



CALIFORNIA’S PLAN FOR STAGED DEVELOPMENT OF A STANDARDS-BASED EPHTN

CONTENTS

1. INTRODUCTION	2
1.1. Plan Purpose and Scope	2
1.2. Network Vision	3
1.3. Plan Audience	3
1.4. Plan Organization	3
1.5. Plan Evolution	3
2. NETWORK FUNCTIONS	4
2.1. Overview	4
2.2. Contribute to Nationally Consistent Datasets	5
2.3. Enhance Data for Spatial Integration	7
2.4. Describe and Discover Data and Services	11
2.5. Integrate Data	13
2.6. Exchange and Visualize Data	17
3. NETWORK TECHNOLOGY	22
3.1. Overview	22
3.2. Data Enhancement Services	22
3.3. Text-Based Linkage Services	27
3.4. Spatiotemporal Linkage Services	31
3.5. Exchange Services	34
3.6. Visualization Services	36
3.7. Metadata Services	40
4. REFERENCES	43
5. APPENDIX - AUTOMATED SPATIOTEMPORAL LINKAGE: VISION, REQUIREMENTS, AND ARCHITECTURE (VERSION 3)	44

1. INTRODUCTION

Recipient activity (j) from CDC PA-02179:

Develop a plan for staged development of a standards-based environmental public health tracking (surveillance) network that allows direct electronic data reporting and linkage within and across health effect, exposure, and hazard data and can interoperate with other public health systems. It is expected that the architecture and information technology functions and specifications used for enhancing existing data systems and developing an overall plan for the Environmental Public Health Tracking (surveillance) Network will be compatible with those being developed under other programs such as NEDSS, Bioterrorism, and EPA's National Environmental Information Exchange Network. (See Appendix II for the Internet addresses and a list of state NEDSS coordinators). The Environmental Public Health Tracking (surveillance) Network should be based on specifications and an environmental public health tracking logic model(s) as addressed in Activity m. These will follow data and technical specifications derived from industry standards for data types, code sets and vocabularies, messages for data exchange, and technology systems standards as available.

1.1. Plan Purpose and Scope

The purpose of this document is to describe the functional requirements of an Environmental Public Health Tracking Network (EPHTN), an architectural framework for addressing those requirements, and development and deployment pathways for implementing a standards-based Network. This plan addresses primarily the technical information systems issues surrounding the development and adoption of standards-based network architecture. Though these issues are dependent on content- and stakeholder-specific domains, this plan treats environmental health content areas generically and focuses heavily on coordination interactions with primary reporting and surveillance systems. Consumers of environmental health tracking information and services are also treated as broad stakeholders of Tracking information and services. This plan does not specifically address or analyze epidemiologic issues of scientific validity or feasibility in the linkage or integration of environmental health data.

1.2. Network Vision

The California Environmental Health Tracking Program (CEHTP) architectural vision assumes that control and management of access to environmental health surveillance data lies with the system owners. The CEHTP encourages, communicates requirements, and assists system owners in securing resources for adopting secure, interoperable, electronic-based information sharing architectures as prescribed by CDC's Public Health Information Network (PHIN) and US EPA's National Environmental Information Exchange Network (NEIEN) federal initiatives. The CEHTP serves as a service provider for assisting system owners in enhancing their data systems and content in ways that ensure maximum levels of electronic interoperability and integration. The CEHTP serves as a standards, specifications, and technology developing partner for record- and systems-level integration.

1.3. Plan Audience

The target audiences of this plan are CDC NEPHTN staff and other informatics professionals, CEHTP staff, CEHTP system owner partners, other CEHTP stakeholders with IT or GIS expertise, and other State Tracking grantee technical experts.

1.4. Plan Organization

This plan has two sections: Network Functions and Network Technology. Network functions are the key potential uses of the Network. Each function addresses the high-level requirements of the Network, both in terms of short- and long-term implementation needs as well as related technology, content, and coordination needs. The Network Technology section examines the technologies that are developed for the Network. Each technology fulfills network functions in different ways. Technologies will be addresses in terms of their benefits, objectives, architecture, and the plan with which they will be developed and deployed.

1.5. Plan Evolution

This plan is meant to be an evolving document. Requirements will change in concert with discoveries of new technologies and as content on the network becomes more accurate and complete. Individual aspects of this plan should be revisited on a regular basis to address and document actual shifts or the need to modify requirements, standards, specifications, methodologies, and technologies.

2. NETWORK FUNCTIONS

2.1. Overview

Network functions are the key potential uses of an implemented Tracking Network in California. These uses were gathered from requirements identified in two Tracking pilot projects, technical needs assessment activities, stakeholder needs assessment activities, industry trends, and various technical discussions at the local, state (inter- and intra-departmental), federal, and private levels. Some of these functions are intended to address short-term requirements, while many incrementally contribute to a long-term, sustainable infrastructure.

There are five identified functions for Tracking Network implementation: 1. Contribute to nationally consistent datasets, 2. Enhance data for integration, 3. Describe and discover data, 4. Integrate data, and 5. Disseminate and visualize data.

Though not all of these functions will be implemented simultaneously, it is expected that each function will be implemented successively to bring each new prioritized environmental or health outcome dataset “online”, or made available for consumption over the Network. Indeed, the availability of environmental health content is dictated by the concurrent implementation of these functions in the context of the content they serve. Standards for functionality and requirements specifications will weigh heavily prior to any implemented function, however, it is expected that, as each environmental or health dataset is brought online to the Tracking Network, evolution of overall function requirements and subsequent elaboration will lead to the need for updates to prior content/functionality implementation iterations.

The implementation phase will largely mirror the model used for bringing a dataset online during the pilot phase. The most notable distinction between pilot and implementation phase is that the contribution of a new online dataset to a nationally consistent dataset is a foremost priority.

2.2. Contribute to Nationally Consistent Datasets

2.2.A. Definition

This network function aims to extend the benefits of California’s environmental health tracking content at a national level. As a result of implementation activities, California will be able to respond to any State or CDC request for available and de-identified environmental hazard, health event, exposure, or pre-linked environmental health integrated data. It is assumed that identified data elements and structure will follow standards of consistency and quality as determined by CDC in cooperation with the Standards and Network Development Workgroup.

The need for compiling nationally consistent datasets is of particular interest to California, because it demonstrates the value of tracking in a holistic sense. The benefit of being able to visualize, disseminate, generate hypotheses, or identify patterns of environmental health association in multi-state or national datasets is undeniable.

2.2.B. Short-Term Needs

The distinction between short- and long-term needs is largely one of transitioning from an ad-hoc, manual, human-dependent model of producing and exchanging datasets to an ongoing, automated, standards- and machine-based model. Preliminary iterations of contributing to larger domain (multi-state or national) datasets will emphasize the preparation and processing of datasets to match consumer requirements. For example, the logic for transforming health or hazard event data into summary aggregations/rates is captured for desired geographic boundaries (census, city, county, etc).

2.2.C. Long-Term Needs

In the latter phases of implementation, needs surrounding the production of national consistent datasets will emphasize standards for data elements and structure, request/response templates, and real-time messaging.

2.2.D. Content Needs

During implementation, CEHTP will establish the ability to consume de-identified tabular- or visualization-based environmental or health data. De-identified tabular datasets are sufficiently aggregated to mask individual-level identifiers. Requests for health events at the individual level will be completely stripped of individual identifiers. Depending on availability and patient confidentiality concerns, individual-level social status variables (age, race, sex, etc) will be included in extract datasets. Visualization-based data are processed datasets for viewing in a GIS environment with continuous values, such as density estimated raster surfaces.

CEHTP will contribute spatiotemporally linked environmental health datasets to multi-state and national compilations. CEHTP will be prepared to handle requests that integrate environmental health events at scales for which data is collected or at any other desired aggregations. It is expected that requests will typically follow a health-centric integration methodology where environmental hazard metrics are aggregated about health events in space and time. For example, ozone and lung cancer linked compilations for a given time period might be marked up from a request that stipulates population rates of lung cancer and average annual ozone concentrations by census block group. Equally, CEHTP will position itself to handle hazard-centric integration requests that aggregate health event metrics around environmental hazard events.

2.2.E. Technology Needs

Basic technology needs in early implementation will emphasize relational database management systems for storing, transforming, and extracting datasets. Traditional desktop GIS applications will be used for producing visualization and spatiotemporal integration products. In early stages of implementation, more traditional mechanisms for transfer will be favored, such as [Secure] File Transfer Protocol, removable media, HTTP download, etc.

In the latter stages of implementation, technology needs will emphasize automated compilations and real-time request/response flows. Message-, integration- and service-based technological tools are required to handle these needs. Network exchange tools developed for the EPA Exchange Network and the CDC Public Health Information Network are likely candidates for servicing message transfer requests.

Flows of geographic data (image-based services or feature-based services) require technological mechanisms that follow vendor-neutral functionality standards and take advantage of web-based protocols. The Open Geospatial Consortium provides two known geospatial streaming standards that the CEHTP will satisfy: Web Map Service (WMS) and Web Feature Service (WFS).

New technological tools are also required to service real-time spatiotemporal integration requests. These tools must provide GIS functionality in an enterprise context, allowing for spatial

analytical procedures (e.g. buffering, distance calculations, intersecting, etc.) in real-time for various input geometry types. Spatiotemporal integration often occurs at scales where individual identifiers are used, like a residential address. These tools must be evocable from a public domain, while the consumable products of the process are de-identified and at the same time derived from confidential inputs that are processed real-time in a secure domain. For more vision, requirements, and architecture discussion surrounding spatiotemporal linkage, see §3.4 and the Appendix.

