FINAL REPORT
Acoustic Metadata Management and Transparent Access to Networked Oceanographic Data Sets

Marie A. Roch
Dept. of Computer Science
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-7720
phone: (619) 594-5830 fax: (619) 594-6746 e-mail: marie.roch@sdsu.edu

Simone Baumann-Pickering & John A. Hildebrand
Scripps Institution of Oceanography
University of California at San Diego
9500 Gilman Dr.
La Jolla, CA 92093-0205
phone: (858) 534-7280 fax: (858) 534-6849 e-mail: sbaumann@ucsd.edu
phone: (858) 534-4069 fax: (858) 534-6849 e-mail: jhildebrand@ucsd.edu

Catherine L. Berchok
Alaska Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
Department of Commerce
7600 Sand Point Way N.E., Building 4
Seattle, Washington 98115-6349
phone: (206) 526-6331 fax: (206) 526-6615 e-mail: Catherine.Berchok@noaa.gov

Erin M. Oleson
Pacific Islands Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
Department of Commerce
1601 Kapiolani Blvd, Ste. 1110
Honolulu, Hawaii 996814
phone: (808) 944-2172 fax: (808) 941-0307 e-mail: Erin.Oleson@noaa.gov

Sofie M. Van Parijs
Northeast Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
Department of Commerce
166 Water Street
Woods Hole, MA 02543
phone: (508) 495-2119 fax: (508) 495-2258 e-mail: sofie.vanparijs@noaa.gov

Award Number: N00014-11-1-0697 / N0001414IP20032, N0001413IP20029, N0001412IP20059
http://tethys.sdsu.edu

LONG-TERM GOALS

The use of bioacoustics to detect, classify, localize, and establish density estimates of marine fauna provides a cost-effective complement or alternative to visual-based methods of study for monitoring and mitigation (Mellinger et al., 2007; Marques et al., 2011) and considerable resources have been invested in the development of bioacoustic analysis methods to accomplish this. As the number of available recordings grows, the ability to manage information derived from these recordings (metadata of the recordings) becomes crucial in order to combine data across disparate studies to provide information at temporal and spatial scales that are meaningful with respect to oceanic, atmospheric, and anthropogenic processes that may affect the health and productivity of various animal stocks. Of particular importance for bioacoustic metadata is the specification of the how the metadata were generated. The period over which effort was invested may not be the same as that of the acoustic data itself. Examples of this include gaps due to instrument failure, analysis of targeted time periods, etc. In addition to specifying the period, the methods used to analyze the acoustic data must be documented in a way that permits scientists to make intelligent decisions about when acoustic metadata from different studies can be combined and when they should not.

In this report, we provide an overview of a set of metadata structuring rules called the Tethys Metadata Schemata that are designed to be consistent yet extensible. Consistency is a clear prerequisite to combining information from multiple studies. This must be balanced with the ability to record new parameters. A researcher studying a specific call type may realize that there are nuances to the call that can be associated with properties such as kinship or other meaningful distinctions (e.g. in birds, see Akçay et al., 2014). Consequently, it is necessary to maintain a balance between consistency and extensibility, and the philosophy of this project is to provide structure whenever possible while allowing for new types of information to be stored in a manner that can become standardized should their use become widespread.

In order to be useful the metadata structuring rules require software to implement them. We have developed an implementation called Tethys Metadata Workbench. It is a client-server model that permits groups of researchers to install a server program that lets individual users store and retrieve acoustic metadata. Tethys metadata servers are currently running at Scripps Institution of Oceanography, The National Oceanic and Atmospheric Administration (NOAA), and Cornell University. In addition to providing database services, the Tethys metadata server also provides access to oceanographic data sets in a consistent manner.

OBJECTIVES

The objectives of this effort are to produce:

1. A database which can flexibly store multiple types of acoustic metadata derived from a variety of acoustic platforms, both stationary and mobile.
2. Standardization of methods to make the data repositories useful to the passive acoustic monitoring community.

3. Access to network available data products in a standard manner (e.g. ephemeris).
4. Secure access on network platforms using industry standard security protocols.
5. Query and visualization primitives in selected analysis and modeling languages (e.g. Matlab, R) for efficient manipulation of spatial-temporal data.
6. Demonstration projects to show the value of the database as a scientific workbench component.

APPROACH

Our approach is broken down into the development of schemata for representing acoustic metadata and the development of a software implementation of the schemata that also addresses incorporating other types of data sources.

