Deep-Sea Research II 56 (2009) 1804–1811

Contents lists available at ScienceDirect

Deep-Sea Research II

journal homepage: www.elsevier.com/locate/dsr2

Bringing together an ocean of information: An extensible data integration framework for biological oceanography

Karen I. Stocks *, Chris Condit, Xufei Qian, Paul E. Brewin, Amarnath Gupta

San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, MC 0505, La Jolla, CA 02093-0505, USA

ARTICLE INFO

Available online 28 May 2009

Keywords:
Seamounts
Information systems
Marine ecology
Biodiversity
Data processing
OBIS

ABSTRACT

As increasing volumes and varieties of data are becoming available online, the challenges of accessing and using heterogeneous data resources are growing. We have developed a mediator-based data integration system called Cartel for biological oceanography data. A mediation approach is appropriate in cases where a single central warehouse is not desirable, such as when the needed data sources change frequently through time, or when there are advantages for holding heterogeneous data in their native formats. Through Cartel, data sources of a variety of types can be registered to the system, and users can query against simplified virtual schemas, without needing to know the underlying schema and computational capabilities of each data source. The system can operate on a variety of relational and geospatial data formats, and can perform joins between formats. We tested the performance of the Cartel mediator in two biological oceanography application areas, and found that the system was able to support the variety of data types needed in a typical ecology study, but that the response times were unacceptably slow when very large databases (i.e. Ocean Biogeographic Information System and the World Ocean Atlas) were used. Indexing and caching are currently being added to the system to improve response times. The mediator is an open-source product, and was developed to be a generic, extensible component available to projects developing oceanography data systems.

1. Introduction

1.1. The need for data integration

Biological oceanography is, by its nature, an integrative science. It seeks to understand and predict biological phenomena in the oceans using information about organisms or their environment. How do temperature-dependent metabolic rates in picoplankton affect global photosynthesis? Do past glaciations impact present-day biogeographic patterns? How do traditional belief systems influence artisanal fishing patterns? These questions span spatial and temporal scales, and even fields of study, but all share similar information challenges. The researcher must collect or find data on the biological phenomenon of interest (photosynthesis, biogeographic pattern, fishing mortality), then collect or find data on the potential explanatory factor (metabolism, plate tectonics, social beliefs), and ensure that the data are "matched" and appropriate for use with respect to spatial and temporal scale, resolution, and quality. Then s/he must bring the data together for some kind of analysis, whether that is a statistical test, a visualization, or another approach, and interpret the results.

This process is data integration, which we define more specifically as the process of combining data residing at different sources and providing the user with a unified view of these data. The unification process might involve creating a new data object—such as a table, map, or graph—that contains information from the different sources. Alternately, it can use computation to aggregate or summarize disparate data elements into a new data object (such as calculating average temperature from a set of temperature sensors).

As the scale of oceanographic research projects increases, and the volume and variety of available data grow, integration becomes increasingly complex. Data centers and data catalogs such as the Ocean Biogeographic Information System (OBIS; www.iobis.org; Zhang and Grassle, 2003), the National Oceanographic Data Center,[1] and the Global Change Master Directory[2] are creating access to ever-growing quantities and varieties of data, as is the internet itself. Very large-scale projects, such as the Global Ocean Ecosystem Dynamics project (GLOBEC),[3] the Ocean Observing Initiative,[4] and the Census of Marine Life[5] address national- or global-scale questions through the work of dozens to hundreds

* Corresponding author. Tel.: +1 858 534 5009; fax: +1 858 534 5117.
E-mail addresses: kstocks@sdsc.edu (K.I. Stocks), condit@sdsc.edu (C. Condit), xqian@sdsc.edu (X. Qian), pebrewin@gmail.com (P.E. Brewin), gupta@sdsc.edu (A. Gupta).

