



**BAA #: HDTRA1-14-CHEM-BIO-BAA**

## **Proposal Title: Mantle: An Internet App. for the Fusion of Global Biosurveillance Big Data**

**Objective:** To develop a multilingual, open source, and open access software application that combines disparate biosurveillance and infectious disease data from multiple foreign governments and international academic sources.

### **Description of Effort:**

- Create a secure identity and access management system for users' data
- Create a an Internet based system to combine disparate data sources in multiple languages
- APIs to ingest the hundreds of existing biosurveillance systems into a single One Health data platform
- Obfuscating human health and agricultural data, from research or surveillance streams or field based research, to maintain regulatory compliance and privacy when necessary (e.g., HIPAA, SOX, etc.)
- Provide a user friendly and simple interface for complex One Health data
- Provide mobile apps for One Health data collection in the field in multiple languages
- Develop a centralized secure web portal for users to upload and store infectious disease datasets, collect raw data, and share datasets with other scientists
- Create an application to automatically clean, sort, and combine biosurveillance data, in multiple languages, with minimal user input and effort

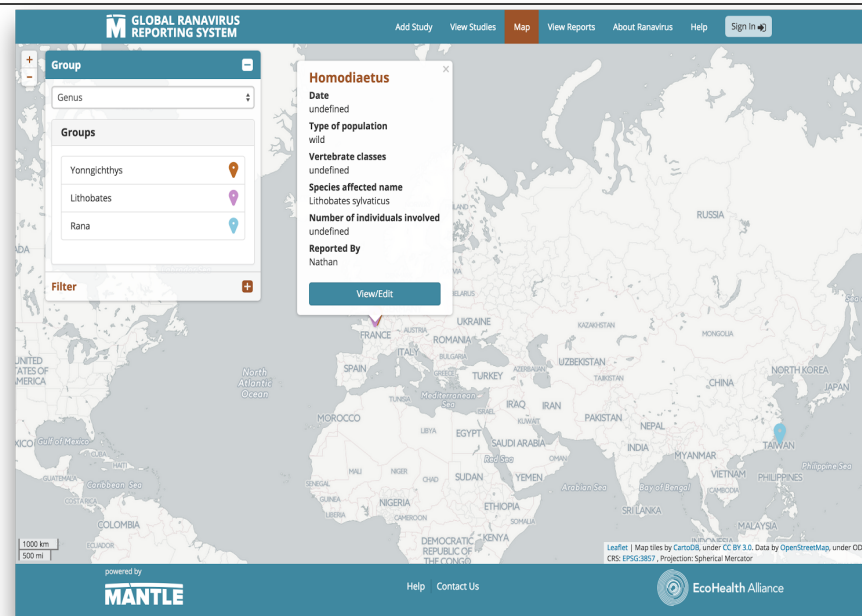
**Benefits of proposed technology:** Mantle will provide the technology to collect and combine biosurveillance data from multiple governments, in multiple languages, that will result in a measurable impact on the bio-event timeline.

**Potential Challenges:** Inaccurate data uploaded by users.

**Maturity of technology:** TRL 3 & MRL 3  
(this project uses and combines existing technologies and R&D)

**Technology goal:** TRL 8 & MRL 5

**Topic:** CBA-06



**Figure 1.** A screen shot of a prototype. Users will be able to select biosurveillance data sets by a combination of location, area, or disease type.

### **Major goals :**

- Establish and promote universal biosurveillance data standards
- Develop a novel open-source and open access (free to all users) cloud-based One Health biosurveillance data fusion platform
- Develop a friendly and efficient user interface for BSVE and international users to interact with worldwide One Health biosurveillance data and predictive models developed by EHA

**Cost:** \$2,296,336.59

**Period of Performance:** 2 years

**Contact Information:** Dr. Andrew Huff Ph# 1.612.743.1265  
huff@ecohealthalliance.org

## PROJECT SUMMARY

The goal of Mantle is to create an Internet app that enables non-technical scientists and governments to easily and efficiently apply metadata to biosurveillance data. Mantle is a free and open-source research and software development project, developed under the Apache License 2.0, dockerized, using the BSVE's SDK. Mantle will be an open-source web platform designed for the curation, integration, and sharing of global public health related data. By incorporating previously developed metadata standards and public health ontologies, and automating their application, Mantle will be designed to meet the data needs of a wide variety of domestic and international public health users. Public health professionals, in the office, field, or the lab, will be able to upload a wide variety of unstructured and structured datasets to Mantle in a variety of commonly used formats and languages. Additionally, Mantle will have an API that can be used to ingest and combine existing and continuously collected biosurveillance data in near real time.