Tethys Schemata
Bioacoustic metadata require context. One might need to know what kind of call was made by which species, where and when it was made, what effort was made in making, detecting, or localizing the call, the methods used, etc. Frequently, this can be seen as a network of heterogeneous data that are related to one another through linkages. One example of this can occur when a comparing the detection of a tonal call which records a sequence of time and frequency parameters to a pulsed call (Figure 1, Roch *et al.*, submitted). While the pulsed call would contain some of the same parameters, a chain of time/frequency points would be inappropriate and other data might be recorded. This type of data network lends itself well to representations that can explicitly capture the network linkages between the data types. While the last forty some odd years of data storage has been dominated by relational databases (see Codd, 1970 for a discussion of the advantages of relational models), the advent of large, heterogeneous datasets has led researchers to consider alternative models of data representation (e.g. Chang *et al.*, 2008; Leavitt, 2010).

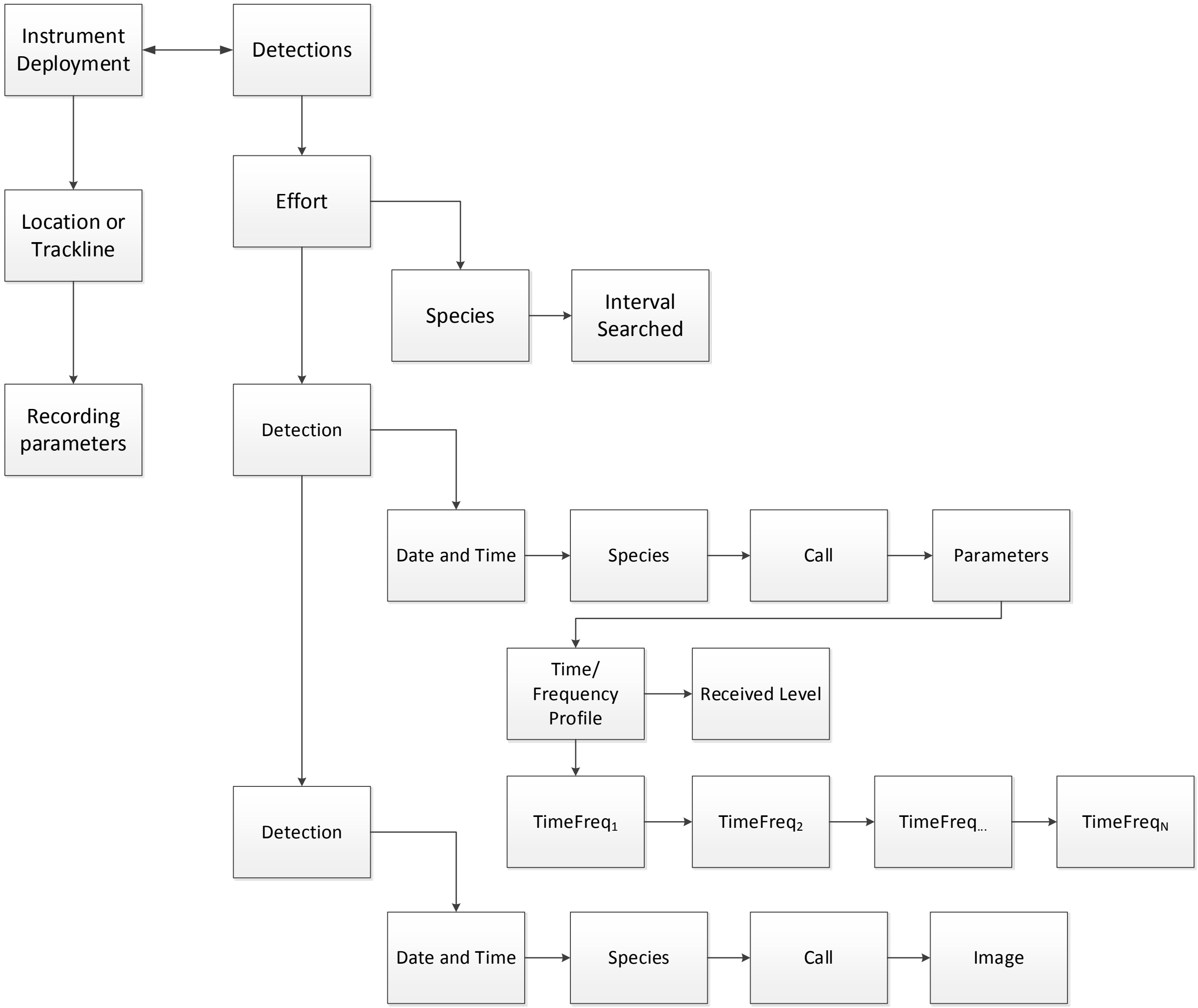

Figure 1 - A network data model provides linkages between different types of data

