Tethys: A workbench and database for passive acoustic metadata

Marie A. Roch
Dept. of Comp. Science
San Diego State University
San Diego, CA USA

Simone Baumann-Pickering, Heidi Batchelor, Daniel Hwang, Ana Širović, and John A. Hildebrand
Scripps Instititution of Oceanography
University of California, San Diego,
La Jolla, CA USA

Catherine L. Berchok, Danielle Cholewiak, Lisa M. Munger, Erin M. Oleson, Sofie Van Parijs, Denise Risch, and Melissa S. Soldevilla
NOAA National Marine Fisheries Fisheries Service,
Fisheries Science Centers
Silver Spring, MD USA

Abstract: **This project proposes a community standard for the representation of passive acoustic metadata along with a freely available software implementation. Our target audience is the marine mammal community, but the concepts are general and are applicable to a wide variety of taxa. In addition, we address the need to analyze acoustic metadata in the context of other environmental and biological parameters. The implementation provides interfaces to access a wide variety of data available from external services, such solar and lunar rise/set times, sea surface temperature, chlorophyll A, etc., thus permitting extensive data exploration in a workbench environment.**

Keywords—bioacoustic, passive acoustic monitoring, metadata, database

I. Introduction

The bioacoustics community requires new techniques for storing and manipulating information about acoustic detections, classifications, and localizations, and to incorporate ancillary information such as oceanographic and other environmental data. In the ocean environment, the need for such a system can be attributed to the success of passive acoustic monitoring (PAM) of marine sounds over the last two decades. PAM techniques have been used to establish presence/absence, abundance and density estimates of vocalizing animals [1], and for furthering knowledge of the population structure [2], or ecology [3] of these species. In addition, PAM has been successfully used in monitoring and mitigation scenarios for Naval activities [4], petroleum exploration [5], alternative energy projects [6], and other scenarios where anthropogenic activity may result in impacts to the viability of marine mammal and fish populations [7].

With the wide variety of PAM applications, the rate of bioacoustic data acquisition is increasing exponentially as sample rates and endurance increases and costs decrease. Practitioners are collecting acoustic data and generating detection, classification, and localization annotations at a rate that exceed most institutions' ability to process and analyze them. In one of our labs, we currently have over 600 TB of acoustic data and will likely collect an additional 100 TB before the end of 2013. Analysis and retention of metadata derived from these acoustic recordings are frequently organized using researcher specific schemes. On a large scale, such schemes are neither manageable nor do they permit efficient scientific discovery of patterns within the data that may have biological, ecological, and management implications. A structured yet flexible database can address both of these challenges.

Several databases and standards have been proposed over the last several years for various taxa as well as geographic information systems. For physical systems, the Open Geospatial Consortium standards [8] provide methods to describe instruments and measurements. ISO 19115 [9] provides similar capabilities. However, both of these standards present significant barriers to reporting analysis effort or varying kinds of biological signals. The reporting of the amount of time spent searching for a particular type of call, possibly in a subset of the full recording, herein called effort, plays a critical role in the ability to make inferences from detections. Without knowing that one looked for a specific type of call between January and December over a period of three years, one could not make credible statements about seasonality if the calls were only detected in winter months. Neither of these standards permits the reporting of effort.

Data specifications for biological organisms are also missing attributes needed for bioacoustic data. Darwin Core [10] does not currently have any method for specifying bioacoustic metadata. The integrated oceanographic information system (IOOS) also lacks extensions for bioacoustic data. Efforts by Fornwell et al. (Bob Gisiner, personal communication) are extending IOOS for visual line transect data which has some commonality with bioacoustic metadata. Most recently, the ocean biogeographic information system spatial ecological analysis of megavertebrate populations (OBIS-SEAMAP) [11] has developed capabilities to include summary bioacoustic data and we have worked with