2.2.F. Coordination Needs

Coordination with CDC and SND during the initial phases of implementation will establish standards for data elements, structure, and functionality. Through iterative development and elaboration, coordination will lead to automated interaction using exchange- and service-based technologies to tie together disparate content and functionality.

2.3. Enhance Data for Spatial Integration

2.3.A. Definition

The results of CEHTP Future Assessment and pilot project planning activities determined at an early stage that the content of environmental health monitoring systems exist at varying degrees of spatial accuracy and resolution. In order to obtain the most scientifically defensible environmental health spatial integration, the spatial component of environmental and health events must be captured at the most accurate and highest resolution possible. Typically, assuring a high quality spatial component for event data is not a mandated activity for surveillance systems, nor is it a funded activity. In many cases, the spatial component of event data is not a reportable data element, in which the flow of its true scale and extent ends at the local level, when the event is recorded and/or key entered. Because CEHTP can provide assistance and services in working towards capturing more accurate spatial event data, CEHTP regards this function as primary utility to implementation of the Tracking Network.

2.3.B. Short-Term Needs

Considerable effort has been expended by CEHTP staff in piloting an enterprise geocoding web service that incorporates multiple reference datasets and CASS¹-certified address standardization/verification tools. Multiple programs within the California Department of Health Services and California Environmental Protection Agency have expressed interest in utilizing its real-time capabilities, while others have already begun consuming its value-added functionality. CEHTP will continue to refine, optimize, and maintain the centralized geocoding system with the concurrent goal of situating at least one instance of the system in an administrative home where there is an abundance of in-house application development capacity, cross-program visibility, and track record in information systems deployments. CEHTP is offering to provide the service's source code to a program meeting these criteria using an Open Source licensing model.

California State and Federal information technology planning organizations have recently identified centralized geocoding as a major interoperability priority. CDC/NEDSS recently entered a contract with GroupOne to provide geocoding services through the GeoStan commercial product. To keep the issue robust, the CEHTP implementation plan will prioritize the investigation of science-based evidence for arguing that real-time centralized geocoding is a key hurdle to feasibly integrating multiple stovepipe surveillance systems. This necessitates a small study or studies that compare traditionally reported surveillance data to that which is enhanced through real-time geocoding.

Whether an environmental health event can be described by a point, line, or area, CEHTP will move quickly to demonstrate to partnering surveillance systems the mutual value in collecting accurate geographic identifiers. One way that CEHTP learned to demonstrate this was to make the dataset quickly available for publicly visualizing in various dissemination mediums. If each of the visualization platforms is easy to navigate, while, at the same time, clearly describes the limitations of the inferior geographic data alongside more accurate data, data stewards are given tangible reasons to work towards improving the geographic accuracy of their data holdings. This methodology showed promise over the last 3 years of the planning phases and will be given priority during the early phases of implementation.

¹ Coding Accuracy Support System (CASS). CASS is a program through which the United States Postal Services approves software vendors and other information service providers to provide certified ZIP+4 and address corrections services to the public.

2.3.C. Long-Term Needs

CEHTP's ultimate goal for enterprise geocoding is that all health surveillance data be attributed with the most accurate spatial representation possible. Labs, hospitals, clinics, schools, or any other location where health surveillance data originates are all possible candidate clients for incorporating centralized geocoding into their reporting architectures. As geocoding reference data is improved by vendors, CEHTP will upgrade and maintain geocoding services to assist stakeholders in shifting focus from street centerline-derived geocodes to parcel-based and building footprint-based extractions. Also, in patient encounter situations, in which a pair of eyes is trained on a computer screen, map services with high resolution aerial and satellite imagery can be utilized to provide heads-up digitization to extract GPS-scale coordinates. CEHTP has already proven that the content and technology are robust enough to accomplish this.

The spatial representation of some environmental health data cannot always be derived or extracted from reference datasets in an enterprise geocoding engine. Particularly with some environmental hazard datasets, such as pesticide use, drinking water, and traffic volumes over road networks, the enhancement of the spatial component requires more novel solutions. The CEHTP needs to align itself to form collaborative partnerships with surveillance systems to assist them in extracting more accurate geography over time. Likely amounting to small-scale IT and GIS infrastructure projects, CEHTP and data system owners will enter joint application development projects to extend existing reporting architectures to incorporate the extraction of higher quality surveillance event geography. There is also a functionality need for providing a feature input service that allows surveillance systems and their local data providers to centralize the activity of importing (from multiple spatial data formats), creating, editing, and exporting spatial feature data. For example, in order to obtain higher resolution pesticide event data, the Department of Pesticide Regulation needs pesticide operators to digitize field boundaries into a central reporting system. This problem is complicated by the fact that some County Agricultural Commissions already maintain their own local field boundary databases. A service that can accept input of both existing feature data of multiple formats and newly inputted data from various client interfaces requires complex solutions. CEHTP refers to this functional requirement as "Enterprise Feature Input" (see §3.2).

2.3.D. Content Needs

Surveillance system owners that utilize spatial enhancement services provided by CEHTP also need to incorporate available content standards for addresses, derived point location coordinates, or other extracted geometries into their existing data models. Providing assistance in this regard will be a priority activity for CEHTP during implementation phases. CEHTP will draw from standards identified in the HL7 Reference Information Model (RIM) as well as functionality-based standards provided from the Open Geospatial Consortium.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

As long as CEHTP remains the administrator of centralized geocoding within the Department of Health Services, CEHTP will also have to maintain sufficient licensing and storage of all geocoding reference datasets and address standardization packages. CEHTP will also bear the cost of maintaining licenses to proprietary mapping services (satellite, aerial, and other basemap imagery used for orientation) used in heads-up digitization implementations.

2.3.E. Technology Needs

In moving into implementation phases with enterprise geocoding and to increase the usability in a spatial enhancement context, a number of technological improvements are necessary to optimize computational speed over the network and to increase spatial accuracy and completeness. Although SOAP-based web services are highly interoperable, there is still a need to simplify client connectivity and functionality in various milieus, particularly in a web browser setting. For heads-up digitization, there is increasing need to provide a technological framework for assembling multiple disparate map image and feature services in single “meta-services”, with the services’ results themselves stemming from standards-based requests. More detail about technological needs, development plans, and deployment plans for enhancement and visualization/dissemination services can be found in §3.2 and §3.6.

2.3.F. Coordination Needs

All of the service-based enhancement initiatives that CEHTP proposes will require marketing and outreach for prospective clients and education and training for consuming clients. Value-added services provided by CEHTP need to be well-documented on an easily accessible website and with the ability to download code, understand code, and discuss how to use it with other users.

Partnerships formed for the purposes of improving the spatial component of non-address geocodable surveillance data, will require novel coordination techniques. Cooperative agreement proposals and incentive-based tool frameworks are two examples of coordination strategies that have shown promise during the planning stages.

2.4. Describe and Discover Data and Services

2.4.A. Definition

This function provides a common framework for allowing Tracking Network users to learn about or be made aware of the existence and quality of environmental health data and services. The implication of this function is that humans or machines can issue standard requests for information about data and services and receive standard responses.

2.4.B. Short-Term Needs

At the onset of implementation, the CEHTP will place particular emphasis on inputting metadata for about 40 California environmental health surveillance datasets identified in the first phase of the Future Assessment activities. These metadata will be inputted according to the template specifications identified by the Metadata Workgroup of SND.

For those identified CEHTP services that already implement standards with global consensus, like Web Map Service and Web Feature Service, CEHTP will input metadata for all mandatory parameters. In the case of WMS and WFS, the XML GetCapabilities document must be filled and available for download under the service endpoint.

Metadata standards must be identified and consensus reached for those services that also require consensus on interface specifications, like enterprise geocoding and spatial linkage. This might amount to establishing a consensus on a formal structure for GetCapabilities documents.

2.4.C. Long-Term Needs

Special attention will be placed on identification of extracts and/or structural attributes within a single system owner's data model that constitute an environmental health "event". These elements and structures comprise the time, place, and any other attributes that influence a health outcome, exposure, or hazard. To consistently document as metadata the subsetting, aggregating, merging, or transforming steps necessary to produce an environmental or health event is an

efficiency benefit to Tracking stakeholders and is a necessary component for automated spatiotemporal integration.

At the latter stages of implementation, CEHTP shall input metadata for all services whose metadata structure was determined earlier in implementation. This could constitute the entry of GetCapabilities documents for centralized geocoding and each spatial linkage service.

2.4.D. Content Needs

The metadata template provided through SND activities shall be evaluated to determine whether any optional or newly identified custom fields should be changed to required fields. For spatiotemporally integrated environmental health data, custom metadata fields must be identified to reference input event data, input auxiliary data, and the method(s) that were used to perform the linkage.

2.4.E. Technology Needs

The CEHTP metadata storage architecture should support distributed and centralized repositories simultaneously. For those data system owners who have the capacity and resources to produce their own metadata, and their metadata meets the required input field specifications, then CEHTP must align its metadata architecture to reference this important remote content. The metadata architecture must also account for situations in which CEHTP requires the augmentation of a remote metadata entry for optional fields that are not maintained by the system owner. For those datasets whose owners do not have capacity, resources, or whose existing metadata does not meet required specifications, then CEHTP shall provide a centralized repository for it.

The CEHTP shall provide a metadata application framework for inputting, searching, browsing, aggregating, and transporting metadata. These functions shall be available generically, whether metadata entries, in whole, or, in part, are situated locally or remotely. The application framework should accommodate the human resources that have control over and contribute to the most updated version of metadata. The framework should also efficiently account for frequent shifts over time in this human resource capacity. Technologies should emphasize a web services framework for interoperability and portability, taking advantage of mechanisms that already exist within the State. For convenience, client application programming interfaces shall be provided, which require minimal server installation overhead.