We use extensible markup language (XML) as a means of representing our metadata. XML is a data annotation method that provides a hierarchical encapsulation of data (Connolly *et al.*, 2007) by enclosing data with structuring elements. Elements are simply pairs of names in angled brackets before and after the data with a leading slash on the latter instance:

<*Call*> boing </*Call*>

Throughout this report, we italicize element names that appear in the main text, but this is only for emphasis and is not required by the XML specification. Elements can be nested:

<*Detection*><*Call*> boing </*Call*><*Species*>…</*Species*> </*Detection*>

and relationships between elements are implied from the nesting structure or via explicit network paths that locate specific elements within XML documents.

In order to provide consistency between disparate detection, classification, and localization efforts, there is a need to provide standardized element names and data formats whenever possible. The XML schema specification (Walmsley, 2002) provides a mechanism to do so. We develop schemata for several concepts related to bioacoustic data:

Table 1. Tethys schemata categories

Schema	Description
Deployments	Characteristics of instrument deployments: when and where they are deployed, how they are configured.
Detections	Descriptions of methods and effort to find biotic and abiotic signals as well as what was detected and characterizations of the detections associated with one or more deployments.
Localizations	Descriptions of methods to find spatial information (e.g. bearing angles, three dimensional location, etc.) from one or more hydrophones.
Ensembles	Groupings of deployed instruments that can be used together in multiple hydrophone applications such as beam forming, localization, etc.
TransferFunction	Descriptions of acoustic instrument calibrations tied to a specific sensor, preamplifier, or instrument.

An important aspect of the XML schema specification is that it can provide for extensibility. The Tethys schemata take advantage of this by strategically placing rules that allow for arbitrary element trees in certain portions of the XML document. Examples of this include the ability to provide new call parameter measurements, specify parameters that are used with detection, classification, and localization algorithm specifications, etc.

A complete review of the schemata are beyond the scope of this report and many details can be found in the Tethys User Manual available at http://tethys.sdsu.edu as well as in (Roch *et al.*, 2013; Roch *et al.*, submitted). To provide a sense of the nature of the schemata, we present the Detections schema at a high level.

Regardless of the schema, each XML document has a top-level enclosing element. For the Detections schema (Figure 2, Roch *et al.*, submitted), this is *Detections*. The stacked squares connecting *Detections* to its children indicate a sequence of elements: *Description*, *DataSource*, *Algorithm*, etc. Mandatory elements are denoted by bold lines. The majority of elements provide structure for child elements (not shown here), such as a group of elements that describe the detection effort. Each element has a data type. With the exception of *UserID*, which has an XML primitive type for alphanumeric data, elements in this figure are Tethys-defined types that are defined elsewhere in the schema.

The optional *Description* element contains children that permit a qualitative description of the goals of this detection effort. *Description* is broken down into children *Objectives*, *Abstract*, and *Method*. This is followed by a *DataSource* element that allows one to uniquely link these detections to a specific deployment or set of deployments (denoted in an ensemble document). The *Algorithm* element permits the specification of the methods used for detection with enough detail to make the effort reproducible. This includes programs used to detect/localize signals (automatically or with analyst assistance), their version, parameter settings, etc.

QualityAssurance contains subelements that specify what quality assurance process was conducted (if any) and contact information for the person responsible for the process. Somewhat related to this, a *UserID* denotes a user identifier of the person who submitted the detection document.

Effort permits the specification of the portion of the deployment that was analyzed. It contains a *Start* and *End* time that must lie within the timespan over which the instrument was deployed. A list of *Kind* elements specifies which species and calls were examined along with a specification of the resolution of call annotations. Three granularities of annotation are allowed:

- call – Each call is recorded individually.
- binned – A time interval is specified in minutes, and detections must fall within specific bins. This is usually used for presence/absence, but the number of calls present within a bin can be recorded if so desired.
- encounter – Detections record acoustic encounters. The start time denotes when the animals were first detected acoustically and the end time indicates when they are no longer producing sound or are no longer within the detection range.