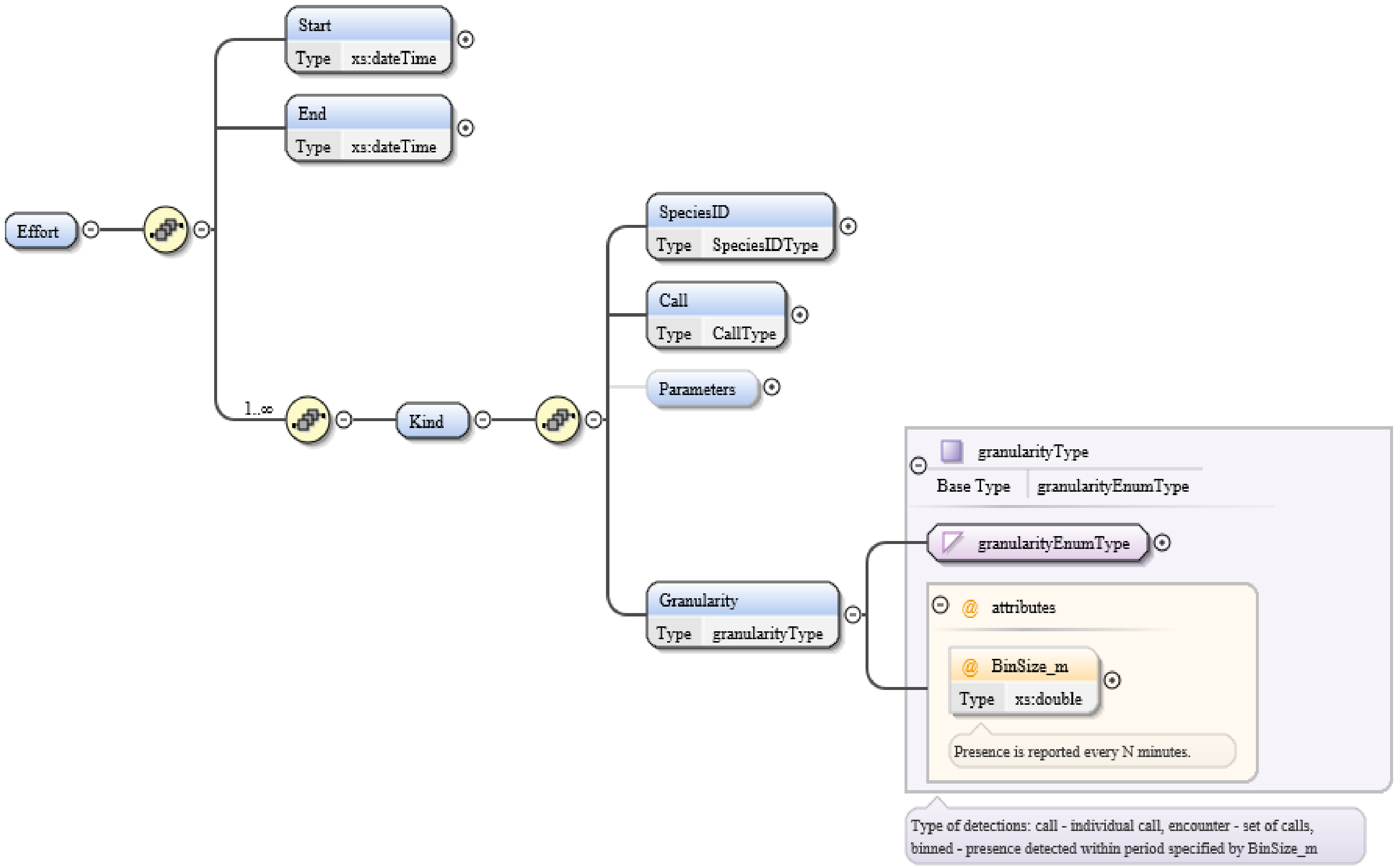

Figure 1 – Specification of detection effort within the Detections schema. Start and End times are followed by a list of one or more Kinds, specifying the species being detected, specific call type (and optional subtype), and an indication of the level of detail (granularity) for the detection. Other elements of the Detections schema indicate the methods used for detection, the data source, and the actual detections themselves. A simplified example of a detection is shown in the text.

Fujioka et al. to help define the vocabulary required for OBIS-SEAMAP for bioacoustic data.

In contrast, the Tethys acoustic metadata system is designed to provide a community standard for representing detections, classifications, and localizations of bioacoustic, ambient, and anthropogenic signals in acoustic data. We standardize common attributes of acoustic metadata, and develop mechanisms for retaining and accessing additional information defined by the scientific user-base of Tethys. It complements systems such as OBIS-SEAMAP by providing a structure for recording fine detail in calls. Summary information can be extracted and exported to OBIS-SEAMAP, a process that will be automated at a later stage of the project.