0967-0645/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dsr2.2009.05.022

[1] http://www.nodc.noaa.gov/.
[2] http://gcmd.nasa.gov/.
[3] http://www.globec.org/.
[4] http://www.oceanleadership.org/ocean_observing.
[5] http://www.coml.org/.

of researchers. Technological advances in sensors (e.g., satellite data and streaming sensors) and computers have allowed the capture and storage of much larger quantities of data. At the same time, advances in processing speed have facilitated computation-intensive research, such as analyses and models that address very large scales, fine resolutions, or many components.

1.2. Data integration challenges

As the number of data sources increases, the heterogeneity of data sources becomes more problematic for integration. Two data providers, both supplying ocean temperature information, can use different data platforms (like OPeNDAP,[6] an Oracle relational database and a MySQL relational database), different data types (floating point numbers vs. image intensity values that need to be transformed), different schemas (one system gives maximum and minimum temperatures in two different columns, while another provides the mean and deviation in two different columns, and yet a third uses four different tables to provide the same information per quarter), different computational capabilities (one system only provides data values, while another provides the ability to compute interpolations), and different spatiotemporal and resolution coverage (one system provides fine-grained data around the Fiji islands and another provides 1° resolution data over the entire globe).

Data from many sources are also more likely to be semantically incongruous. If there are two sources having data from the same set of stations, but one numbers the stations and the other assigns letter codes, there is a need to set up an equivalence relationship between them, i.e. a mapping saying that station 1 is the same as station A, etc. Other relationships besides equivalence may also need to be defined, such as if one data source uses species names and the other uses common names—one common name may "map" to multiple species names, and vice-versa.

1.3. Data integration approaches

There are several approaches to handle data integration across multiple heterogeneous data sources. The most practiced approach is an ad-hoc one: a scientist visits all data sources, inspects them, selects relevant data manually, downloads them onto a local machine, and goes through the steps of sub-setting and transforming through scripts written for specific systems. If they want to perform one more data manipulation or analysis, or if they need to use additional analysis software, they would manually transform the data one more time. This approach is not scalable—it becomes increasingly complex, time-consuming, and error prone as the number and complexity of the data increase.

A more automated approach is data warehousing (Voisard and Juergens, 1999). In this approach, one creates an integrated database schema, that is, a new database (more precisely, a new data warehouse) schema that contains all the data elements needed by the end users from all available data sources. Next, the data are physically extracted from these sources and copied to the data warehouse, thus populating the integrated data warehouse schema. The Ocean Biogeographic Information System, for example, takes this approach (Rees and Zhang, 2007). After the data are populated, data queries and analyses can be run against the warehouse. Data warehouses are optimized to accelerate data retrieval. They are particularly attractive for cases where the original data sources may not continue to maintain and make available the data, or where connections to the data sources are slow or unreliable. OBIS, for example, originally started as a fully distributed system, and moved to a central warehouse for performance reasons—distributed queries were too slow, and connections to the data sources unreliable (Rees and Zhang, 2007).

However, data warehousing suffers from a set of limitations that make it less applicable in some situations. Many data sources change through time, as new data are added, so the warehouse must have a synchronization procedure for updating from each data source. Unless the warehouse plans for downtime, the warehouse must maintain one instance of the data for serving while another instance is updated, which can increase storage costs and management effort. It can be expensive and hard to plan for growth, as the size and number of data sources often expands unpredictably over time. Further, the process of warehousing does not intrinsically address semantic incongruency of data. This is handled in warehouses either by transforming all data from different sources to a common data value dictionary, or by having the application logic (i.e., the software between the user interface and the data services) perform the term reconciliation.

Furthermore, when heterogeneous data are being accessed, it can be valuable to keep data in their native formats, regardless of whether those data are distributed or held centrally. This occurs often in oceanography when integrating table/relational data with geospatial formats such as satellite imagery and bathymetric maps. Keeping geospatial formats in a GIS system, for example allows the capabilities of the GIS system to be used, such as determining slopes or finding data within a buffer around a point or region of interest. Similarly, relational databases have a suite of tools for working with related table data. Converting so that both types of data are in the same format would lose important functionality.