The last two elements, *OnEffort* and *OffEffort* are very similar. Both permit the specification of detections and recording of parameters associated with the detection. Each contain sequences of *Detection* elements (Figure 3, Roch *et al.*, submitted). Detections in the *OnEffort* element must correspond to species, calls, time periods, and granularities specified in the *Effort* element. The *OffEffort* element permits the notation of interesting calls or phenomena that were not searched for systematically.

Children of *Detection* include elements such as the *Start* and *End* times of the call, bin, or encounter, a species identifier from the Integrated Taxonomic Information System (ITIS Organization, 2014), and parameters that describe the call including any user defined ones.

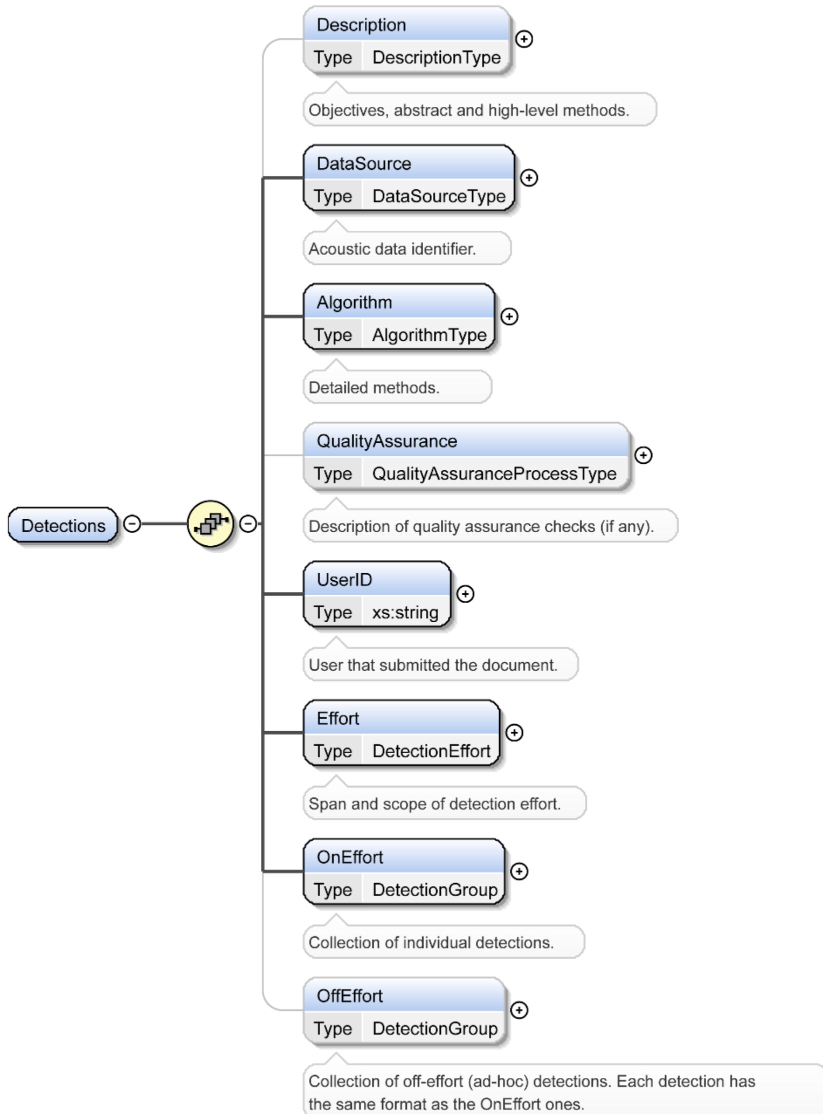

Figure 2 – Top level view of the schema for a detections document.