In addition, Tethys provides interfaces to query and process biological and oceanographic data that are publicly available through web services. Tethys's client-server framework permits science users to work with the data in familiar languages such as Matlab, Java or Python[1]. The integration of PAM metadata with publicly available ecological data sources enables scientists to efficiently answer questions about ecological interactions that would have previously taken weeks or months of work.

II. Data Representation and Storage

Codd's landmark 1970 work [12] moved the database community away from hierarchical and network databases to a more tabular representation that is the core of modern relational database model servers (RDBMS). While RDBMS has many advantages, it is not well suited for loosely formatted data that can have variable structure. A number of projects have turned to the so-called No-SQL databases [13, 14], which offer similar scalability and greater flexibility for data that does not fit in the standard RDBMS format.

One branch of the No-SQL movement is the use of databases that have extended markup language (XML) as their underlying data structure. For efficiency, the data may be decomposed into tables with an underlying RDBMS, but this is done transparently leaving the database designer free to think about an appropriate representation rather than its decomposition. XML permits the structuring of data by text-based elements <denoted with angled brackets> that indicate content, such as this set of elements that we use to denote a B call produced by a blue whale (*Balaenoptera musculus*):

```
<Detection>
  <Start> 2012-01-05T16:00:00 </Start>
  <End> 2012-01-05T16:00:28 </End>
  <SpeciesID> Balaenoptera musculus </SpeciesID>
  <Call> B </Call>
</Detection>
```

[1] An R language interface is in development.

We have proposed a set of XML structuring elements for bioacoustic metadata. The form of an XML document's content can be specified through the use of schema [15]. The constraints provide a list of mandatory and optional fields as well as the format and structure of their content. We have defined schema for the following types of data:

- Deployment – Temporal-spatial and equipment information about an instrument deployment.
- Ensemble – Groupings of deployment instances used together for localization or signal enhancement (e.g. an array).
- Detections – Information about effort, on-effort and off-effort detections for specific deployments or ensembles.
- Localizations – Localization of signals; may be bearings or positions. Linkage to individual detections is permitted.

An example of a schema can be found in a subset of the the Detections schema (Figure 1). Each set of Detections has an Effort element with a start and end time, represented as an ISO8601 time stamp [16]. This is followed by a list indicating which species or phenomena are being examined and the level of detail, or granularity of the reported detection. Valid granularities from most to least specific are 1) call – a specific call, 2) acoustic encounter – reporting the beginning and end time of a series of sounds, and 3) binned – presence/absence during a time period specified by an additional parameter.

XML schemas provide mechanisms that permit extendibility. As an example, the detections schema has an optional <Parameters> element. A number of commonly measured parameters are predefined as well as a <UserDefined> element that may contain any XML elements the user wishes to define. Thus, arbitrary measurements such as a delphinid's whistle time-frequency characterization could be easily stored.

Where possible, we reuse structures from existing standards. The deployments schema borrows elements from the OpenGIS standard, such as structures for representing who should be contacted with respect to a specific instrument deployment or recovery.

III. Mediators

In addition to providing storage of bioacoustic metadata, Tethys provides access to a wide variety of geophysical and biological data sources publicly available through the Internet. The query parser looks for non-native queries which are flagged with an ext: (external) tag and routes these queries to mediator modules that perform the query on the user's behalf. The response is formatted as XML and returned to the user as if the query had been to a local resource.

These modules are referred to as mediators [17]. Mediators have presently been implemented for two web services. NASA Jet Propulsion Laboratory's Horizons ephemeris service [18] provides information about solar and lunar rise/set as well as lunar illumination. NOAA Southwest Fisheries Science Center's ERDDAP project [19] provides access to a wide variety of data products, such as NASA's Ocean Color data products or NOAA's TAO buoy network.

IV. Architecture

Tethys uses a client-server architecture (Figure 2). The server side architecture consists of a set of modules to provide data transport to clients, the database interface, mediators, and the database itself. With the exception of the database, the server-side modules are implemented in Python.

The data transport layer is responsible for communication between clients and the server. It is implemented using XML remote procedure calls (XML-RPC) [20, 21]. Communication can be encrypted using the secure socket layer protocol.

The mediators translate from native queries (see section V) to the appropriate format for each of the supported mediators.