2.4.F. Coordination Needs

Where the required information is lacking from Future Assessment documentation, coordination is required with system owners to identify the structure and elements that determine surveillance events of interest to Tracking stakeholders.

CEHTP will need to coordinate access to metadata from system owners that maintain suitable metadata. For those that do not maintain suitable metadata, CEHTP shall perform suitable outreach and training to system owners so that they can inform their stakeholder base about CEHTP metadata content and application services.

In the event that other states or CDC begin implementing activities for enterprise geocoding, spatial linkage, and other geographic visualization/dissemination services, coordination may be required to determine consistent service metadata structure.

2.5. Integrate Data

2.5.A. Definition

There are two types of integration that the Tracking Network must satisfy: record- and systems-level. Record-level integration involves the combining of records from two disparate databases based on text, spatial, and/or temporal identifiers. Record-level integration is needed, because this is the primary method by which patterns of disparity and associations between environmental and health phenomena are analyzed. Systems-level integration involves bridging multiple disparate information systems in an application interoperability setting. Systems-level integration must occur first before records from remote information systems can be systematically integrated. Advances in network interoperability and service oriented architecture enable record- and systems-level integration to be parallel processes. The ultimate objective is the integration of environmental and health data, and implicit in this objective is systems interoperability and the standards-based exchange of data between systems.

2.5.B. Short-Term Needs

At an early stage, birth and death vital statistics datasets should be extended in text-based linkage services (probabilistic and/or deterministic). Interfaces to the service should be consistent and robust so that other health surveillance system owners (cancer, birth defects, hospital discharge,

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

etc.) can efficiently link patient records back to important vital statistics and, thereby, incorporate the linkage identifiers into their information systems architecture. Interfaces should also accept authenticated requests over encrypted channels so that the service can be broadcasted securely in the public domain. This would enable other external jurisdictions, such as local agencies, other State agencies, other States, or Federal agencies, to integrate their internal individual-level surveillance systems with vital statistics outside their jurisdiction.

For surveillance systems that do not have the capacity or resources to establish their own spatial-enabled database infrastructure, CEHTP will offer hosting for collaborating surveillance systems. In the case of health surveillance systems with confidential individual-level data, hosting will be in a secure DHS-administered domain without public access. For non-confidential data, which is typically the case for environmental hazard content, access to spatial data will be less restrictive. A spatial-enabled database is one that adheres to open standards, stores geometric features as a data type, allows for spatial indexing, and provides native SQL access to spatial operators, spatial functions, spatial predicates, and spatial measurements.

The second phase of the Future Assessment identified 10 priority environmental and health data providers that meet the criteria for environmental health significance, appropriateness, and availability. Spatiotemporal linkage services should be developed and hosted by CEHTP so that these datasets are rapidly extended for analysis on the Tracking Network. Each service should uniquely address issues of spatial dimensionality, spatial resolution, and spatial mismatch so that health and hazard metrics are transformed and aggregated to produce feasible linkage products.

2.5.C. Long-Term Needs

Individual-level event data from health surveillance systems other than vital statistics (cancer, hospital discharge, morbidity, etc.) should be integrated with birth and death identifiers. If other States implement similar birth linkage services, health events having incomplete birth linkage should be attributed with birth identifiers that are obtained from consuming out-of-state services. CEHTP will assist surveillance systems in meeting this requirement.

A deployed Tracking Network is one in which spatiotemporal linkage services are available for all environmental health datasets on the network. A Tracking Network that is properly maintained will account for regular updates to spatiotemporal services to reflect new methods and technologies for integrating environmental health event data.

The long-term objective for integration services on the Tracking Network is to shift the physical hosting location and administrative responsibility of spatial-enabled database and spatiotemporal linkage services to the surveillance system owners. This arrangement will place spatial analysis functions as close to originally compiled source event data as possible, while giving system owners maximum control over their information systems. CEHTP shall assist and provide

available resources to system owners to achieve this objective. One of the most important aspects of spatiotemporal linkage services is that the integration of confidential identifiers be performed in a secure and trusted location, and, that the products of the service not include confidential identifiers.

2.5.D. Content Needs

The process of linking and extracting vital statistics data will be resource intensive. Priority for this activity should be given to diseases that can be better understood by integrating knowledge about cases' perinatal conditions. Health surveillance systems that link individuals to birth and death certificates must restructure their data models to account for birth and death identifier codes.

In the short term, when CEHTP is responsible for hosting spatial-enabled databases, decisions about where to situate non-spatial attributes will have to be made. For example, if CEHTP hosts a spatial-enabled database for geocoded point locations of cancer registry data, should CEHTP also host the rest of the cancer attribute data? Or should CEHTP just provide host unique identifiers and spatial functionality? These decisions will likely be made on a case-by-case basis and will be influenced by a variety of factors. Some of the influencing factors include: surveillance system capacity for electronically exchanging tabular attribute data, update frequency, applications that utilize spatial database, and confidentiality concerns.

CEHTP in cooperation with CDC, system owners, and Tracking stakeholders should identify specific environmental exposures, whether theoretical or not, that require continuous evaluation. For those that are identified, data models must be restructured to account for the incorporation of new exposure metrics. For example, it could be decided that upper atmosphere levels of ozone is an exposure that requires continuous tracking relative to lung cancer incidence. Cancer registries will be able to extract an ozone level through air hazard spatiotemporal linkage services as cases are captured in the registry. A preparatory step, however, is to adapt the cancer content model to include an "ozone exposure" class.

2.5.E. Technology Needs

Text matching software must detect duplicates and match records quickly and easily, especially when the data is dirty, incomplete, inconsistent, and mis-fielded. The service should follow a web services framework and should allow for authenticated access over encrypted channels. Client applications that communicate to the service should require minimal server installation overhead and should allow for single and batch matching, and to a lesser degree, de-duplication.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

The technological requirements for a spatial-enabled database are fairly straightforward. The crux of the technology is an enterprise Relational Database Management System (RDBMS), which stores and manages access to geographic coordinate data. The Open Geospatial Consortium has created a specifications document through community consensus for simple features (points, lines, areas). The specification takes advantage of the indexing, analysis, and query operations that are standard in an RDBMS. There are products on the market that implement the specifications natively in the RDBMS (Oracle, MySQL, PostgreSQL) or through additional application layers that sit on top of the RDBMS (ArcSDE). The most important aspect of the OGC specification is that SQL queries can be written against a spatial-enabled database using topological operators and spatial functions. Also, recordset results can include geographic features.

The development of spatiotemporal linkage services should be based on spatial-enabled databases, an Application Program Interface (API) that implements OGC simple features specification, and web services over Hyper Text Transfer Protocol (HTTP) with optional Secure Sockets Layer. Client applications that communicate with spatiotemporal linkage services should require minimal installation overhead.

2.5.F. Coordination Needs

Birth and death certificate text linkage services will require coordination between CEHTP and the Center for Health Statistics to ensure that human subjects protections are successfully implemented. A website should be built internal to DHS, which documents the goals of the service along with example client usage, code downloads, API documentation, and a user forum. CEHTP shall perform outreach with surveillance system partners to encourage them to link individual level data to birth and death files, when appropriate.

The implementation phases should establish a stakeholder process for identifying health and hazard events that require linkage services, incorporating new spatial linkage methodologies, determining client functionality requirements, and selecting exposures which should be tracked continuously over time. The transition from short- to long-term goals, in which system owners assume a more direct role in providing environmental health tracking services, will require significant outreach, education, and training. CEHTP will encourage and work with members of SND to form consensus on interface specifications for spatiotemporal linkage services.

2.6. Exchange and Visualize Data

2.6.A. Definition

The exchange of data involves two processes that bring together producers and consumers of environmental health surveillance data. Producers of data must provide mechanisms for disseminating or publishing their datasets. Consumers of data take advantage of these publishing mechanisms by implementing mechanisms for accessing or downloading data. The process of publishing and access or production and consumption results in a data exchange. We include the process of visualizing data in the data exchange network function, because visualization is a special case of exchange in which disseminated data is more easily comprehensible to humans.

2.6.B. Short-Term Needs

The CEHTP has identified the early need to implement client and server-based EPA Exchange Network Node tools. Though there are currently no specific identified environmental hazard datasets requiring production or consumption, CEHTP would like to position itself to assist environmental surveillance systems quickly should they wish to publish their official hazard datasets to the Exchange Network. Equally, it would serve CEHTP well to have in place the capacity to quickly consume hazard data through a Node Client, if one of environmental health tracking priority became available. In the coming years, multiple consumable environmental hazard datasets (e.g. drinking water, pesticides) will come online, and CEHTP will be positioned to quickly take advantage of this availability.

The same holds true for CEHTP's capacity to utilize the PHIN Messaging System. In the near term, childhood lead surveillance and confidential morbidity reporting surveillance data will be available over PHIN. CEHTP can prepare for these events by having the capacity to utilize the PHIN Messaging System.

CEHTP does not assume the role of disseminating PHIN or Exchange Network content to the public, especially not confidential data. The default assumption is that surveillance systems will publish and manage secure, role-based access to official analytical datasets. Should surveillance systems desire assistance in this important specification, CEHTP will offer collaborative and infrastructural resources. Should the opportunity exist in the intervening time to publish de-identified official datasets, CEHTP will take the prerogative in doing so.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

Short-term implementation should be marked by the production of useful non-confidential dissemination and visualization products. CEHTP will utilize its capacity for developing static and dynamic Web Map Services, Web Feature Services, and tabular- and chart-based dissemination tools. The content and services that comprise these products will be hosted through CEHTP's web portal.