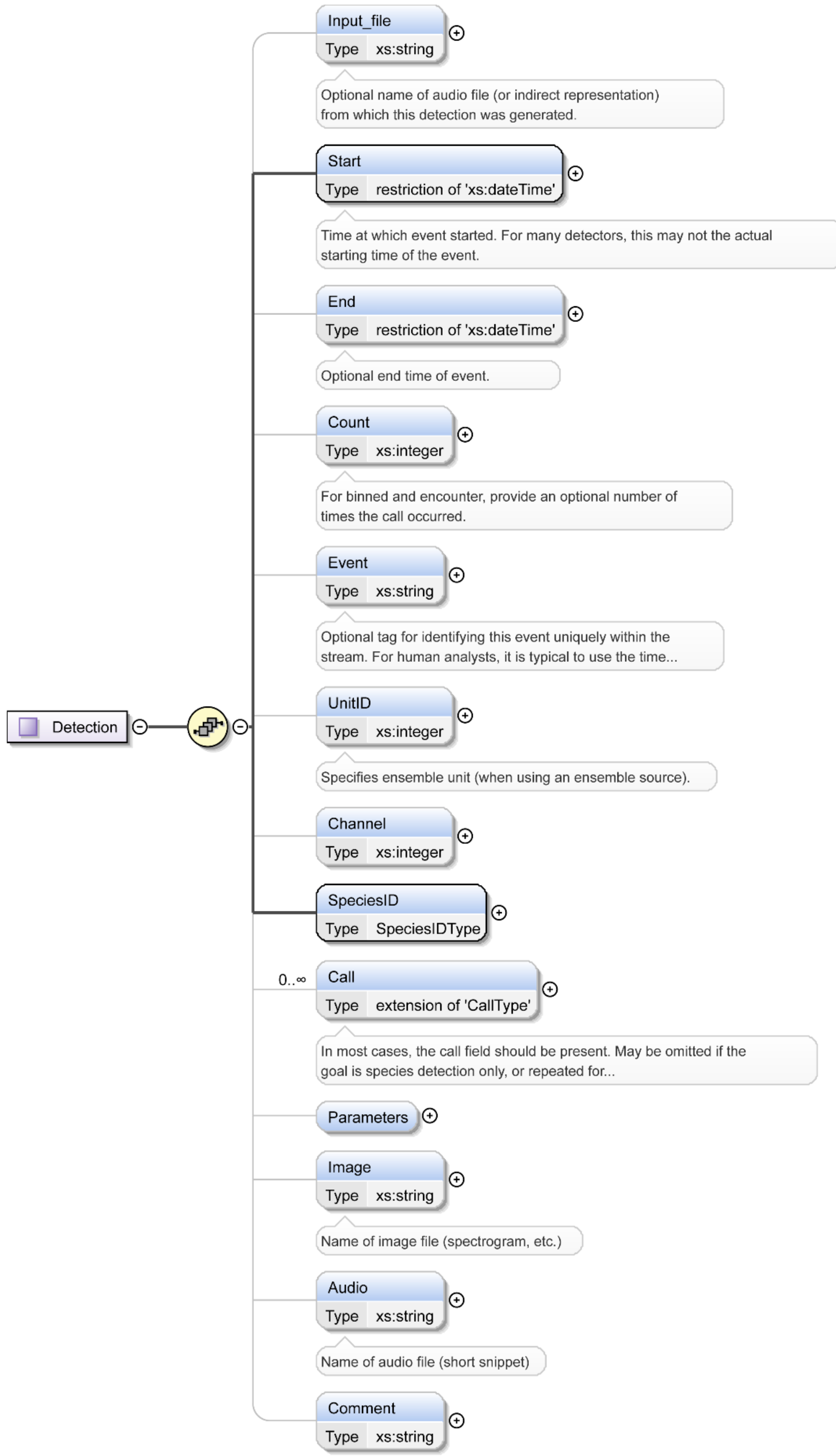

Figure 3 – Schema for individual detections.

Tethys Reference Implementation

The Tethys schemata are open to the public and can be implemented by any vendor. We provide a supported open source reference implementation that is freely available and runs on the Windows operating system (Microsoft Corporation, Redmond WA). Much of the system is portable and could be run on other operating systems with little effort, but the data import module relies heavily on Microsoft-specific technologies for processing Excel spreadsheets and Access databases.

The architecture (Figure 4) is based on a client-server model implementing the RESTful model (Fielding, 2000) which relies on a simple set of http protocol (web) operations between client and server that do not require the server to retain information about a client's state. The server's controller and data transport modules are implemented in Python using the open source CherryPy web framework (CherryPy Team, 2014). A security module permits the use of encrypted data transmission between the server and client programs.

Acoustic metadata are stored using the Berkeley dbxml product, an open source XML database maintained by Oracle Corporation (Redwood Shores, CA). While other database vendors were considered, Berkeley dbxml provided a good combination of performance (Manegold, 2008) balanced with the stability of a large and well known leading database developer. Like most XML databases, the XQuery language (Walmsley, 2006) is used to interrogate the database through the RESTful network interface.

Clients were developed to query and import data from a variety of languages: Matlab, Java, and Python. A primitive R client has been developed, but development effort focused on other areas in which the user community were more interested such as more sophisticated methods of importing data and the addition of quality assurance support. The Matlab client implements methods to enable users to query the database without learning XQuery and provides several visualization tools.