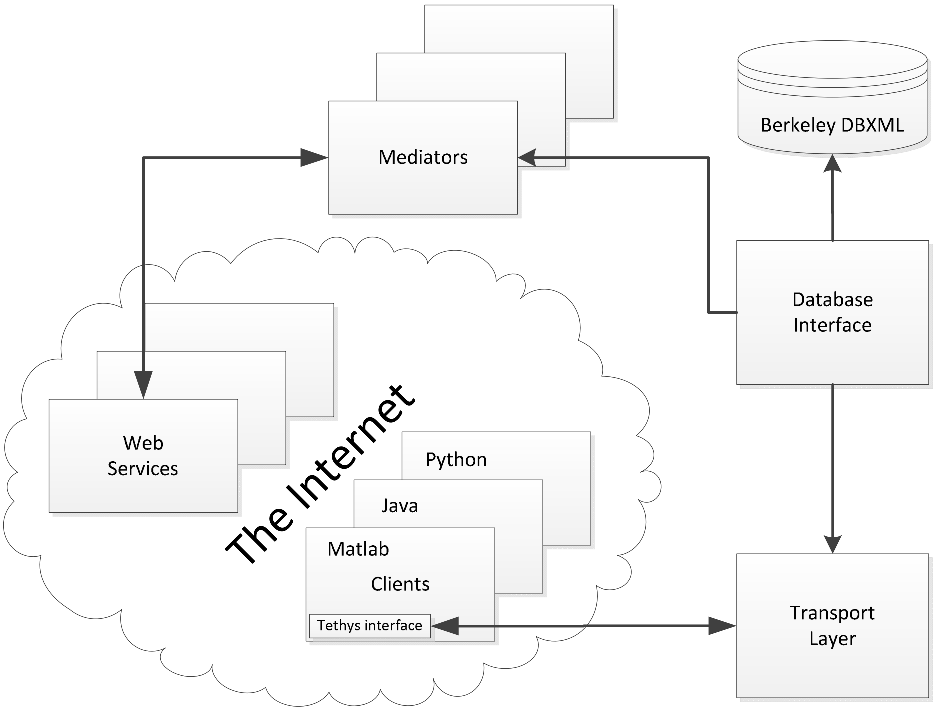

Figure 2 – Tethys client-server architecture. Server modules are responsible for connecting with clients, interfacing with the database store (Berkeley DBXML) and processing requests for data from other Internet services (mediators). Clients can access Tethys from a variety of languages.

The results are translated to XML and placed in a temporary XML document. The query is then updated to reference the temporary document rather than the mediator resource and query processing proceeds as a normal query.

The database module is provided by Oracle's (Redwood Shores, CA) freely available Berkeley DBXML as a native XML database server. Berkeley DBXML was chosen due to its freely redistributable source code, the stability of Oracle corporation, and a study by Manigold [22] that demonstrated Berkeley DBXML to be a capable native XML database that compared well with its peers in benchmark tests.

V. Data Access

Data is accessed via XQuery [23], a functional database language for querying XML. Queries written in XQuery must reflect the schema of the database. Like all network database languages, a more intricate knowledge of the data layout is needed than with a traditional RDBMS where one only needs to know tables, their attributes, and relations.

While very complex queries are best written in XQuery, the majority of the scientific user-base of Tethys is not well-versed in database languages in general, and we provide a wide variety of functions in the Matlab interface that permit users to write moderately complex queries by invoking functions with optional arguments. Each such function uses an XQuery skeleton query whose selection criteria are set based on values provided to the Matlab function in the form of order independent keyword/value pairs in the arguments. As an example, to query effort from deployments 30-35 at a site referred to as "M," the following Matlab function would be invoked:

```
effort = dbGetEffort(queryHandler,
  'DeploymentID', {'>=', 30},
  'DeploymentID', {'<=', 35}, 'Site', 'M');
```

The function substitutes criteria in the appropriate places in the query. In addition to the Tethys documentation, we provide users with a "cookbook" document that provides concrete examples of how to query Tethys.

Data is returned as XML documents that must be parsed by the user (most languages have libraries for processing XML) with the exception of the Matlab interface where functionality is provided to translate into native Matlab data structures.

VI. Data Import

Data can be added to Tethys in a variety of manners. Native XML that conforms to Tethys schema can be added without translation. Import filters are available for comma separated value files, spreadsheets, and databases. The import filters use XML specifications to map from the source material to Tethys schema. Whenever data is added to Tethys, a copy of the source material is retained along with the XML representation, permitting database rebuild in the case of catastrophic failure.