Previous discussion established spatiotemporal linkage as a priority network function in terms of record- and system-level integration. Spatiotemporal linkage should also address issues of data exchange and visualization. In the early stages of implementation, the spatiotemporal integration architecture should emphasize visualization. For example, following the deployment of a dataset-specific linkage engine, the CEHTP website that provides information about the service should host a real-time visualization demonstration with contextual comparison information. A user who executes a service for a specific time and space event domain can compare the results to precompiled comparison information for nearby event domains and other related aggregations. This comparison information will amount to a 10% sample of the California population and should be assembled, geo-referenced, and attributed with hazard- and health-centric linkage metrics. For example, a point-based coverage of 10% of the California population will be attributed with preterm birth rates, term low birth weight rates, pesticide usage, and traffic rates. This sample will be applied to the entire population and summarized for various administrative regions (block groups, tracts, cities, counties, etc.). A real-time request for spatiotemporal linkage at any time and space event domain can be compared to the 10% sample results to give the user an idea of how the spatiotemporal domain of their specific requests compares to that of the nearby or aggregated population.

2.6.C. Long-Term Needs

CEHTP's ultimate strategy for dissemination of official surveillance event data is for system owner's to assume primary responsibility for utilizing the technological tools provided through EPA Exchange Network or Public Health Information Network to expose their official surveillance data holdings to stakeholders who have a right and need to consume the data. Whether official surveillance data holds confidential identifiers or not, system owners will be encouraged and assisted by CEHTP to develop authentication and role-based access mechanisms that meet the standards and specifications stipulated by the federal data sharing initiatives.

CEHTP views dissemination as a very important function, but the ability to ensure that surveillance data is being disseminated to the widest audience possible does not rest solely with CEHTP. Local data intermediaries and topic-specific organizations/agencies are viewed as data sharing partners who increase the visibility and infusion of surveillance data to widely dispersed stakeholders over the network. Through the course of implementation phases, CEHTP will investigate methodologies for efficiently disseminating surveillance data to these "meta-portals". For example, CEHTP will collaborate with and assist organizations such as Pesticide Action

Network, InfoAlameda, and Neighborhood Knowledge California to implement client data exchange tools so that they can in turn exchange the very same surveillance data to specialized and targeted audiences through their own dissemination and visualization applications.

While early stages of implementation will emphasize visualization of spatiotemporal linkage services, the mature stages of implementation will ensure that these services are reaching the widest audience possible. CEHTP will work toward developing robust client invocation tools that have low installation overhead and can perform batch integration requests on diverse input data. If spatiotemporal linkage can be aptly visualized and its functionality can be easily consumed, then it will have the highest potential of reaching the stakeholders that need it the most.

2.6.D. Content Needs

The official dataset provided by surveillance systems for external consumption often follows a format and structure not easily adaptable to environmental health investigations. Historically, environmental health personnel devote a significant amount of time to subsetting, transforming, and aggregating the official dataset to match the requirements of their investigation. There is a strong need to minimize resources expended on this process such that official dataset extracts match environmental health requirements as close to the point of data exchange as possible.

Health surveillance data are often legally protected to prevent the disclosure of individual-level identifiers. At the same time, there is a great need to visualize patterns within a health outcome domain. For this reason, it is a primary goal of CEHTP to develop and deploy methods that mask the identity of individuals, but at the same time demonstrate group-level patterns at useful spatial and temporal scales. Raster surfaces derived from kernel estimation and for use in a GIS are one such solution which allows the visualization of confidential data. CEHTP will continue to examine ways in which confidential data can be processed and visualized so that individual-level identifiers are secured, but patterns are revealed. CEHTP has made considerable strides with cohort data, such as birth records, but more development is needed for population-level data. One of the key complicating factors is using Census-level population data as a denominator in kernel estimates of standard incidence rates.

The content needs surrounding the visualization of spatiotemporal linkage services concern the development of a comparison dataset for identified hazard- and health-centric scenarios. As described previously, this would amount to a point-based coverage constituting a 10% stratified random sample of the California residential population as derived from the Census, and refined from parcel-based land-use or general plan data, tax assessor data, and geocoding reference files. The comparison dataset will be updated with hypothetical exposure metrics on a schedule that depends on the selected spatiotemporal linkage services of interest and the update cycle of each services' underlying surveillance data.

The development of Web Map Service visualization tools depends on content needs that are specific to the dataset being visualized and general attributes of the target audience consuming the service. Oftentimes, a useful visualization scenario depends on target-audience specified requirements for aggregations of the dataset. Moreover, for datasets with highly resolved spatial and temporal event domains, precompiled aggregations are necessary, because processing at request time would lead to significant computational deficiencies.

2.6.E. Technology Needs

A dominant requirement identified by CEHTP for implementation of data exchange services is that CEHTP retain in-house expertise for implementing an Exchange Network Node and the PHIN Messaging System. This requirement does not presuppose that all dataset-specific exchange mechanisms developed in coordination with CEHTP also be developed by CEHTP. It merely ensures that CEHTP has substantial knowledge into the architecture of both exchange mechanisms should surveillance system owners entrust CEHTP to perform that function or should CEHTP wish to scrupulously procure the service from a third party. For any implementation, whether developed in-house or through a third party, the components that comprise the exchange mechanism will be evaluated in terms of modularity, vendor neutrality and adherence to industry and framework (government data sharing initiative) standards. For example, a low risk implementation is viewed to be one that can execute in the most operating system platforms, can interface with the most RDBMS platforms, follows functional standards for web application services, and implements the majority of PHIN or NEIEN specifications.

Technological requirements for visualization and dissemination services center around three areas: server-side components, client-side components, and functional interfaces that link them. On the server-side, it is particularly important that components are portable and interoperable, such that future platform migration is possible and not prohibitively difficult. This is an important requirement, because many of the services that CEHTP develops are built for a single surveillance dataset and the assumption that the system owner might one day take on the responsibility of administering and hosting the service. In exposing visualization and dissemination services, CEHTP has also recognized the need to follow functional interface specifications and standard transfer protocols so that clients can consume services in a platform-neutral manner. Yet on the client-side, the ease of consuming a service is as important a requirement. Visualization and dissemination services must balance the need for ubiquitous interoperability, while, at the same time, providing interfaces that are easy to consume.

2.6.F. Coordination Needs

Once CEHTP has developed internal skills to deploy EPA and CDC data exchange infrastructure and has actually deployed one each of an Exchange Network Node and PHIN MS, we will perform a marketing, outreach, and training program to system owners to encourage them to deploy their own messaging architecture or allow CEHTP to assist them in deploying one. If partnerships are established between CEHTP and system owners to deploy messaging mechanisms, then considerable coordination will be required to determine requirements and reach a mutually beneficial architecture.

In situations where CEHTP is developing visualization and dissemination services, coordination will be required with system owners to apprise them of the activity and with potential meta-portals to determine interoperability requirements.

CEHTP needs to communicate with target audiences and system owners to identify useful aggregation schemes or other derived/computed products to be used in visualization services.

CEHTP needs to identify and work with the relevant players that have a stake in determining which spatiotemporal linkage services require automated update to a population comparison dataset.

3. NETWORK TECHNOLOGY

3.1. Overview

A critical component to the EPHTN is the technology infrastructure. Technology may be simple, such as using email to send data to an entity that compiles those data and redistributes. Or it may be significantly more complex with disparate surveillance systems, corresponding schemas, virtual and physical data repositories, and robust client interfaces that result from agreed upon protocols and methods of operation. Though CEHTP is situated in a large health department, where many surveillance data systems exist, CEHTP does not have immediate control or access to these data resources. Because CEHTP is not the “owner” of these surveillance databases and, yet the EPHTN vision is to integrate these data sources together in a surveillance setting, CEHTP views itself as a service provider. For that reason, each of the technological components that comprise the EPHTN in a service-oriented architecture are “services” themselves. These components are: data enhancement services, text-based linkage services, spatiotemporal linkage services, exchange services, visualization services, and metadata services.

3.2. Data Enhancement Services

3.2.A. Overview

The data enhancement services that the CEHTP aims to provide for system owners (and local reporting jurisdictions) address completeness and accuracy issues in the geospatial realm. A major theme in the development of these services is the ultimate goal of visualizing and integrating surveillance data. There are two identified technological paths that satisfy data enhancement requirements: enterprise (centralized) geocoding and enterprise feature input. Enterprise geocoding is an address broker that extracts geographic coordinates, such as longitude/latitude or census region identifiers, for various users and applications across a networked enterprise. The address broker provides address standardization, verification, geocoding, and overlay with auxiliary datasets. Enterprise feature input is a server- and client-based architecture for importing, creating, editing, and exporting geographic features across a

networked enterprise. Implicit in both services is standards and specifications for application interoperability.

3.2.B. Architecture

Enterprise Geocoding

Enterprise geocoding is handled by a unit transaction, in which a client sends a request to the server, the server processes the request, and the server responds with the output results to the client. In the case of geocoding, a request can include one or more addresses. Equally, the response can include output for one or more geocode results. CEHTP has designed this operation around the web service standard. Web services are desirable in this case, because XML provides an interoperability standard for messaging over the WWW, and SOAP provides an interface to remote methods that can input or output serializable objects. Therefore, client and server implementations do not have vendor-dependent platform restrictions; for example, Microsoft can talk to Java can talk to ESRI.

