

RESEARCH STRATEGY

A. SIGNIFICANCE

Data integration could alleviate many of the problems facing the biosurveillance and health fields. The process of data integration takes semantically incompatible data, from disparate sources, and in different file types, and merges them into one widely understandable format with metadata to make these data discoverable and available to scientists broadly (1). A successful data integration system must solve multiple problems (1-4) and Mantle will address these key unresolved issues:

Issue 1. Dataset availability: Many datasets are owned by entities who keep them private for reasons related to intellectual property and research concerns, or for security or privacy reasons (5-7). However, even public datasets often lack discoverability, are not stored in a centralized location, and do not have a searchable index of datasets due to the time and difficulty formatting data to the required standards (8). Mantle will incentivize researchers to make datasets available and will link public data in a centralized and searchable database.

Issue 2. Structural heterogeneity: Data are stored in different software formats and structures. Tabular data may be stored in Excel spreadsheets or CSV (or other character-delimited text files), or in a variety of relational database systems. Spatial data may be stored in a tabular format, or may be in a large number of spatial formats or semi structured formats used by various GIS systems and their pre and post processing software (9). Mantle will combine disparate forms of data and will provide homogenous data to users.

Issue 3. Semantic incompatibility: Data, even in the same format, are incompatible in a number of ways. Numerical (continuous or ordinal) variables, like cases of a disease, or temperature, are generally aggregated (summed, averaged) by some period of time and/or some spatial area. The level of aggregation is not (and should not be) standardized, but there is no standard way to refer to the level of organization. Nominal values refer to the same semantic entity (i.e., a particular species) are differently represented in individual dataset schemas and are irreconcilable without mediation. Semantic integration is a challenging and complex problem (4, 10). Broadly speaking, semantic integration involves *mapping* the language in which individual datasets are expressed, the *local schemas*, with a *global schema* (1-4). Two prevailing methods, *global-centric* or *global-as-view* and *source-centric* or *local-as-view*, approach the problem slightly differently, and trade advantages and disadvantages in ease of querying items in the global database, schema flexibility, and others. Mantle will use *global-centric* or *global-as-view*, and *source-centric* or *local-as-view*, to combine previously incompatible datasets.

Issue 4. Ontologies: Ontologies are essential to data integration. They provide a structured, logical definition of a domain's concepts and their relationships (11). In data integration, they serve as mediators, mapping the relationships between heterogeneous representations of concepts in individual datasets (3, 11-14). Ontologies themselves are not standardized. However, methods exist for aligning and evolving ontologies, including the work of Stanford's

CEDAR group, bioontology.org, and Protégé (15-17) . Additionally, the specification of ontologies is codified in a set of web standards centered around linked data, including Resource Documentation Framework (RDF; 18), Web Ontology Language (OWL; 19-20), and others. A new standard, recently approved, standardizes a metadata for tabular data within this framework) (19). As technology coalesces around these standards for ontology specification, and ontologies are developed to codify scientific data collection (12), new types of software are possible, using online, standardized, curated libraries of ontologies to integrate disparate datasets and these ontologies make it possible to query vast amounts of data in a unified interface. Mantle will use machine learning and curated libraries to assign metadata.

Barriers in scientific and institutional culture: Solving the conceptual and technological challenge of matching datasets to ontologies will not itself address the problem of data sharing. Cultural and institutional barriers must be addressed, both programmatically (e.g., by grantors mandating data sharing and metadata annotation policies) and by providing tools and education to make the concept of metadata annotation accessible to scientists (21-22, 16). In one survey of 1329 scientists (23), only 26% were satisfied with tools for metadata preparation. 32% reported dissatisfaction, and 42% neither agreed nor disagreed, indicating “either [that] they truly are indifferent or they are unsure about what metadata means”. The authors speculated the latter is true—46% answered that they do not currently use metadata to describe their dataset. Some of the inertia in the adoption of data-sharing practices can be explained by the fact that large majorities of the scientists surveyed report being satisfied with their data collection, searching, cataloguing, and short-term storage processes. Conversely, it is promising that even larger numbers (78%) report that they would be willing to openly share some of their data in a central repository; fewer report that they would be willing to openly share all of their data.