While unrelated to the Tethys schemata, the server implementation supports an architecture for importing data from other sources and providing it back to the user in XML as if it were part of a Tethys database. These modules mediate between Tethys and other data services, and two mediators have been implemented:

1. Ephemeris server – A mediator permits the retrieval of sun and moon positions, illumination, sun/moon rise/set events, etc. The mediator connects to The National Aeronautics and Space Administration (NASA) Jet Propulsion Lab's Horizons service (Giorgini *et al.*, 1996).
2. Oceanographic server – A mediator interfaces with The NOAA Environmental Research Division's Data Access Program (Simons, 2011) which enables access to a wide variety of data products such as the NOAA Tropical Atmospheric Ocean buoys or NASA's Ocean Color.

Data can be added to the database from a wide variety of sources such as spreadsheets, XML documents, or database queries. Translation is accomplished via an XML specification that maps field names from the user's analysis tool to those expected by Tethys. When importing acoustic metadata from databases, sophisticated queries are possible that include referencing the context of an earlier query. For users developing new tools, an application programming interface is provided that can permit the generation of XML documents directly, thus skipping the need for translation.

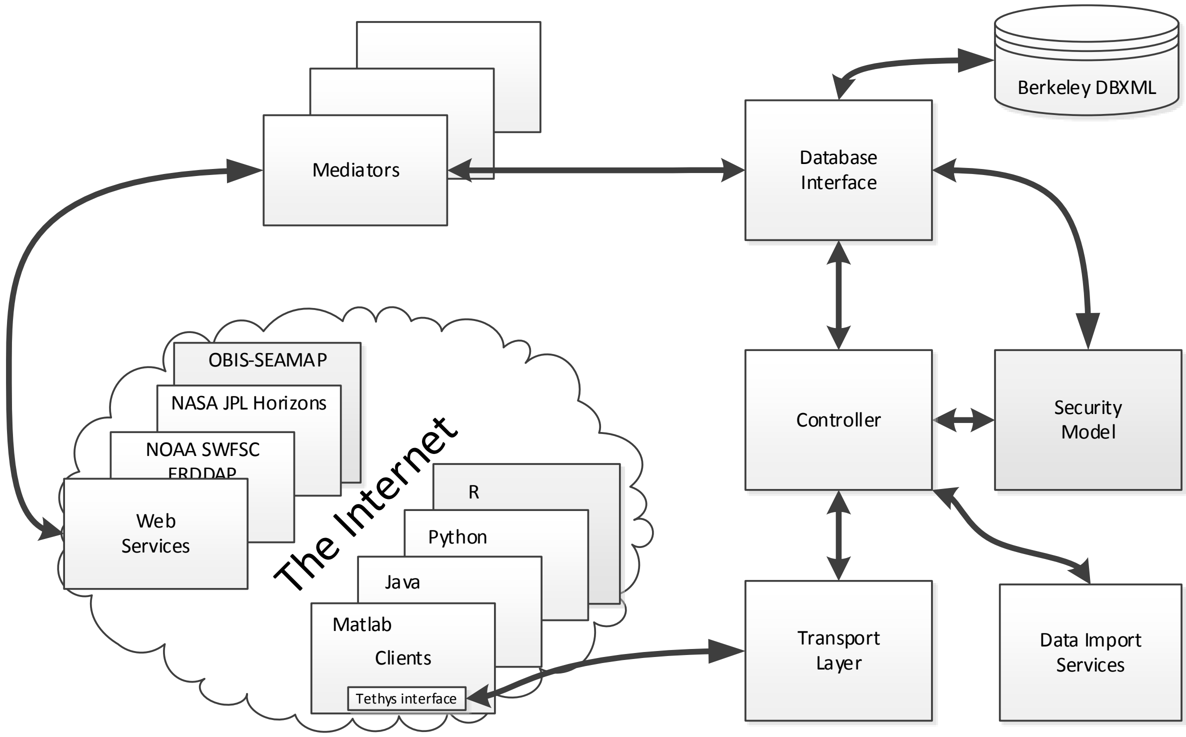

Figure 4 – Tethys reference implementation architecture

WORK COMPLETED

Version 2.2 Tethys schemata and implementation have been released on the project web site. Major improvements made in the last year came fom the output of the final Tethys workshop and include significant enhancements to the import facilities permitting more sophisticated data import (nested queries), the incorporation of the ability to represent quality assurance processes within the schemata, a National Center for Environmental Information trial with NOAA Northeast and Alaska Fisheries Science Centers to use Tethys deployment metadata in archiving Fisheries Science Center acoustic data, and experiments demonstrating the ability of the system to represent metadata in other domains.