The serializable objects or “types” used in enterprise geocoding are:

- Address – street (prefix, number, street name, type, suffix, etc), zip code, city name, and error (error codes from CASS-certified address standardization/verification package). Used in both input and output to geocoding engine.
- GeocodeOptions – Client customizable options for how addresses are geocoded in a session. GeocodeOptions are sent to the geocoding service prior to sending addresses for geocoding.
 - doStreetId (Boolean) – Flag whether street segment ID is returned
 - doStandardizedAddress (Boolean) – Flag whether standardized street address is returned
 - doRegionId (Boolean) – Flag whether region IDs from overlaid auxiliary datasets are returned
 - doZipAsZone (Boolean) – Flag whether zip code should be used as index key
 - doCityAsZone (Boolean) – Flag whether city soundex code should be used as index key
 - doFirstMatchingCoordOnly (Boolean) – For multiple street databases used, flag whether the first matching geocoded coordinate is returned
 - doMultiServiceErrorMetrics (Boolean) -- For multiple street databases used, flag whether distance from centroid of coordinates (average error) is returned
 - spellingSensitivity (int) – Percentage score as indicator for geocoding engine of sensitivity to spelling sensitivity

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

- `minimumMatchScore` (int) – Minimum score of potential candidates to be considered a result.
- `sideOffset` (int) – For interpolated coordinates along a street centerline, distance from centerline, right or left, to offset geocoded coordinate
- `sideOffsetUnits` (String) – Units of `sideOffset`
- `streetResources` (String []) – Sequential list of resource IDs for street centerline and other geocoding reference files to use in coordinate extraction
- `standardizationResources` (String []) – Sequential list of resource IDs for standardization packages (ZP4, USPS, etc) to be used in address verification and standardization
- `regionResources` (String []) – Sequential list of resource IDs corresponding to auxiliary polygonal datasets to be overlaid with coordinate results
- **GeocodeRecord** – The processed result of a geocoding transaction that is returned to the client. Many of the fields in `GeocodeRecord` are lists or arrays (marked as a []) to indicate sequential values for each `streetResource` used in the geocode processing.
 - `status` (String []) – The status of the geocode: matched (M), unmatched (U), or tied (T); Tied and unmatched records have a score of 0 and the rest of the fields below are nullified
 - `score` (int []) – Value assigned by geocoding engine to indicate matching degree of success
 - `side` (String []) – Side of the street (L for left or R for right), if the `streetResource` is a street centerline reference dataset
 - `x` (double []) – Longitude value
 - `y` (double []) – Latitude value
 - `streetID` (String []) – Geocoded street reference identifier
 - `regionIDs` (RegionIDs) – See below
 - `metadataID` (String []) – An identifier to indicate which street reference resource this geocode result came from, including a code for whether a zip index or city soundex index was used
 - `averageError` (double []) – Distance of this coordinate from centroid of all other coordinates found for this request
 - `standardizedAddress` (Address) – Address returned from first successful address standardization sub-request.
- **RegionIDs** – A list of extracted regions for a single geocoded coordinate that is embedded in a `GeocodeRecord`

The SOAP interface methods used in centralized geocoding are:

- `getCapabilities()` – The `getCapabilities` method is not yet implemented. This method returns a `Capabilities` document describing the geocoding service. The `Capabilities` document includes the WSDL output of the web service. However, it should also include metadata about required/optional `GeocodeOptions` and values to those options as well as metadata for each of the resources (reference, standardization, and region overlay) currently available to the service.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

- initializeTool (String, String) – The initializeTool method takes a username and password over Secure Sockets Layer only and checks the tool registry to authenticate the requested user for a specific tool. If successfully authenticated, then each of the geocoding service's methods below can be successfully executed. Otherwise, the methods below will fault with any invocation
- setGeocodeOptions (GeocodeOptions) – Writes the client instance of the GeocodeOptions object to the session.
- findAddress (Address) : GeocodeRecord – Sends a client instance of the Address object to the geocoding engine, processes the Address according to the instructions provided by the GeocodeOptions object, and returns a GeocodeRecord result to the client
- findAddresses (Address []) : GeocodeRecord [] – Same as findAddress method, only multiple Address objects are sent to the server in a single request

Enterprise Feature Input

The architecture of an enterprise feature input service is largely undetermined. This requirement was discovered late in the planning stages of the Tracking grant. There are, however, key aspects of an architecture that can be described here. The most important aspect is that a central spatial-enabled database is required to satisfy any central feature input service requests. In addition, the OGC Web Feature Service specification would be the standard for which features are streamed to or queried by the client, new feature instances are created, and existing features are deleted, updated, or locked. The primary difficulty lies in allowing clients to utilize their existing GIS platforms, including Computer Aided Design platforms, to interact with the service, while, at the same time, allowing more standards-based clients (such as browser-based clients) to simultaneously consume the service. Fortunately, most GIS platforms recognize the power of OGC standards and are incrementally implementing them within their products. Over the past year, there has been a blossoming of activity on the browser-based front. With the introduction of the standards-based Google Maps client, a user community has concentrated activities on harnessing this platform in concert with WFS services to create rich data streaming applications in a browser, while not requiring download and installation of vendor-dependent plug-ins. What is needed now is the ability to perform updates through the browser in a similar way.

The other aspect to a feature input architecture is to provide a service that accepts vendor-dependent spatial formats. The client's objective would be inserting features into a larger feature dataset. The service, however, would provide a staging area within the spatial-enabled database to verify first whether the features are suitable for inputting. The flow-process would start by the client providing the service datastore parameters to access the features, coordinate reference system identifier, and a subsetting conditions. For file-based features, it would also include an upload step to the server. Following this, inputted features to the service can be viewed and

edited through a WFS client. Once the features are determined to be suitable for insertion, the service would copy from the staging area into the larger production dataset.

3.2.C. Development Plan

Enterprise Geocoding

Since the piloting of enterprise geocoding was a primary activity during planning phases, many aspects of its development have already reached a mature stage. However, in moving into implementation phases with enterprise geocoding, the completion of the `getCapabilities` method is necessary, and a number of new optimization and accuracy enhancements have been identified that will prepare this service for full-scale deployment. The most important function is an optimized run-time spatial database connection framework that continuously invokes and refreshes connections to reference datasets and indices. Currently, spatial connections are obtained at a slower rate than desired after periods of inactivity. Another requirement is to migrate storage of geocoding reference datasets and indices to an OGC-compliant spatial-enabled database structure, in which geometric shapes are stored as native data types. It should be noted that there is a lack of commercial technology that supports native data types simultaneously with geocoding support to reference files that are not based on street centerlines, such as discreet (not interpolated) georeferencing to points, lines, or areas. For optimal maintenance, centralized geocoding requires the incorporation of an address standardization package with lower maintenance overhead, such as USPS Address Information APIs. The system would also benefit from a pre-geocoded central address file and a reporting system that accepts input from local postmasters who have post office box-to-address translational information.

Enterprise Feature Input

The details of feature input service development are still under consideration. It is clear, however, that a SOAP-based web API will be needed at least for transport and input of vendor-specific data formats from client sites. SOAP is also listed as a potential transport mechanism for Web Feature Service communications. The GeoTools Java open source GIS toolkit would provide the methods for converting various input formats to the format required for insertion into the central spatial-enabled database. On the browser-based client side, a platform for visualizing and editing WFS content will need to be developed. The Google Maps [javascript] API could provide a good foundation for this type of implementation, though much more enhancement and tweaking would be necessary to allow it to consume and edit WFS content.

3.2.D. Deployment Plan

Enterprise Geocoding

Currently, CEHTP enterprise geocoding is deployed in a secure location within CDHS LAN, and it is deployed in a public location outside the CDHS LAN. Two agencies within the Health and Human Services Agency have expressed interest in assuming an ownership role for the service. CEHTP will encourage, assist, and devote resources for any top-level agency to perform this role. Having the service running from multiple locations is not a drawback and could actually lend to distributed processing in a failover clustering or redundancy scenario.

Central to the success of enterprise geocoding is ensuring that it is utilized by all of the stakeholders that need it. During the planning phases of the Tracking grant, we experimented with various client implementations to the service. For example, stakeholders have utilized downloadable software for installation that interfaces with a local database, communicates with the remote service, and updates an address table with the geocoding results. In another client implementation, new records inserted into a database table triggered the geocoding process. The CEHTP believes, however, that each of these implementations and any other newly identified ones should be downloadable and documented at a main website. The website will have other additional features that encourage the knowledgeable utilization of the service. The most important feature is the service API documentation describing the interface methods, objects, and metadata. For the lay visitor, these descriptions will be paraphrased in a frequently-asked-questions section. The geocoding site will also include a user registration, authentication, and dynamic data import (upload), process (geocode), and export (download) utility. Multiple import formats will be supported for uploading and import into the system. For software developers, there will be example implementations, client code downloads, and a moderated user forum.

Enterprise Feature Input

The deployment plan for enterprise feature input services is under consideration. Like other CEHTP services, it can be surmised that ownership of this service will likely be shifted towards system owners requiring control of its functionality.

3.3. Text-Based Linkage Services

3.3.A. Overview

A text-based linkage service developed by CEHTP will provide stakeholders, mostly health surveillance systems, with the ability to match one or more inexact and non-unique record-level

attributes to a large cohort dataset, such as birth or death vital statistics data. The benefit of this network technology is record-level integration. In some integration situations, where temporal and spatial identifiers cannot be used for record-level integration, text-based linkage can be used. The ideal case of the service's usage would be a surveillance system that centralizes reportable health outcomes, with records being automatically updated of record-matching output, transparent to the user, and in real-time as data is entered into the system. Later, down the road, when the surveillance data is consumed it can be easily integrated with the matched cohort data through identifier-to-identifier joins.

3.3.B. Architecture

Like other CEHTP services, the architecture of text-based linkage services will follow a client-server model. A typical request-response flow will originate with a client request for matching on one or more records in one or more reference datasets. The server responds with output results having one match identifier for each reference dataset requested. Web services are suitable for this architecture, because XML provides an interoperability standard for messaging over the WWW, and SOAP provides an interface to remote methods that can input or output types that can easily be marshaled and sent over the network.