Identified problems: Data owners have financial, political and other reasons for keeping data private (5). It is not possible to integrate these private datasets with public data, and even for dataset owners integration is still beset by the existing challenges. Also, scientists with domain-specific expertise lack formal training in computer science skills to conduct complex multi-dataset merges (14). Current software solutions do not bridge the skills or understanding gap between scientists and the data problems they must solve. (23-24)

The metadata and data integration problem remains unsolved: Existing software packages do not provide tools to apply metadata to both health and biosurveillance data. Some services allow scientists to upload and store datasets, but treat datasets as monolithic chunks of data. (25). Dryad, for example, is an open-source archive of datasets. These services generally host detailed metadata for each dataset, provide DOI numbers for datasets for publication purposes, and provide a search interface. However, they often only allow the annotation of dataset-level metadata, not allowing the use of ontologies for data integration (e.g., Dryad) or use metadata standards which do not conform to current linked data specifications (e.g., KNB’s Ecological Metadata Language; 26) Both examples given are part of DataONE (27), an NSF-funded collaboration working toward better data practices in science. While these formats provide

valuable services for data portability and sharing, they lack the data integration aspect that is crucial to furthering the objectives of biomedical science.

Previous attempts to create metadata systems encountered lack of awareness and acceptance of metadata standards (23-24). For example, Ecological Metadata Language (EML) is a metadata standard defined as a large XML schema encoding properties of ecological datasets of various types. In 2008, the Long-Term Ecological Research (LTER) program mandated a move to EML for all datasets (24). This move, however, has been notably slow. Scientists involved with its application were interviewed, and they found numerous points of friction (24). Sites were given the standard as an XML specification—250 pages long—and software provided for its application were often incompatible with previously existing systems, and how to reconcile it with other existing metadata systems and data structures was not always clear. Despite its complexity, other researchers found it too limited. Applying EML to existing datasets was a “mostly unfunded mandate”, and had no immediate payoff for scientists; despite voicing support for EML and metadata in general, “...when the rubber hits the road, an unfunded mandate to be altruistic (and simultaneously to lose one’s own tried-and-true local bricolage with data structures) does not prove highly attractive” (24).

Metadata are noted as a “product”, a monolithic structure, which does not account for the iterative and ad-hoc nature of metadata use in day-to-day research (24). Scientists use datasets for different purposes, and in doing so, will describe their structure and composition in discourse; this constitutes “metadata as process”, and is not a single, comprehensive structure. Attempts to impose concrete metadata structures on scientific data of a certain type have, then, met with at best tepid success and “almost-use” (24). For metadata annotation to become common practice among scientists, it must become approachable and offer a value proposition to scientists. Mantle will offer both an understandable, accessible way to annotate datasets, and add facilitate the process of collecting and managing data, so that scientists adopt them as part of their commonplace data workflows.

Broader Impacts: Mantle is currently aimed at biosurveillance and health data. Many components will have potential use beyond biosurveillance, so Mantle will be developed in a generalizable, reusable, and scalable manner. As with any data integration platform, data security must be addressed. As data become more portable, accessible and integrated, systems must be hardened against malicious attacks (28-30) Therefore sensitive data must be safeguarded, including personally identifying information, and security will be incorporated into the design of Mantle from the outset. Databases which incorporate public data must also protect against the injection of false data (31). With proper security measures in place, Mantle will be useful for broad scale ecological and land use data, health data, and human behavior and demographics data. Furthermore, Mantle could serve as a novel way to integrated ecological and social data to improve understanding of how human and natural systems interact to change health outcomes and affect disease emergence.

B. INNOVATION

This project is an innovative fusion of software engineering, data science, and public health research. The incorporation of existing vocabularies and ontologies developed by CEDAR, will help establish universal data standards universally and has not yet been attempted. This approach will allow us to better understand gaps in the creation, and assignment of ontologies, across languages. For about the past decade, data portability and availability has been pursued by scientists and mandated by governments, but has not fundamentally improved (16). Previous studies have found that structural semantic heterogeneity are significant obstacles to overcome when combining data and when assigning metadata (32, 33) and these problems are exacerbated by the lack of formal training in data integration of most scientists in biomedical fields and across academia (14). Mantle directly addresses these technical problems and human deficiencies by automating the metadata application processes where possible and guiding users elsewhere, using machine-learning algorithms trained on existing data and crowdsourced dataset annotations. By automating these processes, we hope that Mantle will be used outside of biomedical research, as it directly addresses a problem that is common to throughout scientific disciplines. For example, many fields of science are becoming data intensive, and thus reliant on cyberinfrastructure. An example is the use of databases as virtual laboratories in astronomy, where an astronomer can make and record a large number of virtual observations (14).