These limitations have prompted the use of on-demand information integration engines, also called mediators (Wiederhold, 1992; Gupta et al., 2007). A mediator works by: (1) maintaining a registry of the schemas of the data sources it needs to integrate; (2) allowing the creation of "virtual schemas" over the data sources to be integrated, to bring together fields of interest from across multiple sources, based on a common element like date or location; (3) decomposing complex queries on the virtual schemas into queries sent to the component data sources to retrieve the appropriate data (using wrappers to translate the query appropriately for each data source); and (4) combining the data results from the component data sources and returning them to the user. The user's task is only to pose the query to the virtual schema, and the mediator handles the translation, optimization, and distribution of the query to the relevant sources.

When mediators are used with distributed data sets, they reduce the need for data synchronization because data are not copied; rather, the original source is queried on-demand. Updating is only needed when the capabilities or structure of the data source change, not when records are added, removed, or edited. For distributed or centralized systems, mediators allow data to be held in their native formats, and can use the capabilities of those platforms. Further, they permit joins across the data sources, allowing data in one source to be used to define the query in another data source (as explained further in Section 2.1). When data are held locally in heterogeneous formats, a mediator can supply the needed data registry and wrappers.

All of these approaches have strengths and weaknesses (Jones et al., 2006). We note that the data integration approaches given above—ad-hoc, warehousing, and mediation—are not mutually exclusive, and can be combined in one system. For example, a single system may warehouse large data sources, but query on-demand smaller sources that change frequently. Or a system may centralize, optimize, and standardize data sources as far as is

[6] Open-source Project for Network Data Access Protocol; http://opendap.org/.

reasonable, but use mediator components for integrating across the remaining formats.

We describe here an extensible mediator-based system designed for biological oceanography information. Specifically, we wanted to develop a system that integrates species distribution data (i.e. records of the geographic locations where a species has been observed, plus additional information such as the abundance found) with environmental information such as temperature, nutrient levels, bathymetry, and modeled ocean currents.

To test the performance of the system in biological oceanography, we created prototype portals for the Ocean Biogeographic Information System and for a seamount ecology research project. By applying the system in real scientific projects, we were able to assess whether it was able to meet the researchers' needs, whether the response times were appropriate, and whether additional functionality is needed or desired. For the OBIS testbed, we assessed whether the system response times were acceptable when dealing with large species distribution data sets (11 million records in OBIS) and large physical raster data sets (global World Ocean Atlas data with 1° horizontal resolution and 33 vertical layers). For the seamount study, we used the data integration engine to draw together data to predict the likelihood of oceanographic retention over various seamounts, and determine whether retention impacts specific aspects of the benthic community structure (Brewin et al., 2009). It represented a test of the system's ability to access a wide range of data types.

2. The Cartel mediator

Based on our prior work in spatial information integration (Gupta et al., 1999; Zaslavsky et al., 2000), we have developed a flexible, extensible mediator-based data integration system. The basic mediator was first developed for neurobiology data in the Biomedical Informatics Research Network (BIRN; http://www.nbirn.net/index.shtm), and has been implemented in the online BIRN Data Portal. It originally operated on relational databases only. To extend the mediator to biological oceanography data, we created wrappers and integration strategies for different data types and data sources needed for marine biogeographic research, particularly a suite of geospatial data types.

The mediator is open-source software that is freely available to projects developing oceanographic data systems. We are not developing our own information portal—the testbed projects are designed to demonstrate and test the capabilities of the system, not to represent fully-finished public portals. The mediator and all of the wrappers are developed in Java, and so are platform-independent. First we describe the generic components of the system, and then we detail the specific enhancements made for oceanographic data.