RESULTS

The Tethys metadata system is beginning to gain traction with users outside of the principal investigators. Peter Wrege and Sara Keen (Cornell University Bioacoustics Research Program) are using the system for forest elephants (*Loxondonta cyclotis*) and Cornell plans on developing front-end graphical user interfaces for the system. Jasco Ltd. announced at the 2015 International Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals that they planned on building a Tethys interface into their visualization system. The Tethys metadata system has been described in an IEEE Oceans paper (Roch *et al.*, 2013) and an expanded journal-length manuscript representing the most recent developments is under review (Roch *et al.*, submitted).

The strength of this system is the type of questions that one can ask when one has an analytical engine that can automate the integration of acoustic metadata with environmental information. The system has permitted spatio-temporal analsysis of beaked whales across the Pacific revealing possible acoustic signatures for several species of beaked whales (Baumann-Pickering *et al.*, 2014) and revealed spatial and temporal patterns in habitat use for fin and blue whales (Širović *et al.*, 2015). The abiltity to track details of equipment such as calibration curves proved useful in a study that examined performance degradation of species identification algorithms in the face of equipment and site differences and proposed techniques to mitigate for this (Roch *et al.*, 2015). Other studies that will use this system for analyzing marine mammals with respect to oceanographic conditions and anthropogenic sources (e.g. sonar, habitat models) are underway and are expected to produce additional Tethys-enabled publications.

IMPACT/APPLICATIONS

The Tethys Metadata Workbench has been used to represent over 300 years of detection effort across multiple species and many deployments, recording millions of detections in the labs of the authors. Visualization capabilities permit the exploratory data visualization, at times making patterns or the lack thereof easy to detect. It has been used in the production of scholarly journal publications and reports to the US Navy.

TRANSITIONS

This project has matured to the point that it is being transitioned to funding by US Navy Living Marine Resources.

RELATED PROJECTS

N39430-15-C-1712 – Tethys, a workbench for passive acoustic monitoring metadata. PI Marie Roch, Simone Baumann-Pickering, Ana Širović. This new start represents a transition of the current project towards Fleet use.

N00014-15-1-2299 – Unsupervised learning (clustering) of odontocete echolocation clicks. PI Marie Roch, Simone Baumann-Pickering, Margareta Ackerman. Project uses Tethys for maintaining acoustic metadata information.

ONR N00014-13-IP20051– Advanced Methods for Passive Acoustic Detection, Classification, and Localization of Marine Mammals. PI Jonathan Klay, Dave Mellinger, Dave Moretti, Steve Martin and Marie A. Roch. Some of the work in this grant makes use of Tethys and has overlapping key personnel.

N00014-12-1-0273 – Modeling of Habitat and Foraging Behavior of Beaked Whales in the Southern California Bight, PI John Hildebrand, Simone Baumann-Pickering – The work performed in this grant makes use of Tethys and has overlapping key personnel.

N000141210904 – Blue and fin whale habitat modeling from long-term year-round passive acoustic data from the Southern California Bight, PI John Hildebrand, Ana Širović. – The work performed in this grant makes use of Tethys and has overlapping key personnel.

N000141310641 – ESME workbench enhancements – PI David Mountain – ESME provides acoustic modeling, simulated animal movements, and environmental data visualization.

NSF-OCE-11-38046 – OBIS-SEAMAP, PI Patrick N. Halpin – OBIS-SEAMAP collects visual and acoustic detection information for marine mammals, sea birds, and sea turtles. We have worked with Ei Fujioka to integrate acoustic detections into their platform to permit transfer of data summaries from Tethys to OBIS-SEAMAP.

REFERENCES

Baumann-Pickering, S., Roch, M. A., Brownell Jr, R. L., Simonis, A. E., McDonald, M. A., Solsona-Berga, A., Oleson, E. M., Wiggins, S. M., and Hildebrand, J. A. (2014). "Spatio-Temporal Patterns of Beaked Whale Echolocation Signals in the North Pacific," Plos One 9(1), pp. 17, doi:10.1371/journal.pone.0086072.g001.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E. (2008). "Bigtable: A Distributed Storage System for Structured Data," ACM Trans. Comput. Syst. 26(2), 1-26.

CherryPy Team (2014). "CherryPy - A Minimalist Python Web Framework," Accessed July 28, 2014. http://www.cherrypy.org.