The “types” to be used in text matching services are:

- **MatchInput** – The MatchInput class is used as input to the text linkage service. It holds multiple fields of data, the universe of which is determined by the reference dataset, but is limited by the field schema of the input dataset.
- **MatchOptions** – MatchOptions are client customizable options for how records are matched in a session and for the structure of responses. It includes specifications for which reference datasets are desired in the matching process. MatchOptions are sent to the text linkage service prior to uploading records for matching.
- **MatchOutput** – The MatchOutput object is sent back to the client. It includes the processed results of text linkage. For each requested reference dataset, it holds the corresponding unique identifier and match success metric of the matched record.

The SOAP interface methods used in text linkage are:

- **getCapabilities()** – The getCapabilities method returns a Capabilities document describing the text matching service. The Capabilities document includes the WSDL output of the web service. It also includes metadata about required/optional MatchOptions and allowed values for those options as well as metadata for each of the

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

reference datasets currently available to the service and fields used for matching within those datasets.

- initializeTool (String, String) – The initializeTool method takes a username and password over Secure Sockets Layer only and checks the tool registry to authenticate the requested user for the text-matching tool. If successfully authenticated, then each of the text match service’s methods below can be successfully executed. Otherwise, the methods below will fault with any subsequent invocation.
- setMatchOptions (MatchOptions) – Writes the client instance of the MatchOptions object to the session.
- getMatch(MatchInput) : MatchOutput – Sends a client instance of the MatchInput object to the text match engine, matches the record to each reference dataset specified by the instructions provided in the MatchOptions object, and returns a MatchOutput result to the client
- getMatches(MatchInput []) : MatchOutput [] – Same as getMatch method, only multiple MatchInput objects are sent to the server in a single request

Because automated processing is desired in a machine-to-machine setting, there are no plans currently to allow the service to operate in “candidate” or “interactive” mode. An “interactive” mode would potentially allow the service to return to the client possible candidates for a disputed match. To accomplish this task, the service would have to return detailed attribute information from the reference dataset for each candidate. This is not a desired function from the view of CEHTP, because it requires the output of extremely confidential attribute information to external clients. At the same time, outputting just an identifier from a reference dataset presupposes that the matching process is very conservative in determining a single gold-standard match. Although this is the planned architecture, there is risk for certain matching conditions resulting in a preponderance of false negatives.

3.3.C. Development Plan

Developing a text-based matching service involves selection of a commercial text matching product. The criteria for selection are:

- Processing speed relative to base table size – Are there limits to the size of a reference dataset? How is match speed affected by size of reference dataset?
- Cost – Are up-front cost and maintenance costs prohibitive?

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

- Platform support – Does the package run under multiple operating systems, network protocols, and database management systems?
- API support – What type of application development languages interface with the package?
- Recognition – What do customers, preferably within EPHTN, say about the package?

Once a text matching package is chosen, the second step involves coding to the interface schema described above utilizing the API provided by the text matching package. Once a web service exposes the necessary interface methods, sample client requests are sent to the service to test speed and stability. At this stage, configuration tweaks and iteration are performed to optimize computation speed.

3.3.D. Deployment Plan

In most value-added service implementations, CEHTP has emphasized the requirement that control or ownership of the service be transferred to the system owner's domain. In the case of text-based linkage services, there is less emphasis on this requirement. Though still a potential unrealized benefit, currently CEHTP sees no apparent value, and, in fact, disadvantages, to the migration of ownership of text linkage services to system owners. Some of the drawbacks include:

- Licensing – Record-matching packages are expensive, both in terms of initial costs and ongoing costs. If CEHTP is to provide resources for obtaining and maintaining a license on which multiple reference datasets will depend, then it does not make sense to disperse copies of the software to individual surveillance systems where conditions for additional license renewals might be required.
- Processing Architecture – In other value-added service scenarios, it is of extreme benefit to situate the processing components of the service as close to the physical storage location of the official updated dataset on which the service depends. In the case of text matching, entire reference datasets, including large index derivations, are typically held in memory allocations to achieve the most optimal cross-reference capacity. There is, hence, a data lookup and exchange step of incorporating physical records from a reference dataset into the text matching package. This step cancels the benefit of situating application-tier components in proximity to “raw” data.

Therefore, the case of deploying text-based matching services requires the centralization of processing. CEHTP will install, configure and maintain custom libraries and third party packages on devoted hardware within CDHS. Initially, the service will only be available to clients on the

CDHS local area network. Once it reaches a critical stage of maturity, stability and usage, and, if there are external requests for consuming it, CEHTP will coordinate with information technology oversight agencies within the State to expose its functionality to outside audiences.

The main objective of deploying a text-based linkage service is to integrate large health surveillance systems. On a case-by-case basis, CEHTP will coordinate with surveillance system owners to configure their unique reporting systems architecture to accept as input the linkage identifiers provided through matching of reference datasets. These partnerships will accomplish the task of integrating future updates to surveillance systems. There is, however, the issue of linking historical surveillance data. Because there are a limited number of surveillance systems where historical data require matching, CEHTP will accept and execute informal one-time requests.

3.4. Spatiotemporal Linkage Services

3.4.A. Overview

Spatiotemporal linkage services aim to provide automated record-level integration of environmental hazard and health surveillance data over the Network using temporal and spatial identifiers. When spatial linkage systems are developed and deployed using consistent standards and specifications, the following is possible:

1. Spatiotemporal integration can be invoked in the same way for state- and national-level tracking systems across any environmental or health content area.
2. Spatiotemporal integration can result from standard requests originating in an insecure domain, while actual processing occurs in a secure domain. Environmental health tracking stakeholders have the opportunity of integrating environmental and health data without having to physically access confidential individual-level identifiers.
3. Creating new spatiotemporal integration services are easier with each dataset-specific implementation.
4. Following similar web-based standards and specifications, spatiotemporal integration can be suitably incorporated into existing data sharing/interoperability architectural frameworks.
5. Associations between environmental and health phenomena can be tracked in real-time as surveillance data are reported

3.4.B. Architecture

The spatiotemporal linkage architecture treats environmental and health data generically by establishing one as the primary dataset and the other as the secondary dataset. Spatial and temporal identifiers of the primary dataset are submitted to the system in individual transactions. These identifiers are used to transform attributes of the secondary dataset to create a linkage product that is returned to the client. For example, in a health-centric scenario, health surveillance data is considered the primary dataset and environmental hazard data is the secondary dataset. For any given health event submitted to this system, environmental hazard events that correspond to the health event in space and time are transformed and summarized to create a final environmental hazard metric/product that can be related to the health event. This means that specific spatiotemporal linkage implementations are created with primary datasets in mind as a runtime consideration, but should be developed physically near and with tight functional coupling to the secondary dataset. This architecture is described in ample detail in the Appendix.

3.4.C. Development Plan

Many aspects of the spatial linkage architecture have already been developed and demonstrated during the planning phases of the Tracking grant. The implementation phases will be marked by enhancements to existing code libraries, support for additional application and database platforms, the implementation of more content-specific services, and the development of low overhead client solutions.

The current libraries being used for spatial linkage rely on a limited set of geometric input parameters that interface with a vendor-dependent spatial datastore model. Both of these deficiencies can be solved by moving to an API that implements a standard geometry model, hopefully provided by OGC, and can connect to multiple datastores. GeoTools is an open source Java GIS toolkit, implementing the OGC Simple Features specification (i.e. standard geometry model) and providing generic interface access to over 10 widely used spatial data formats or spatial-enabled databases. To be able to support a range of GIS data platforms is an incentive for system owners to buy-in to services that CEHTP provides. CEHTP already uses GeoTools to a small degree; this usage will increase through implementation.

Many of the components of the existing spatial linkage library were written in the Java platform. Because CDHS IT hosting services supports Microsoft-only web applications, it would be beneficial to port the spatial linkage libraries to the Microsoft platform. During implementation, CEHTP will investigate the use of IKVM.NET to implement Java for the .NET Framework. The tools provided by this package allow the use of Java libraries in .NET applications and vice-versa. Having both a .NET and Java code library available for assisting surveillance system owners in

implementing spatial linkage services is a time-saving incentive that could lead to markedly faster development and deployment timelines.

During the planning phases of Tracking, multiple spatial linkage services were developed. Each of these was health-centric. The implementation phases of Tracking will see early development of hazard-centric services. This endeavor will lead to iteration and enhancement of spatial linkage libraries as well as requirements evolution.

Spatial linkage services are SOAP-based web services. By definition, SOAP is interoperable. Any client that can generate an XML message and accept an XML response can consume a SOAP-based service. Some application platforms provide automated tools for consuming a web service. Using these tools is relatively easy for technically-skilled personnel, yet it is hoped that spatial linkage services can be consumed by audiences more far-reaching. CEHTP will pay close attention to emerging industry trends surrounding client-based functionality and interoperability. For example, in the last year there has been a renaissance in development of client-based web mapping sites surrounding the release of Google Maps API. This API is based on a composite technology called Asynchronous Java and XML (AJAX). Though in its infancy and by no means consistently supported across web browsers, this technology allows for the real time streaming of web service data without having to reload a page. This trend holds enormous potential for providing low overhead client functionality vis-à-vis spatial linkage services. At least in browser-based clients, there is the possibility that a client implementing AJAX functionality can receive robust CEHTP linkage services without any server installations, and a minimum application development skill level.