VII. Analysis and Visualization

The Matlab client provides a variety of functions for analysis and visualization. Multiple data sets can be retrieved and superimposed. In Figure 3, acoustic encounters of Risso's dolphins (*Grampus griseus*) near the equator are shown superimposed over shading showing day and night and nighttime lunar illumination. Risso's dolphins are known to be night-time foragers [24] and the associated vocalization pattern is reflected in the visualization. It is also interesting to note that these dolphins appear to be echolocating more when the moon is not present, a pattern that we have seen over longer time scales and at different locations (unpublished data). Functions are also provided to transform detections into presence/absence data matrices, permitting the use of correlation analysis and other statistical techniques.

VIII. Summary

We have established a community standard for PAM data that incorporates a need for structure and flexibility. This includes schema for characterizing deployments, detections, classifications, localizations, and arrays. The use of a community standard will not prevent data summarization to other data stores such as OBIS-SEAMAP and we will be

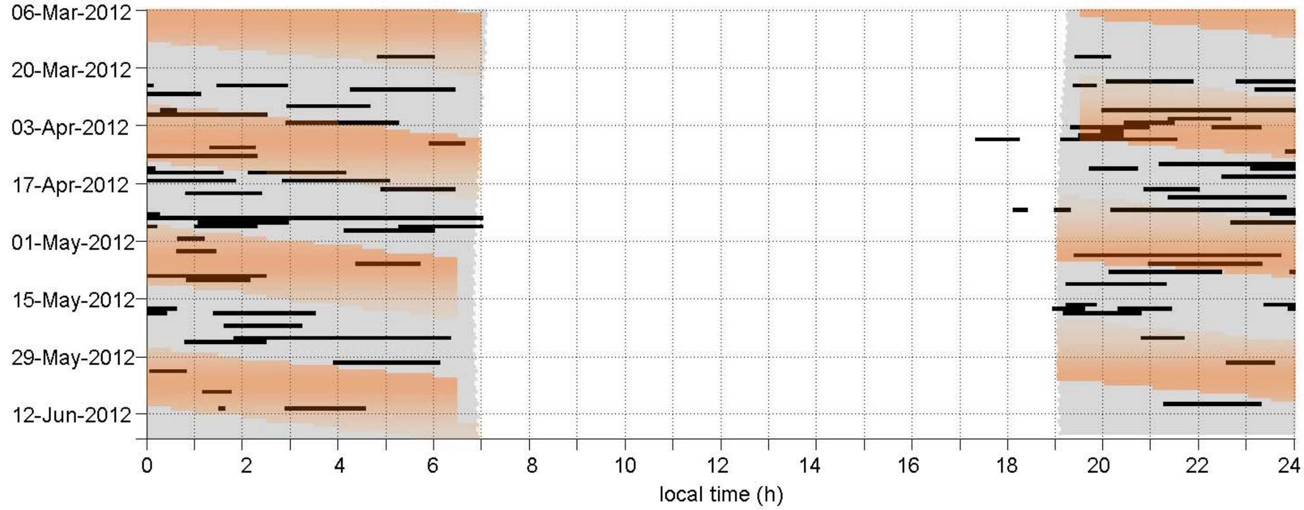

Figure 3 – Daily detections of Risso's dolphin echolocation clicks near the equator between March and May 2012. Grey shading indicates night, and nightly lunar illumination is shown in orange. Black lines indicate detections of echolocation activity which is primarily at night and most frequently during periods of time where lunar illumination is not present. Cloud cover is unknown.

working with Ei Fujioka and Pat Halpin (OBIS-SEAMAP) to establish a data summarization protocol.

In summary, for our labs, Tethys has presented opportunities to address new types of research questions. We have focused on marine mammals, fish, and anthropogenic signals, but the framework is general and could be easily used in other contexts. It has enhanced meta-analyses over large spatial and temporal scales and allows researchers to examine acoustic behavior from an ecological perspective. Tethys is publicly available from http://tethys.sdsu.edu.

ACKNOWLEDGMENTS

The authors wish to thank the National Oceanographic Partnership Program for their support (award N00014-11-1-0697), as well as The Office of Naval Research, Michael Weise, and the Bureau of Energy Management, Jill Lewandowski and Jim Price. Analysis of the Risso's dolphin data was funded by National Geographic. We would also like to thank Ei Fujioka, Bob Gisiner, Shane Guan, Hassan Moustahfid, and Shannon Rankin for productive and helpful conversations as we have developed this project.