We hope that Mantle will succeed in overcoming current metadata practices by integrating Mantle with an API for data upload and download. This means that other developers can extend the system, perhaps directly uploading datasets from mobile devices or importing directly into an analysis application. We will develop secure mechanisms to obfuscate sensitive data. This will make Mantle compliant with regulations for sharing health data, broadening its set of use-cases. These mechanisms will also enable the use of Mantle with data which cannot be shared for other reasons, and facilitate the *partial* sharing of such datasets, so that they can maximally contribute to other scientific endeavors and the public good. Furthermore, Mantle's metadata assignment features will exist in a user-focused, community-based platform. This will mean that disparate datasets, which are part of Mantle's system and have been assigned ontologies, can be merged and aggregated with unprecedented ease. Past interfaces have not been user-focused, and have relied on scientists with no expertise in metadata application to do the technical data management work. We are designing ways to enable scientists with domain knowledge and no expertise to contribute and to describe their data in flexible ways that make it interoperable with other similar datasets.

C. APPROACH

Share, combine, and export cleaned and joined data: Mantle will use a cloud-based federated database to handle the integration of disparate data storage types. The front-end graphic user interface will provide the portal for users to input, query, and download data. Mantle will use the Resource Description Framework (RDF), which is a standard data interchange model for web-based applications (18). Mantle will use XML syntax to define both the relationship between two entities and the two ends of any link to accommodate evolving data schemas. These Uniform Resources Identifiers (URIs) will allow Mantle to organize and

index data from sources (e.g., tabular data), based on best practices developed by the W3C working group (34). In the case of a CSV file, all of the columns, rows, and cells will be converted to an annotated tabular data model in Mantle. Mantle will be a Linked Data Platform Resource (LDPR) (35), which is an HTTP resource that can be modified and accessed using HTTP code and is managed through a LDP server (35). It will store and retrieve big data using Amazon's NoSQL service, Amazon's Relational Database Service, and MongoDB cluster hosted on ec2 instances. Data sharing *will occur via the* traditional server/client model in Amazon S3 and via an open source peer-to-peer model using Bittorrent protocol and open source compression algorithms to increase speed of transfer from a distributed network of repositories. The entire Mantle environment can be templated into a public Amazon Machine Image for others to use on demand. Amazon provides multiple tools to allow API calls from many different languages, and development environments, thereby enabling scientists from around the world to use Mantle. Overall, these solutions enable researchers to outsource information management and administration and enables them to focus on their primary research internationally.

To enable efficient querying across many linked datasets, Mantle will need to implement a triplestore (36) or leverage an existing implementation of one. Triplestores are designed specifically for the purpose of querying the subject–predicate–object expressions that form the basis of the RDF data model. SPARQL is the most prevalent query language for triplestores (37). Although complicated, it is a very powerful language in its capacity to use many types of data at once. For example, long, trailing queries with many, nested prepositions like the following can be expressed with relative economy: “find all the symptoms of diseases transmitted by bats that migrate to locations with a longitude greater 30 degrees that patient X traveled to in the last 3 months.” Mantle will have in interface that simplifies the process of building these types of complex queries for the user.

Mantle will become a Digital Object Identifier (DOI) provisioner so that hosted datasets can be cited and used in published works. DOIs may include permanent URLs that allow users to locate referenced data. We will likely use DataCite (38) to register DOIs for datasets that are uploaded to Mantle because they use an XML schema. Existing tools like Dryad currently use DataCite for this purpose (38). Data will be stored in the cloud hosted by Amazon Web Services on the GovCloud for enhanced security and privacy. See Figure 1 for a detailed illustration of Mantle's schematic.

