

c. We should separate routine and ad hoc indicators. Routine indicators should be computed continuously as outcome of VS threads, and allow for analysis over time and cross-jurisdictional comparison. Ad hoc indicators should be possible to construct from the VS data platform, but they don't have to be traced over time.

d. Indicators for key concepts (productivity, well being) should be organized into groups (vectors of indicators).

e. Portal may be a more appropriate term for the front-end. The portal – implemented in some common portal framework such as Drupal – may include modules that expose VS data (via lookup/browse, search

and download), support maps/charts/narratives, and may also include operational modules/indicator alerts (depending on use case)

f. Types of indicators and capabilities to be considered for dashboard: threshold alerts (though VS is not per se a warning system), simple "calculators" (a'la http://www.fool.com/calcs/calculators.htm)

i) Specific design recommendations: Archiving and dissemination

a. VS-collected datasets should be made available as downloadable Vital Signs products, in standard formats, with accompanying documentation and use agreements.

b. Derived datasets should be made available subject to licensing agreements for original data, and Vital Signs use agreements.

c. Data download agreements, including disclaimers as to accuracy and fitness for use, should be in place.

j) System components that VS information system could leverage

a. To lower costs, existing technologies should be used as much as possible.

b. ODK. The system has been already deployed as part of VS data collection workflows. In other projects, it was extended to database uploads and service interfaces, i.e. similar to how it is proposed in the VS blueprint. ODK can handle off-line data collection for many types of data, including forms and imagery, working on incorporating handwritten notes (another option for that is Captricity.com). ODK can validate data against several types of errors, including systematic encoder's bias. ODK also supports synchronization between local and central databases.

c. Ushahidi developed a crowd sourced mapping platform (crowdmap), which may be used to assimilate and present georeferenced data, from emails, tweets, photos, texts. Alerts can be generated from the data. It is open source. Another development is BRCK, which can be used to connect up to 28 sensors, and communicate measurements; has Ethernet and dual SIM card. SwiftRiver is their analysis platform (the Ushahidi presentation is available at the workshop web site).

d. ESRI. The Gates Foundation is in communication with ESRI on potential use of the ArcGIS platform and services.

k) Support for in situ and ex situ data collection

a. To effectively trace sample information, the IGSN (International Geo Sample Number) system may be used, with samples pre-printed before going in the field. Since samples are being numbered using internal data collection protocols, a mapping from these numbers to IGSNs would be needed.

b. In situ sensor data: uploaded once a quarter (e.g. water levels).

c. To manage both types of data, ODM2 (Observations Data Model v 2) may be used. Its promise is to provide a common data model for sensor measurements and samples. It is still under development (there is a respective NSF project).

d. Standard-compliance is preferred, but having well-documented data collection and dissemination is a must.

l) Potential partners mentioned

a. IBM (Nairobi), Thoughtworks, Ashesi Univ (Ghana). South Africa has a number of high tech companies. Ihub (Nairobi, 125 companies, including a Supercomputer Cluster, and a consulting and research arm). Afrilabs.com (has a marketplace). http://www.praekelt.com/ (mobile campaigns)

APPENDIX 1: Workshop Participants

Name	Affiliation	Interests and Expertise
Lee Cooper	UCLA Anderson School	Management in information rich environments, action research
Jan Dempewolf	University of Maryland	RS and GIS for agriculture, ecosystems, LULC dynamics
Marc Levy	Center for International Earth Science Information Network (CIESIN)	Sustainability metrics; natural-social science integration and modeling
Nathaniel Manning	Ushahidi	Strategy, Collaborative development, Open data
Nithya Ramanathan	UCLA, Nexleaf	Mobile technology for health and environmental impact studies
David Ribes	Georgetown University	Cyberinfrastructure, IT and social organization
Stan Wood	Gates Foundation	Agricultural policy, statistics, data and analysis systems
Kate Schneider	Gates Foundation	Agriculture, policy analysis, program evaluation
Sandy Andelman	Conservation International	Ecosystems, ecology, conservation
Cheryl Palm	Earth Institute, Columbia University	Agriculture, food systems env impact, soils
Bob Scholes	CSIR South Africa	Ecosystem science, GEOSS
Ilya Zaslavsky	UCSD	Cyberinfrastructure, EarthCube, Hydrology
Elena Yulaeva	UCSD	Climate modeling, downscaling
Ravic Nijbroek	Conservation International	Remote sensing, ecosystem analysis
Sara Barbour	Conservation International	Project facilitator

APPENDIX 2: Workshop Agenda

Time	Theme	Lead
9:00 – 9:10	Introduction to VS and the workshop	Sandy
9:10 – 9:20	Participant Introductions	All
9:20 – 9:50	Review of VS threads and user requirements. Expected workshop outcomes	Sandy, Cheryl, Bob
9:50 – 10:10	Information infrastructure blueprint	Ilya
10:10 – 10:30	Q&A on threads and infrastructure blueprint	Sandy, Cheryl, Bob, Ilya
10:30 – 10:45	Coffee	
	25-min discussions. Focus: Designing data workflows, from raw data to indicators to archiving	Sandy, Cheryl, Bob, Ilya
10:45 – 11:10	Organizing statistical and household surveys data in VS	
11:10 – 11:35	Information system support for in situ and ex situ measurements	
11:35 – 12:00	Remote sensing workflows: infrastructure for deriving products, data volumes and update frequencies, retention policies	
12:00 – 12:25	Managing agriculture data, from data capture to indicators	
12:30 - 1:30	Lunch	
	25-min discussions. Focus: model and indicator layers; deployment in Africa	Sandy, Cheryl, Bob, Ilya
1:30 – 1:55	Data collection with mobile devices, validation and crowdsourcing	
1:55– 2:20	Availability and characteristics of models; connecting models with data	
2:20 – 2:45	Indicators and design for decision-makers: towards an effective dashboard and multi-scale multi-modal visualizations	
2:45 – 3:10	Near real time access to data and model computations: operational and cloud deployment issues	
3:10 – 3:25	Coffee	
3:25 – 3:50	Implementation in Africa: specifics, strategies and risks. Systems to leverage	Nathaniel
3:50 – 4:30	Discussion of VS infrastructure, estimates of data flows, construction costs, operations costs	Sandy, Cheryl, Bob
4:30 - 5:00	Discussion of implementation partners, phased implementation, risk mitigation, and next steps	Sandy, Cheryl, Bob

Note: Part of the afternoon discussion (1:30-3:10) was structured around three topics selected from the initial list of 25-min discussions above: 1) Information system support for in situ and ex situ measurements; 2) Organizing statistical and household survey data in VS; 3) Availability and characteristics of models, connecting models with data.