Mantle's users can belong to governments, private organizations, and individuals, and individual datasets can be grouped together into larger projects, all with group-level access permissions. These features enable public health professionals to collaborate across geographic, institutional, and disciplinary boundaries to accomplish large-scale data collection efforts not otherwise possible. Mantle will also include a number of open-access datasets from EcoHealth Alliance's partners, and biosurveillance data, that with Mantle, will be openly available for users to combine with their own data or content.

Users of Mantle will be able to set fine-grained sharing and privacy controls on uploaded datasets to share or protect their data, and industry best practices will be employed to protect all data uploaded to Mantle. Once users create and sign into their user account, Mantle users will be able to examine publicly available and obfuscated datasets (to protect privacy) in a number of views appropriate to their content, including tables, maps, and charts. Additionally, Mantle will display datasets from different data sources alongside one another and save and export combined datasets. Users with export privileges will be able to download data in a number of formats for use with external software (e.g., .xlsx, .csv, .txt, .shp, .shx, .dbf, etc.).

Mantle will uniquely provide free access to high fidelity infectious disease data, which will help enable scientists, public health practitioners, and policymakers to tackle the world's biggest infectious disease threats. Furthermore, Mantle will enable identification and faster response to infectious disease threats as data can be continuously uploaded, validated, and contextualized via Mantle's API, rather than waiting for data to be collected and integrated after infectious disease threats are identified via traditional biosurveillance mechanisms (by significantly reducing data communication, data cleaning, and language translation time). Open access health data and open source biosurveillance software will help infectious disease and biosurveillance research advance, and Mantle will fill a critical gap in emerging infectious disease knowledge and infectious disease preparedness. Mantle will generalize across scientific fields as more big data ontologies are created, and will be able to be used broadly.

## 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 biosurveillance data issues:

**Issue 1. Dataset availability:** Many datasets are owned by entities that 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. 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 on 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), 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. Data owners have financial, political, and other reasons for keeping biosurveillance data private (5). 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 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). Software provided for EML 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. 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 facilitate the process of collecting and managing data, so that scientists adopt them as part of their commonplace data workflows. 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.

**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 that 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 integrate ecological and social data to improve understanding of how human and natural systems interact to change health outcomes and affect disease emergence.

Mantle is an innovative fusion of software engineering, data science, and public health research. 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. 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. Furthermore, Mantle's metadata assignment features will exist in a user-focused, community-based platform, enabling 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.