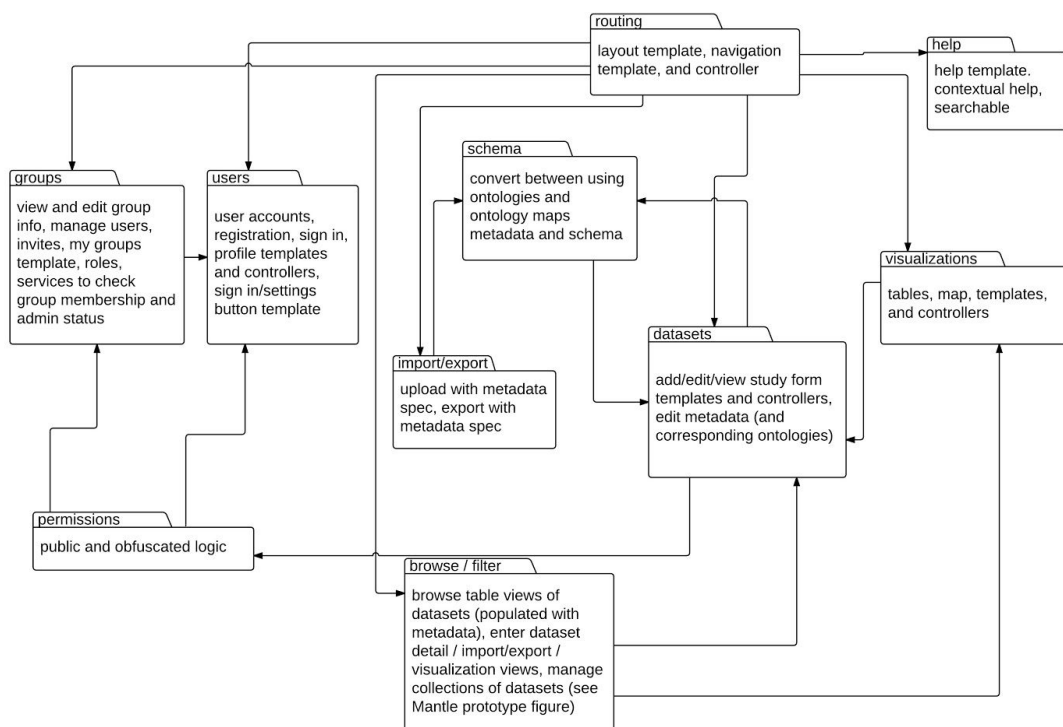


Figure 1. Diagram of components of Mantle's functions, permissions, and data visualizations.

Metadata: Mantle must solve the data integration challenges including dataset availability, structural heterogeneity, and semantic heterogeneity. Mantle will facilitate the collection, storage, versioning, and sharing of data, as well as fitting into existing analytical toolchains, reducing friction in scientific workflows. We will allow scientists to upload datasets in a number of common formats (**Aim 1**). In particular, we will support tabular data in Excel, .csv, and other character-delimited files, spatial raster and vector data in formats supported by the Geospatial Data Abstraction Library (GDAL) software package, and textual data (**Aim 1**). Users will be able to export data tabular and spatial data in supported file formats, and always have access to the original file (**Aim 1**). Uploaded datasets will be converted to Mantle's native data types (WCSV, JSON-LD, and other RDF-compatible formats) (**Aim 1, 3**).

Users will be guided through an interactive process of matching their dataset with cloud-hosted ontologies (**Aim 1**). We will develop and implement machine learning algorithms to suggest ontologies that might fit the contents of the dataset (**Aim 2**). We will present possible ontology matches to users and allow them to select an appropriate one (**Aim 1, 2, 3**). Matches selected by Mantle's users will be used to update machine learning algorithms to improve future guesses (**Aim 2**), and applied to datasets as metadata (**Aim 1**).

Our ontology-matching engine will be developed in collaboration with CEDAR, a BD2K funded project and a leader in ontology applications. We will work with them to interface with their APIs

for searching cloud-hosted libraries of ontologies (37), assist with the curation and creation of biosurveillance-related ontologies in industry-standard formats, and call upon their experience in ontology searching and alignment (39) for the development and training of NLP feature extraction, information retrieval, and machine learning classifiers (**Aim 1, 2**).

Ontology searches will draw on a sequence of techniques. The corpus of curated ontology terms and their synonyms will be used as the basis of textual feature extraction using regular expressions and other text-matching methods. These features will be combined with features of ontologies, such as view counts (39), as predictors for an ensemble of machine learning classifiers, trained on previously-uploaded datasets and optimized using Mechanical Turk or other crowdsourcing methods (**Aim 1, 2**). These techniques will select the most likely ontological match for each entity in a dataset—specifically, data types and the values of nominal variables. Users will be able to easily view and select alternate matches, and these corrections will be used to update classifier training (**Aim 1, 2, 3**).