3.4.D. Deployment Plan

There are two primary areas of technological importance in deploying spatial linkage services. The first is the plan for physical hosting, administration, and maintenance of linkage services and the second is the plan for ensuring that linkage services are consumed by stakeholders that need it.

At the point when the business rules and surveillance data have been coupled through application development into a deployable service, there are many factors which influence the decision on how to determine ownership of a linkage service. Ownership in this case deconstructs to the physical positioning of the software components and the frequent administration hurdles of maintaining upgrades, equipment migration, and server uptime. One very important influencing factor is that CEHTP is usually the only application developer involved in a linkage project. In other instances, CEHTP may enter joint application development ventures with a system owner. Another important factor is that a linkage application should be physically situated close to the surveillance data on which it depends; concurrently, surveillance systems are moving toward electronic exchange mechanisms. Security of data resources is also a major factor.

Whatever the factors are, it will be CEHTP's goal to establish ownership and control of linkage services at the relevant surveillance systems' local domains. Hence, typically, ownership of a linkage service will initiate with CEHTP, and, perhaps, the secondary data on which it depends will be copied and maintained within the CEHTP domain. This ownership situation, however, will shift as the linkage process stabilizes, matures and reaches a critical number of stakeholders. Through support and resources provided by CEHTP, the system owner should realize the benefit in maintaining the service within its own domain. CEHTP and system owner linkage partners have already found uses for linkage services beyond just environmental health integration. In alternate deployment scenarios, CEHTP is prepared for instances where the system owner does not have the incentive for providing local ownership of a linkage service, in which case CEHTP will maintain ownership through the life of the service. Or in another scenario, CEHTP is prepared to relinquish considerable control of the linkage deployment (as well as development) should a system owner raise the requirement that the linkage service be wholly deployed at the system owner's local domain from the start of the project.

Because spatiotemporal linkage services have an architecture that emphasizes back-end mechanisms, it is necessary to address the front-end (or presentation) issues of deployment so that it is clear these services will reach their desired audiences. One of the most important aspects of each spatial linkage service is documentation. CEHTP will establish a boiler-plate documentation website for each linkage implementation. API documentation and frequently-asked-questions with extensive discussion into method caveats, intentions, and drawbacks are the standard for documentation sites. The sites, however, will not only include instructions on how to consume the services in various client scenarios, it will also provide routines for outputting ToolMetadata and DefinitionMetadata list extracts. There will be utilities for tool user registration and authentication. It will also include dynamic presentation for invoking the service, visualizations of its results relative to comparison datasets and on top of aggregated visualization products, and the ability to import input datasets and export output results to common data editing mediums (text file, MS Excel, Access, etc). For software developers, these sites will include example implementations, client code downloads, and a moderated user forum.

3.5. Exchange Services

3.5.A. Overview

The objective of exchange services for the CEHTP is electronic data exchange that implements the specifications prescribed by the corresponding national data sharing and interoperability initiatives: National Environmental Information Exchange Network and Public Health Information Network.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

During the planning phases of the Tracking grant, the CEHTP did not have the opportunity to electronically exchange public health data through PHIN-compliant messaging mechanisms (PHIN-MS), nor did it have the opportunity to consume environmental hazard data through NEIEN-compliant messaging mechanisms (Node Client). This is due, in part, to the fact that outcomes and hazards investigated by pilot projects utilized data from surveillance systems that do not have any activities established around the federal data sharing initiatives. This lack of activity is also due to the fact that many environmental and health data systems that are performing electronic data sharing activities are either in planning or development stages, and/or the content of production systems is not of environmental health relation or priority.

There are, however, two environmental flows under development, pesticide illness reporting and pesticide use, that have significant environmental health priority, and, as it stands, were funded with CEHTP as a partner.

On the health side, flows under development that are of environmental health priority tend to favor local-to-state or laboratory-to-state reporting. It is clear that electronic data exchange within the health department is in its infancy. CEHTP awaits and will encourage and/or assist surveillance systems in enhancing their surveillance information systems so that Tracking stakeholders can electronically consume official health event datasets.

Pending the availability of health and environmental event data over NEIEN and PHIN, CEHTP will investigate during implementation phases the use of PHIN-MS, Exchange Network Node server components, and other data sharing mechanisms to exchange environmental health integrated data. This objective includes the coupling of real-time automated spatiotemporal linkage with PHIN and NEIEN data transfer tools.

3.5.B. Architecture

There are two architectural pathways that CEHTP envisions for developing an electronic data exchange infrastructure for environmental health tracking. The first is one that has already been laid out by EPA and CDC through their corresponding data sharing and interoperability initiatives, each applying to and is the responsibility of individual environmental and health surveillance systems. The second is the incorporation of real-time environmental health record-level integration with existing data sharing frameworks. In order to devise the second framework, detailed knowledge and implementation experience concerning the first must be acquired and developed by CEHTP staff.

3.5.C. Development Plan

In the short term, there are two identified opportunities for developing data exchange capacity that incorporate the ability to link environmental and health data on a record-by-record basis. The first concerns joint PHIN and NEIEN development activities surrounding NEIEN-based Pesticide Illness Reporting system, the PHIN-based Web Confidential Morbidity Reporting system, and the NEIEN-based Pesticide Use Reporting system. The second involves the systematic and ongoing flow of environmental health integrated datasets to be incorporated into centralized and consistent datasets maintained by the CDC Tracking Branch.

While pesticide use (hazard event) and illness (outcome event) data will flow as ebXML and SOAP exchange messages within and between individual health and environmental programs at local and state levels, the responsibility of performing record-level integration on these data rests with the CEHTP. And the establishment of this process as a systematic process with tie-in to existing data exchange tools has been expressed as a requirement by the Cal/EPA Office of Environmental Health Hazard Assessment (OEHHA) to the extent that they are mandated to perform an assessment of each pesticide illness report, surveying suspected or hypothetical pathways of exposure. CEHTP, in cooperation with OEHHA and the Department of Pesticide Regulation, will investigate development methodologies for systematically attributing pesticide illness health event records with pesticide use information that is dynamically calculated and exchanged using existing tools.

Though the detailed requirements for contributing to consistent national datasets have not been established, it is clear that States will be reporting both environmental and health data in formats that are feasible for analyzing in light of each other. This requirement is satisfied in part by spatiotemporal linkage. The piece that is missing is the ability to flow the products of this integration to CDC using a consistent and shared methodology. For this reason, CEHTP will work towards developing the methods and vocabulary by which health data and hazard-derived spatiotemporal linkage products can be sent to CDC by an ebXML Client in asynchronous, synchronous or route-not-read mode.

3.6. Visualization Services

3.6.A. Overview

For visualization services, CEHTP places primary emphasis in the Internet mapping technological arena. Though services which would provide charts, graphs, and tables are of equal importance to spatial representations, many detailed aspects of implementing these traditional visualization

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

products have underpinnings in PHIN and NEIEN activities that many individual surveillance systems have already addressed. Moreover, CEHTP has made specific targeted investments in human and infrastructural resources to assist surveillance systems in implementing spatial-based visualization services. One large aspect of this investment has concerned the implementation of standards so that the benefits of the service reach the largest number of stakeholders under conditions where risk is minimized and interoperability maximized.

One of the main objectives of providing spatial visualization services is to encourage the evolution from the recent (and traditional) Internet mapping model where a system is built that physically warehouses, maintains, and establishes styles of visual expression for all of the geographic layers that comprise a final mapping visualization system. The newly evolving model takes into account the heterogeneous networked environment and utilizes open standards for spatial visualization to reduce the responsibility and workload of each individual data provider such that they need only provide service interfaces to just the data layers for which they have a mandate to maintain. The final mapping visualization product is one which integrates multiple map services from disparate networked locations and overlays them in a single client visualization platform. In the evolving model, the burden of storage, maintenance, and even the styles for layer expression rests only with the owners of the data.

There are great benefits to the visualization technology that CEHTP is developing. These services implement standards that were reached by global consensus, ensuring product viability and interoperability. System owners and their stakeholders will find hidden information in environmental health data products as spatial visualizations bring to light patterns within their data that could lead to hypotheses about underlying processes and causes of distribution. CEHTP assumes the major burden of developing visualization services, while system owners enjoy the benefit of controlling services after they are deployed. There are multiple web application deployment scenarios that are available to clients, and most implementations require little, if any, technology investment.

3.6.B. Architecture

The architecture of CEHTP-envisioned spatial visualization services implements the WMS (OGC, 2004) specification. This is generally described in the specification as

A Web Map Service (WMS) produces maps of spatially referenced data dynamically from geographic information. This International Standard defines a "map" to be a portrayal of geographic information as a digital image file suitable for display on a computer screen. A map is not the data itself. WMS-produced maps are generally rendered in a pictorial format such as PNG, GIF or JPEG, or occasionally as vector-based graphical elements in Scalable Vector Graphics (SVG) or Web Computer Graphics Metafile (WebCGM) formats.

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

This International Standard defines three operations: one returns service-level metadata; another returns a map whose geographic and dimensional parameters are well-defined; and an optional third operation returns information about particular features shown on a map. Web Map Service operations can be invoked using a standard web browser by submitting requests in the form of Uniform Resource Locators (URLs). The content of such URLs depends on which operation is requested. In particular, when requesting a map the URL indicates what information is to be shown on the map, what portion of the Earth is to be mapped, the desired coordinate reference system, and the output image width and height. When two or more maps are produced with the same geographic parameters and output size, the results can be accurately overlaid to produce a composite map. The use of image formats that support transparent backgrounds (e.g., GIF or PNG) allows underlying maps to be visible. Furthermore, individual maps can be requested from different servers. The Web Map Service thus enables the creation of a network of distributed map servers from which clients can build customized maps.