References

1. Marques, T.A., et al., *Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales.* J. Acous. Soc. Am., 2009. **125**(4): p. 1982-1994.
2. McDonald, M.A., S.L. Mesnick, and J.A. Hildebrand, *Biogeographic characterisation of blue whale song worldwide: using song to identify populations.* J. Cetacean Res. Mgt., 2006. **8**(1): p. 55-65.
3. Soldevilla, M.S., S.M. Wiggins, and J.A. Hildebrand, *Spatio-temporal comparison of Pacific white-sided dolphin echolocation click types.* Aquat. Biol., 2010. **9**(1): p. 49-62.
4. Širović , A., et al., *Marine Mammal Demographics of the Outer Washington Coast During 2008-2009*. 2011, Naval Postgraduate School: Monterey, CA.Tech. Memo. NPS-OC-11-004CR.
5. Blackwell, S.B., et al., *Effects of airgun sounds on bowhead whale calling rates in the Alaskan Beaufort Sea.* Mar. Mammal Sci., 2013: p. 24.
6. Tougaard, J., et al., *Pile driving zone of responsiveness extends beyond 20 km for harbor porpoises (Phocoena phocoena (L.)).* J. Acous. Soc. Am., 2009. **126**(1): p. 11.
7. Clark, C.W. and D.B. Peters, *Acoustic System Monitors and Mitigates Harm to Marine Mammals in Real Time.* Sea Technol., 2009. **50**(8): p. 10-14.
8. Open Geospatial Consortium. *OGC Standards and Supporting Documents*. 1994 [cited 2010 June 1, 2010]; Available from: http://www.opengeospatial.org/standards.
9. International Standards Organization, *Geographic Information - Metadata*. 2003, International Standards Organization: Geneva.Tech. Memo. ISO 19115:2003, p. 140.
10. Wieczorek, J., et al., *Darwin Core*. 2009, Taxonomic Databases Working Group: London.Tech. Memo. 2009-09-23.
11. Fujioka, E., et al., *Integration of Passive Acoustic Monitoring Data into OBIS-SEAMAP, a Global Biogeographic Database, to Advance Spatially-Explicit Ecological Assessments.* Ecol. Inform., in review.
12. Codd, E.F., *A relational model of data for large shared data banks.* Comm. ACM, 1970. **13**(6): p. 377-387.
13. Chang, F., et al., *Bigtable: A Distributed Storage System for Structured Data.* ACM Trans. Comput. Syst., 2008. **26**(2): p. 1-26.
14. Leavitt, N., *Will NoSQL Databases Live Up to Their Promise?* Computer, 2010. **43**(2): p. 12-14.
15. Walmsley, P., *Definitive XML Schema*. 2002, Upper Saddle River, NJ: Prentice Hall PTR. 528.
16. International Standards Organization, *Data Elements and Interchange Formats - Information Interchange - Representation of Dates and Times*. 2004, International Standards Organization: Geneva.Tech. Memo. ISO 8601:2004, p. 33.
17. Wiederhold, G., *Mediators in the architecture of future information-systems.* Computer, 1992. **25**(3): p. 38-49.
18. Giorgini, J.D., et al., *JPL's On-Line Solar System Data Service.* B. Am. Astron. Soc., 1996. **28**(3): p. 1158.
19. Simons, R.A. *ERDDAP - - The Environmental Research Division's Data Access Program*. 2011 [cited 2012 Feb 1]; Available from: http://coastwatch.pfeg.noaa.gov/erddap.
20. Winer, D. *XML-RPC Specification*. 1999 [cited 2011 June 15]; Available from: http://xmlrpc.scripting.com/spec.html.
21. St. Laurent, S., J. Johnston, and E. Dumbill, *Programming Web Services with XML-RPC*. 2001, Cambridge, MA: O'Reilly Media.
22. Manegold, S., *An empirical evaluation of XQuery processors.* Inform. Syst., 2008. **33**(2): p. 203-220.
23. Walmsley, P., *XQuery*. 2006, O'Reilly: Farnham, UK.
24. Soldevilla, M.S., S.M. Wiggins, and J.A. Hildebrand, *Spatial and temporal patterns of Risso's dolphin echolocation in the Southern California Bight.* J. Acous. Soc. Am., 2010. **127**(1): p. 124-132.