Overcoming barriers in in scientific and institutional culture: To achieve greater success, Mantle’s goal is to provide an environment for data management that researchers *want* to use. The merit of Mantle is that it will enable users to assign metadata and ontologies to their data seamlessly and enable them to combine with other publicly available data. This will save Mantle’s users significant amounts of time by not having to learn how to clean and structure data to combine with other disparate data.

Security and Privacy: While outside the scope of this proposal, we understand that security is an integral part of systems like Mantle. Systems should enable the core security features of role-based access, passwords, and audit trails.

Data integration: As described above, datasets often suffer from various types of heterogeneity. For example, dataset heterogeneity, especially structural and semantic heterogeneity, differences in file structure, and how concepts are referenced. These are the primary barriers to better data sharing practices across science. Overcoming these barriers can facilitate and incite multidisciplinary research, which is becoming an increasingly important part of the study of human infectious disease, as human health is inextricably bound to ecological, environmental, and animal health factors (40-41). By facilitating the integration of heterogeneous, multidisciplinary data (e.g., infectious disease observations and samples collected in the field, laboratory tests, sociodemographic and economic data, environmental data, data on human and animal population and movement, various other spatiotemporal data types), Mantle will contribute to more effective public health research and disease forecasting.

Focusing on the field of biosurveillance and infectious disease research, we will explore various methods for applying metadata to datasets, in collaboration with leaders in the field of data integration and biosurveillance. Our project will explore cutting-edge methods to facilitate scientists applying metadata annotations, based on a curated selection of structured, web-standard ontologies, to datasets. Technologies for data integration by consistently applying metadata about ontologies will not solve the problem of data heterogeneity unless scientists use it. We are thus developing Mantle, a platform to facilitate the storage, sharing, and analysis of

data, with the capability to apply ontological metadata as part of the capture and upload processes. Mantle will, at each step of the way, use a combination of heuristics and machine-learning algorithms to present users with its best guesses for data type and metadata fields, allow users to correct its guesses, and use corrected data to update its guesses for future computations.

Preliminary data & prototype development: Last year, EcoHealth Alliance was contracted by the U.S. Forest Service to build an infectious disease surveillance system for amphibian diseases. The resultant project was a focused prototype of Mantle that implemented a basic subset the full project's planned functionality, allowing users to upload tabular datasets of amphibian disease, browse similar datasets, and export combined datasets in CSV format. When not logged in, visitors to the site can browse a list of publicly-shared datasets, view them on a map, and export records from merged datasets. When users sign up, they gain the ability to upload datasets to the system. The Mantle prototype only accepts data in one format, a CSV file structured around an ontology created specifically for the project. Users must provide metadata for datasets, including contact information for data owners and sharing preferences (public, private, or public with certain information hidden). This process is made easier by the user filling out a profile, which provides default values for a new dataset's ownership metadata. Logged-in users can browse their own private data, as well as publicly-shared data, and export a merged CSV file combining their datasets with other users' shared data. The prototype does not include a flexible upload or export engine, flexible ontologies for datasets, or machine-learning algorithms for intelligent metadata assignment.