Codd, E. F. (1970). "A relational model of data for large shared data banks," Comm. ACM 13(6), 377-387.

Connolly, D., Khare, R., and Rifkin, A. (2007). "The Evolution of Web Documents: The Ascent of XML," World Wide Web J. 2(4), 119-128.

Fielding, R. T. (2000). "Architectural Styles and the Design of Network-based Software Architectures," *Computer Science*, The University of CA, Irvine, Irvine, CA, pp. 162.

Giorgini, J. D., Yeomans, D. K., Chamberlin, A. B., Chodas, P. W., Jacobson, R. A., Keesey, M. S., Lieske, J. H., Ostro, S. J., Standish, E. M., and Wimberly, R. N. (1996). "JPL's On-Line Solar System Data Service," B. Am. Astron. Soc. 28(3), 1158.

ITIS Organization (2014). "Integrated Taxonomic Information System (ITIS)," Accessed July 4, 2014. http://www.itis.gov.

Leavitt, N. (2010). "Will NoSQL Databases Live Up to Their Promise?," Computer 43(2), 12-14.

Manegold, S. (2008). "An empirical evaluation of XQuery processors," Inform. Syst. 33(2), 203-220.

Roch, M. A., Batchelor, H., Baumann-Pickering, S., Berchock, C. L., Cholewiak, D., Fujioka, E., Garland , E. C., Herbert, S., Hildebrand, J. A., Munger, L. M., Oleson, E. M., Van Parijs, S. M., Risch, D., and Širović , A. (submitted). "Management of Acoustic Metadata for Bioacoustics."

Roch, M. A., Baumann-Pickering, S., Batchelor, H., Hwang, D., Sirovic, A., Hildebrand, J. A., Berchok, C. L., Cholewiak, D., Munger, L. M., Oleson, E. M., Van Parijs, S., Risch, D., and Soldevilla, M. S. (2013). "Tethys: a workbench and database for passive acoustic metadata," Oceans 2013, 5 pp.

Roch, M. A., Stinner-Sloan, J., Baumann-Pickering, S., and Wiggins, S. M. (2015). "Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks," J. Acous. Soc. Am. 137(1), 22.

Simons, R. A. (2011). "ERDDAP - The Environmental Research Division's Data Access Program," Accessed Feb 1, 2012. http://coastwatch.pfeg.noaa.gov/erddap.

Širović , A., Rice, A., Chou, E., Hildebrand, J. A., and Roch, M. A. (2015). "Seven years of blue and fin whale call abundance in Southern California," Endang. Species Res. 28, 61-75.

Walmsley, P. (2002). *Definitive XML Schema* (Prentice Hall PTR, Upper Saddle River, NJ).

Walmsley, P. (2006). "XQuery," (O'Reilly, Farnham, UK).

PUBLICATIONS

Baumann-Pickering, S., Roch, M. A., Brownell Jr, R. L., Simonis, A. E., McDonald, M. A., Solsona-Berga, A., Oleson, E. M., Wiggins, S. M., and Hildebrand, J. A. (2014). "Spatio-Temporal Patterns of Beaked Whale Echolocation Signals in the North Pacific," Plos One 9(1), pp. 17, doi:10.1371/journal.pone.0086072.g001. [published]

Roch, M. A., Baumann-Pickering, S., Batchelor, H., Hwang, D., Sirovic, A., Hildebrand, J. A., Berchok, C. L., Cholewiak, D., Munger, L. M., Oleson, E. M., Van Parijs, S., Risch, D., and Soldevilla, M. S. (2013). "Tethys: a workbench and database for passive acoustic metadata," Oceans 2013, 5 pp. [publsihed]

Roch, M. A., Stinner-Sloan, J., Baumann-Pickering, S., and Wiggins, S. M. (2015). "Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks," J. Acous. Soc. Am. 137(1), 22. [published]

Širović , A., Rice, A., Chou, E., Hildebrand, J. A., and Roch, M. A. (2015). "Seven years of blue and fin whale call abundance in Southern California," Endang. Species Res. 28, 61-75. [published]

Roch, M. A., Batchelor, H., Baumann-Pickering, S., Berchock, C. L., Cholewiak, D., Fujioka, E., Garland , E. C., Herbert, S., Hildebrand, J. A., Munger, L. M., Oleson, E. M., Van Parijs, S. M., Risch, D., and Širović , A. "Management of Acoustic Metadata for Bioacoustics." [submitted]