2.1. Generic mediator components

The core of the mediator is (1) a registry of the schemas of all the data sources it needs to integrate, (2) a capability to have one or more virtual schemas defined over the data sources, and (3) a set of wrappers that operate on each participating data source. The purpose of the virtual schema is to effectively hide the heterogeneity of the individual data sources from the user. A virtual schema is much like a regular relational database schema, except that one or more "tables" in the schema is defined as an integrated view over multiple sources—these are called "virtual tables". More technically, a virtual schema is often defined as a collection of integrated views that do not have an explicit primary key–foreign key relationship and are connected through the application. Each virtual table is created by selecting attributes that come from one or more data sources. For example, assume that there are two data sources *Temp* and *Sal*, each holding the following fields:

> *Temp*: date, latitude, latitude-accuracy, longitude, longitude-accuracy, min-depth, max-depth, and temperature value, plus a unique identifier for each row.
> *Sal*: date, max-latitude, min-latitude, max-longitude, min-longitude, max-depth, min-depth, and salinity value, plus a unique identifier for each record.

Note that both the temperature and the salinity values are declared within spatial cubes given by the boundaries or accuracies on the latitude, longitude, and depth.

Suppose we want to create a virtual data set on the summer temperature and salinity from these two sources. We can create a virtual table, *Physical_Data*, that holds the date, new-lat-min, new-lat-max, new-lon-min, new-lon-max, new-depth-min, new-depth-max, temperature, and salinity. In this virtual table *date* comes from taking the common dates from both tables, and then sub-setting only the summer months from them. The field's *new-lat-max* and *new-lat-min* are defined by creating a spatial intersection of the latitude boxes of each record of each table, and the fields *temperature* and *salinity* are taken from the *Temp* and *Sal* tables, respectively. The virtual schema may have more complexity. If two sources represent temperature in two different units, the virtual table will define the appropriate conversions. In database terms, such a virtual table, created by defining the output tables in terms of the source tables is called an *integrated view*.

The integrated view is currently specified using an XML format that closely resembles a relational schema (a more relational representation of the same information has been implemented recently in the version of Cartel used by the Biomedical Informatics Research Network). In the testbeds, the integrated view was developed by a biological oceanographer working with an application engineer. The biological oceanographer described the queries and results that were desired, and the application programmer implemented the integrated views. A similar approach has been successfully used in developing the BIRN neurobiology data portal. A tool called Fuente (Astakhov et al., 2006) can assist with developing integrated views, though more experienced application engineers often prefer to develop without the tool. In the future, it would be desirable to develop a graphical user interface that allows users of the system to define new integrated queries and schemas on the fly without needing technical skills.

The wrappers are the components that translate a query on a virtual schema into actual queries sent to the data sources. For example, consider the query "find the average salinity over all data within a 200 mile buffer of Australia." If one data source provides country boundaries in a shapefile stored in the Oracle Spatial[7] system, then the part of the query that needs to compute the 200 mile buffer must be use the specific primitives of the Oracle Spatial system. This translation from the user's query to the mediator into a query in Oracle Spatial syntax is performed at the wrapper. If the data source was ArcGIS[8] instead, the translation would involve translating the query into the function calls that the ArcGIS application programming interface (API) provides. The second translation involves data. The data representation used by the mediator may not match a data source's

[7] http://www.oracle.com/technology/products/spatial/index.html.
[8] http://www.esri.com/software/arcgis/.

representation of the same data. The format translation process is performed by the wrapper. It is important to recognize that what scientists do today amounts to the development of conversion tools from every format to every other format. Data integration systems reduce this by translating formats from every source format to the mediator-recognized format.

The mediator operates by decomposing a query on the virtual schema into its constituent parts. Consider the query "find the average salinity in the 200–1000 m depth range for winter months in regions where the temperature at the same depth and time is below 0 °C, and show the regions on a map". The mediator finds regions where the temperature conditions are satisfied by sending this part of the query to *Temp*. Given the results of this partial query, the system keeps a copy of these regions, and also sends these regions to *Sal* to get the average salinity values computed. Then these regions and the average salinity values come back to the mediator. The mediator sends the regions to a map server and the average value back to the user.