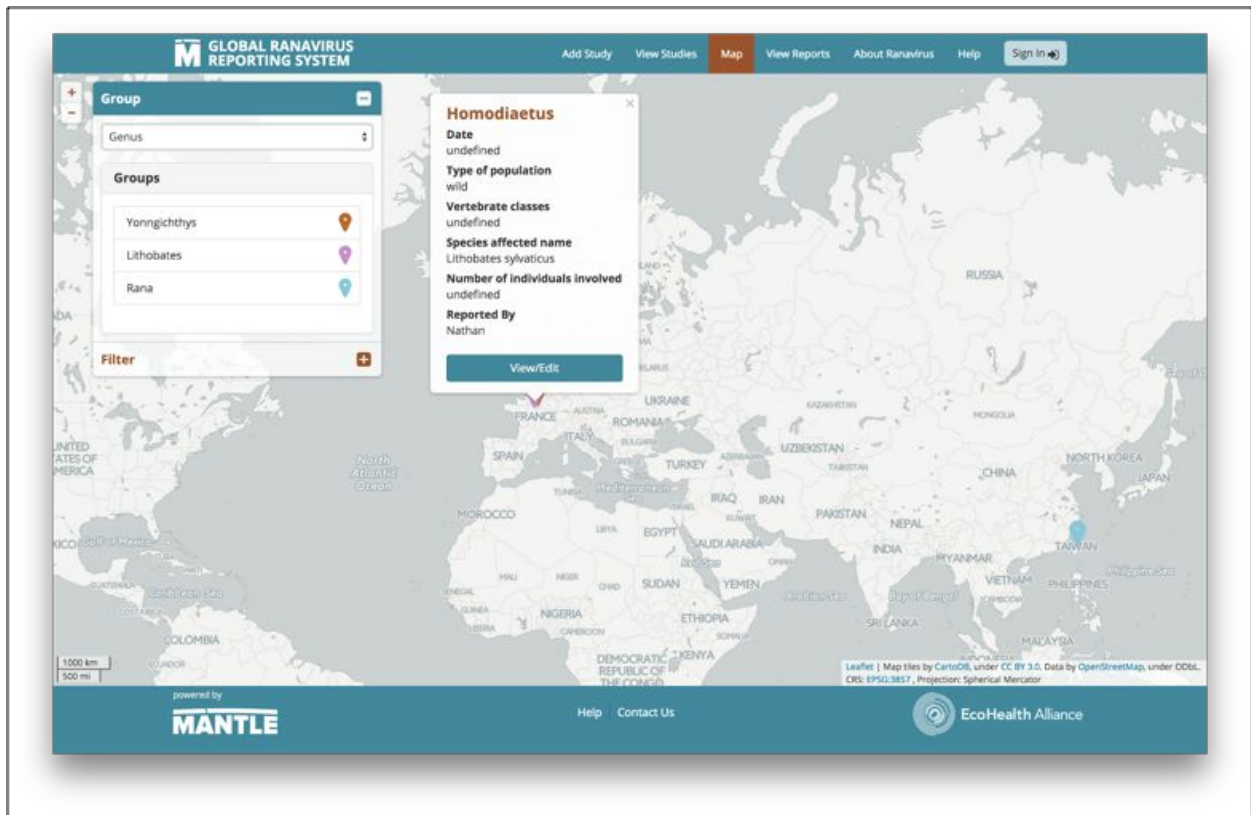


Figure 2. A screenshot of a prototype of Mantle’s graphic user interface and web portal. This view shows the amphibian disease biosurveillance system, a “proof of concept” which allows users to upload disease reports in CSV format, merge tabular datasets using a specific, tailored ontology, and query data spatially (developed for the U.S. Forest Service).

D. TIMELINE & MANAGEMENT PLAN

EcoHealth Alliance will be the primary organization leading this project with Dr. Andrew Huff as the project lead. EcoHealth Alliance leads cutting-edge research into the critical connections between human and wildlife health and delicate ecosystems. Project staff includes software developers, data scientists, public health scientists, clinicians, diagnostic laboratory personnel, veterinarians, information and cyber security experts, and administrative support (Figure 3). Our management plan blends strong scientific expertise in global emerging infectious disease (EID) surveillance, using Agile software methodologies for rapid application development. The project will be managed by our team of data scientists and software developers at EcoHealth Alliance, in consultation with thought leaders in the field of biosurveillance (EHA, ISDS), infused with innovative technologies DIT Inc., developers of leading edge, high quality software, and secured by in the information and cyber security experts at Clango Inc (Figure 4).

		Base Year				Option 1				Option 2			
Mantle Timeline		Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer
Planning	Conduct Mantle kickoff meeting	x	x										
	Daily scrums and planning sessions	x	x	x	x	x	x	x	x	x	x	x	x
	Construct data collection plan	x	x										
	Conduct user advisory group meeting	x		x		x		x		x		x	
	Construct web hosting development plan	x	x										
	Construct research & development plan	x	x										
	Information & cyber security plan	x	x	x									
	Establish data & information management plan	x	x	x									
	Establish marketing and communication strategy	x	x	x									
Implementation	Develop application	x	x	x	x	x	x	x	x	x	x		
	Software testing			x	x	x	x	x	x	x	x	x	
	Field test application			x	x			x	x	x	x		
	Project management meetings	x		x		x		x		x		x	
	Software refinement and stress testing					x	x	x	x	x	x	x	x
	Host a Mantle launch event											x	
	Mantle training and workshops				x		x		x		x		x
	User focus groups and feedback			x	x	x	x	x	x	x	x	x	x
	Marketing and outreach			x	x	x	x	x	x	x	x	x	x
Evaluation and Deliverables	Prepare peer-reviewed journal articles				x			x	x			x	x
	Attend regional, national, and international conferences to present results		x			x	x			x	x		
	Financial reports	x	x	x	x	x	x	x	x	x	x	x	x
	Copyright (no patent required as open source open access application)				x				x				x
	Progress & technical reports		x				x				x		
	Property reports				x				x				x
	Analyze user traffic, user accounts, and data uploads									x	x	x	x
	Annual audit			x				x				x	

Figure 3. Planning, implementation, project evaluation, and deliverables for Mantle.