## REFERENCES CITED

1. M. Lenzerini, Data integration: a theoretical perspective, in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*. 14 (2002).
2. A.P. Sheth, Changing focus on interoperability in information systems: From system, syntax, structure to semantics, in *Interoperating Geographic Information Systems* M. Goodchild, M. Egenhofer, R. Fegeas, C. Kottman, Ed. (Springer, Boston, Massachusetts, 1999) pp. 5–29.
3. I. F. Cruz, H. Xiao, The role of ontologies in data integration. *Engineering intelligent systems for electrical engineering and communications*. 13.4, 245 (2005)
4. A. Cali, D. Calvanese, G. De Giacomo, M. Lenzerini, Data integration under integrity constraints. *Information Systems*. 29(2), 147-163 (2004).
5. J. Polonetsky, O. Tene, Privacy and Big Data: Making Ends Meet, in *Big Data and Privacy: Making Ends Meet - Future of Privacy Forum and Stanford Law School The Center for Internet and Society*. pp. 1-6 (2013).
6. M. Birnhack, Big Data as A New Informational Privacy Paradigm, in *Big Data and Privacy: Making Ends Meet - Future of Privacy Forum and Stanford Law School The Center for Internet and Society*. pp. 7-10 (2013).
7. J. Brookman, G. S. Hans, Surveillance as a De Facto Privacy Harm, in *Big Data and Privacy: Making Ends Meet - Future of Privacy Forum and Stanford Law School The Center for Internet and Society*. pp. 11-14 (2013).
8. K. Braunschweig, J. Eberius, M. Thiele, W. Lehner, The state of open data, in *WWW2012, Lyon, France: ACM*. (2012).
9. Warmerdam, F. (2008). The geospatial data abstraction library. In *Open Source Approaches in Spatial Data Handling* (pp. 87-104). Springer Berlin Heidelberg.
10. M. Uschold, M. Gruninger, Ontologies and semantics for seamless connectivity, in *ACM SIGMod Record*. 33(4), 58-64 (2004).
11. J. S. Madin, S. Bowers, M. P. Schildhauer, M. B. Jones, Advancing ecological research with ontologies. *Trends Ecol Evol (Amst)*. 23(3), 159–68 (2008).
12. J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, *et al.* An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*. 2, 279-296 (2007).
13. N. Schuurman, A. Leszczynski, Ontology□Based Metadata. *Transactions in GIS*. 10(5), 709–26 (2006).
14. T. Arrison, S. Weidman, Ed. *Steps Toward Large-Scale Data Integration in the Sciences: Summary of a Workshop* (National Academies Press, Washington, DC, 2010).
15. P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, *et al.* BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res*. 39(Web Server issue), 541-5 (2011)
16. M.A. Musen, C. A. Bean, K-H. Cheung, M. Dumontier, K. A. Durante, *et al.* The center for expanded data annotation and retrieval. *J Am Med Inform Assoc*. (2015)
17. T. Tudorache, S. Falconer, N. F. Noy, C. Nyulas, T. B. Üstün TB, *et al.* Ontology Development for the Masses: Creating ICD-11 in WebProtégé, in *Knowledge Engineering and Management by the Masses* (Springer, Berlin Heidelberg, 2010) pp. 74–89.
18. K. S. Candan, H. Liu, R. Suvana, Resource description framework: metadata and its applications. *ACM SIGKDD Explorations Newsletter*. 3(1), 6-9 (2001).
19. M. J. O'Connor, C. Halaschek-Wiener, M. A. Musen, Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL, in *The Semantic Web – ISWC 2010* P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, *et al.* Ed. (Springer, Berlin Heidelberg, 2010) pp. 194–208.
20. G. Antoniou, E. Franconi, F. V. Harmelen, Introduction to Semantic Web Ontology Languages, in *Reasoning web* N. Eisinger, J. Maluszynski, Ed. (Springer, Berlin Heidelberg, 2005).

21. B. Notay, Improved Metadata for tracking Research Output across the fields. *Euroscientist*. (2013)
22. P. Diviacco, P. Fox, C. Pshenichy, A. Leadbetter, *Collaborative Knowledge in Scientific Research Networks* (IGI Global, Hershey, Pennsylvania, 2015).
23. C. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, *et al.* Data sharing by scientists: practices and perceptions. *PLoS ONE*. 6(6) (2011).
24. P. N. Edwards, M. S. Mayernik, A. L. Batcheller, G. C. Bowker, C. L. Borgman, Science friction: data, metadata, and collaboration. *Social Studies of Science*. 41(5), 667-90 (2011).
25. J. Greenberg, Theoretical Considerations of Lifecycle Modeling: An Analysis of the Dryad Repository Demonstrating Automatic Metadata Propagation, Inheritance, and Value System Adoption. *Cataloging & Classification Quarterly. Metadata and Open Access Repositories*. 47(3-4), 380-402 (2009).
26. E. H. Fegraus, S. Andelman, M. B. Jones, M. Schildhauer, Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull. Ecol. Soc. Am.* 86(3), 158–168 (2005).
27. W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse, G. Janee, DataOne: Data Observation Network for Earth - Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine*. 17(1/2), (2011).
28. C. Tankard, Big Data Security. *Network Security*. 2012(7), 5-8 (2012).
29. B. Brown, M. Chui, J. Manyika, Are you ready for the era of “big data”? *McKinsley Quarterly*. 4, 24-35 (2011).
30. J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, *et al.* The Mobile Data Challenge: Big Data for Mobile Computing Research. *Pervasive Computing*. (2012).
31. S. Rosset, E. Aharoni, H. Neuvirth, Novel Statistical Tools for Management of Public Databases Facilitate Community-Wide Replicability and Control of False Discovery. *Genetic epidemiology*. 38.5, 477-481 (2014).
32. L. Seligman, A. Rosenthal, A metadata resource to promote data integration, in *Proceedings of the IEEE Metadata Conference, Silver Spring*. (1996).
33. A. Rosenthal, E. Sciore, S. Renner, Toward unified metadata for the department of defense, in *IEEE Metadata Workshop*. (1997).