This International Standard applies to a Web Map Service that publishes its ability to produce maps rather than its ability to access specific data holdings. A basic WMS classifies its geographic information holdings into "Layers" and offers a finite number of predefined "Styles" in which to display those layers. This International Standard supports only named Layers and Styles, and does not include a mechanism for user-defined symbolization of feature data.

In the case of basic WMS GetMap requests, CEHTP will expose the most common aggregations of environmental health data in space and time over a finite and limited set of layer specifications. For example, a basic pesticide use WMS will provide annual aggregates of all pesticides within townships and sections for the whole State and for individual counties. This would deconstruct to perhaps a few dozen layers, a statewide layer and one layer for each county, that can be documented easily and according to the WMS specifications in a GetCapabilities document. However, in more complex scenarios, CEHTP will support customized dynamic crop site and pesticide formulation-specific aggregations for different intensity metrics (poundage, application frequency, and application rates). The number of layers quickly increases to hundreds of thousands of possibilities to support this requirement. Although these "custom" or "dynamic" WMS implementations will still provide documentation to the range of choices allowed for breaking out spatial visualizations, clients that can only support a GetCapabilities document with a finite [and limited] choice of WMS layer specifications will likely fault.

Another important aspect of the spatial visualization architecture is the ability to combine services and layers from multiple locations or domains. The specifications for satisfying this requirement still follow the WMS specification, only in this scenario we see support for recursion where a WMS can call a WMS can call a WMS and so on. The key in supporting this, technologically speaking, is in creating a "meta-WMS" in which the LAYERS parameter of a WMS is itself layers from other WMS servers. Once again, the problem in conforming to the WMS specification is properly documenting the LAYERS in the GetCapabilities document. If the LAYERS are themselves WMS, the GetCapabilities document would hold a series of references to other WMS GetCapabilities documents. There is a companion OGC specification to WMS, called the Web Map Context Specification, which assists clients in retrieving descriptive context information about servers and layers involved in the construction of multi-source map images.

3.6.C. Development Plan

The development of “static” and “dynamic” map service visualizations requires a priori aggregation/summarization of individual environmental health datasets. This involves the deliberate development of query structures that take into account dataset-specific schema and consistently result in a tabular structure on which the WMS can depend. For example, a traffic map service requires development of an RDBMS stored procedure to aggregate individual traffic volume monitoring events into monthly and annual average daily traffic volume metrics.

The development of dataset-specific “static” Web Map Services is accomplished using ArcIMS or, the open source equivalent, University of Minnesota MapServer. For ArcIMS implementations, ArcXML map configuration files are developed for extending a typical ArcIMS map service as a WMS using the ArcIMS/WMS Connector. MapServer/WMS implementations allow for the embedding of existing remote WMS. In some cases this will be a desired configuration, however, it is unclear whether MapServer’s “Mapfile” configuration file can allow for the accurate documenting of recursed WMS layers.

The development of “dynamic” or “real-time custom” WMS is accomplished through utilization of the Java/ArcIMS Connector API, Java 2 Enterprise Edition (J2EE/Servlets), and GeoTools Java GIS toolkit. The Java Connector is harnessed to build dynamic layers, queries, and symbolized layer renderings at request time. Servlets accept HTTP requests, parse query parameters, invoke Java Connector operations, accept ArcIMS output, and stream image output to client. GeoTools is applied for conformance to WMS request parameters and exceptions.

The development of meta-WMS is accomplished through utilization of J2EE Servlets, GeoTools, and Java Advanced Imaging (JAI). Like dynamic WMS implementation, Servlets and GeoTools are used to manage HTTP request/response flows and conformance to WMS specifications. JAI is used for mosaicking multiple transparent and/or opaque WMS output images and producing a single output image. Threading is used so that requests to map services at different locations occur simultaneously and such that the longest client response wait time is not the sum of all image production times, but the length of the longest single image production time. Finally, a framework for server-side versioned caching/storing of tiled output from single WMS servers is required to improve processing time.

3.6.D. Deployment Plan

Because spatial visualization services closely depend on the data that comprise the service, it is CEHTP’s primary objective to migrate deployment of dataset-specific services to system owners’ domains. CEHTP will direct resources towards ensuring that system owners’ maintain control over their visualization services. This also should be particularly attractive to system owners,

because they can control aggregation schemes and the graphical symbolization of layers to meet the needs of stakeholders beyond just environmental health tracking.

Since WMS is a specification that encourages interoperability, the universe of possible client implementations is too large to enumerate. CEHTP, however, will concentrate activities on integrating WMS content into existing environmental health-related meta-portal sites. For example, the Pesticide Action Network hosts a pesticide use dissemination web application called PesticideInfo.org. While this site allows for dynamic querying of tabular- and chart-based visualizations, there is a lack of dynamic mapping visualizations. During Tracking planning phases there has been considerable effort undertaken to integrate a pesticide use WMS in the existing PesticideInfo.org visualization architecture. This model will be repeated for other dataset-specific visualizations and the meta-portals that handle that content.

But for this process to move smoothly, the client tools available for consuming a WMS must be robust, user-friendly, and easy to integrate. AJAX-based mapping clients similar to Google Maps viewer stand as worthy candidates for handling this task. The Google Maps API is particularly attractive, because it can simply consume a WMS endpoint, while simultaneously overlaying it with Google basemap and satellite imagery with ease.

3.7. Metadata Services

3.7.A. Overview

Environmental health tracking content and services must be documented in a way that maximizes their usability. Usability in this case is an indicator of knowledgeable use and ease of discovery. Users would like the ability to specify certain search parameters and view the possible search results over the EPHTN. Before utilizing the resource itself, users need to have the ability to view information about the resource to understand its purpose, schema structure, completeness and accuracy, and use/access restrictions. Metadata is the key to documenting this information and metadata services are the tools by which the metadata content is created, edited, queried, and disseminated. The CEHTP is devising an architecture and infrastructure that allows data owners and other Tracking stakeholders to interact with Tracking metadata and ensures that it reaches as many audiences and access mediums as possible.

3.7.B. Architecture

There are two key components to metadata services architecture: storage site and access pathways. CEHTP envisions a physical storage structure that supports both federated, centralized, and hybrid distributed storage scenarios. This range of physical storage architectures is not determined by a need to support overly complex systems. It is based on the fact that the entities who maintain metadata currency or accuracy are not always in the same location, especially when environmental health tracking datasets start inheriting attributes that are increasingly derived from interconnected integrations. Depending on the dataset or service in question, physical storage architecture for gold-standard metadata will include

1. metadata documents or databases that are referenced from repositories in remote locations,
2. metadata content that is stored in a centralized CEHTP repository
3. content that is partially centralized and partially dispersed

The means of access to metadata content, or the mechanisms that control creation, update, and query, follow an n-tiered (multiple database, application, and presentation tiers) web services model with session tracking and role-based access.

3.7.C. Development Plan

The creation of metadata content is relatively straightforward for datasets, since a standard template has already been established for the EPHTN. Initially CEHTP will query partner surveillance systems to collect any existing metadata content that follows a standard electronically consumable structure. These entries will then be checked for harmonization to the EPHTN template, performing transformations when feasible. For those that do not already exist, metadata content will be key-entered in coordination with partner data systems.

Metadata content creation for CEHTP services is less straightforward, because no standard EPHTN template currently exists. This is complicated by the fact that not all services have been developed or deployed, nor has there been an effort at the national scale to determine functional interface specification for EPHTN services, particularly in the area of geocoding and spatiotemporal linkage services. CEHTP will take a more active role in soliciting coordination among Tracking grantees and the CDC to solve these difficult problems. In the meantime, CEHTP will work towards creating service-based pilot metadata templates.

Whether metadata content describes a dataset or a service, there are 5 operations that must be supported:

CALIFORNIA ENVIRONMENTAL HEALTH TRACKING PROGRAM

1. Authenticate – Determine the user for establishing session- and role-based customization
2. Create – Method which initiates the metadata document
3. Update – Method which allow the editing of an existing metadata document
4. Distribute – Method which pushes a metadata document to associated remote repositories
5. Query – Method which returns zero or more metadata documents, depending submitted query parameters. Query searches include support for geographic parameters.

The service must also support rule-based triggers that alert stakeholders of new entries or edits to existing items. Rules are based on geography, time, subject area, or any other relevant search category.

3.7.D. Deployment Plan

In similar ways as other CEHTP services, the primary factor to ensure the usability of metadata services is in deploying client tools that reach the widest audience possible. A browser-based web API that requires few, if any, plug-in installations and requires little technical expertise in implementing is one that will have the most success. Implementations that make use of AJAX or utilize similar methodologies are likely to give rise to the most successful deployment scenarios.

Also central to the implementation of metadata services is the coupling of its interfaces with methods from other CEHTP services. All other network technologies depend on the tight integration of metadata services. For example, there are two complex types (ToolMetadata and DefinitionMetadata) within the spatiotemporal linkage interface specification that depend on interoperability with Metadata services.

4. REFERENCES

Open Geospatial Consortium, “OpenGIS® Implementation Specification for Geographic information – Simple feature access – Part 1: Common architecture”, 11/30/2005.

Open Geospatial Consortium, “OpenGIS® Implementation Specification for Web Feature Service Implementation Specification”, 5/3/2005.

Open Geospatial Consortium, “OpenGIS® Web Map Context Implementation Specification”, 5/3/2005.

Open Geospatial Consortium, “OpenGIS® Web Map Service (WMS) Implementation Specification”, 8/2/2004.

**5. APPENDIX - AUTOMATED
SPATIOTEMPORAL LINKAGE:
VISION, REQUIREMENTS, AND
ARCHITECTURE (VERSION 3)**