Clango will be responsible for the security aspects of Mantle. Clango has 15 years experience in identity and access management, anti-fraud solutions, governance, and advisory services worldwide. Clango has deployed numerous identity and access management (IAM) capabilities like user registration and lifecycle management; adaptive and federated authentication; privileged administration and access governance. Clango has worked in finance, healthcare, higher education, and across federal, state, and local governments. Clango assures that only authorized identities have access to the right data at the right time.

The **International Society for Disease Surveillance (ISDS)**, works to improve population health by advancing the science and practice of surveillance to support timely and effective prevention and response. The International Society for Disease Surveillance (ISDS) is a 501(c) 3 nonprofit organization founded in 2005 and dedicated to the improvement of population health by advancing the science and practice of disease surveillance. ISDS' membership represents professional and academic subject matter experts in the fields of public health

surveillance, clinical practice, health informatics, health policy, and other areas related to national and global health surveillance. ISDS works toward a vision of timely, effective, and coordinated disease prevention and response among a skilled public health workforce through programs that position us at the vanguard of the disease surveillance field.

	EcoHealth Alliance							ISDS	Clango			
AREAS OF EXPERTISE	Andrew Huff	Peter Daszak	Nathan Breit	Brock Arnold	Jonathan Gooley	Noam Ross	Frederico Rosario	Toph Allen	Daniel Sullivan	Laura Streichert	Anun Kothamath	Steven Hawkins
Emerging Infectious Disease Modeling												
Data Management	X	X	X	X		X	X	X		X	X	X
Informatics	X	X	X	X		X	X	X		X	X	X
Modeling and Simulation	X	X	X	X		X	X	X		X	X	X
Visualization	X		X	X	X	X		X		X		
Software Development												
Applications Development	X		X	X	X	X	X	X			X	X
Functionality & Design	X		X	X	X	X	X	X			X	X
Mobile Applications			X	X	X		X	X			X	X
Scientific Development	X	X	X	X	X	X	X	X			X	X
Systems Development	X	X	X	X	X	X	X	X			X	X
Testing & Automation	X		X	X	X	X	X	X			X	X
Cyber & Information Security												
Application and Systems Development Security	X		X	X			X		X		X	X
Business Continuity and Disaster Recovery Planning	X						X				X	X
Cryptography			X								X	X
High Availability Systems	X						X		X		X	X
Identity and Access Management / Control			X								X	X
Laws, Investigation, and Ethics	X										X	X
Physical Security	X						X				X	X
Security Management Practices	X		X				X				X	X
Telecommunications and Networking Security			X				X				X	X
One Health												
Bacteriology		X								X		
Biology		X				X						
Biostatistics	X	X				X		X		X		
Biosurveillance	X	X		X		X		X		X		
Clinical Laboratory Science		X				X				X		
Environmental Health Science	X	X				X		X				
Epidemiology	X	X						X		X		
Field Surveillance		X						X		X		
Food Systems	X											
Health Systems Research	X	X						X		X		
Infectious Disease Ecology	X	X				X		X				
Medicine		X								X		
Parasitology		X								X		
Plant Pathology		X				X						
Policy	X	X								X		
Veterinary Medicine		X										
Virology		X								X		
Zoology / Wildlife		X										
Other												
Cartography	X					X		X				
Data Science	X		X	X		X	X	X			X	X
Geographic Information Systems	X					X	X	X				X
Governance		X								X	X	X
Implementation Science	X		X	X		X	X	X			X	X
Legal and Regulatory	X	X								X	X	X
Linguistics / Ontology	X		X	X		X		X	X	X	X	X
Spatial Databases	X		X	X	X	X	X	X			X	X
Statistics	X			X		X		X				
Social Science												
Anthropology												
Psychology	X											
Sociology								X				

Figure 4. Areas of expertise and capacity of staff.