Note that this query represents a join across data sources. The ability to create joins between data sets is not provided by other technologies commonly used for scientific data system, such as OPeNDAP, ODBC[9] or DiGIR.[10] Further, finding data within a 200-mile buffer represents an example of using the capabilities of the native data source, a capability that may not be available if all data sources are converted into a single format and platform. While these capabilities can be developed in custom systems (i.e. the ad-hoc approach described in the introduction), the advantage of the mediator is that it is reusable and scalable—once a wrapper is developed for a data type, new data sources can be added with minimal effort, and new systems can be developed with much less effort than coding from scratch.

2.2. Wrappers for oceanographic data

The mediator was originally developed to operate across data in different relational databases. In biological oceanographic applications, however, data of interest are often stored in a variety of non-relational formats. To allow the mediator to operate across these data types, we created the following additional wrappers and integration strategies.

- A software suite to convert any OPeNDAP data source to a relational database with minimal human intervention. OPeNDAP[11] is client and server software for making data accessible over the internet. In our conversion process, the tasks of traversing the directory hierarchy to parse metadata and data files, and the creation and population of the database, are performed automatically. The system needs human input only to give meaningful names to data and metadata tables and to data attributes. Once an OPeNDAP repository is converted, the system treats it as a standard relational resource.
- Software routines to generate Map Algebra[12] requests from the mediator for raster data sources stored in ESRI systems. Our system does not generate all possible Map Algebra requests, but can create translations for a subset of Map Algebra functions.
- Wrappers that accept spatial data in shapefiles and ingest them as Oracle Spatial objects. Wrappers have also been developed to query spatial data stored in Oracle and PostGIS databases.
- Software to get raster imagery from data sources like MODIS,[13] and index it using new indexing schemes for faster data retrieval.
- A library of software routines to perform tasks like various unit conversions, distance computations, geometric tests like point-in-polygon and so on. This library helps us to perform quick computations at the mediator and increases efficiency by avoiding access to remote computational resources.
- An efficient strategy to display a large number of spatial data—we use a map server to convert the spatial data into images, which can be more efficiently sent as a raster layer to a web-based map viewer like Google Maps. The same map server is also used to perform simple spatial computations such as creating polar projections of query results.

3. Testing the mediator—methods

3.1. OBIS testbed

The OBIS testbed was used to evaluate the system's performance on large data sets. We developed an implementation of the mediator to integrate the contents of the Ocean Biogeographic Information System with the World Ocean Atlas data (Levitus, 2006) and vector data sets of boundaries such as Exclusive Economic Zones (EEZ), Marine Protected Areas, and the Major Fishing Areas of the UN Food and Agriculture Organization. OBIS is an international project providing access to georeferenced marine species location records—the core data are records of a particular taxonomic name at a particular latitude, longitude, and depth, plus some optional additional information about that observation or collection. At the time of this test, OBIS was serving 11.5 million data points. The World Ocean Atlas 2005 (WOA) is a global data set of ocean parameters, such as temperature, salinity, and nutrient levels, with 1° latitude/longitude horizontal resolution and 33 vertical layers. Our test queries were to (1) extract biological data for a given taxon falling within a given polygon (e.g. "return all tuna records found within Namibia's Exclusive Economic Zone); (2) provide summaries of physical conditions in which a taxon is found (e.g. "provide the mean and standard deviation for the temperature and nitrogen levels of the locations where *Hoplostethus atlanticus* has been recorded"); and (3) extract biological data points based on physical conditions (e.g. "return a list of all taxa, and their location records, found in surface waters of 17–27 °C"). All data were downloaded to the local system, as the intent was to test the mediator's capabilities, and not the internet connectivity speed to the source data sets. The system used for testing was a Quad Core 3 GHz Intel Xeon with 4 GB RAM running Windows Server 2003.

To demonstrate the mediator's capabilities, we developed a prototype portal. Note that the mediator is not tied to a particular interface—the intent of the Cartel mediator project is to provide a component that can be built into other projects' portals and data systems—this interface was developed just to demonstrate the query functionality. Fig. 1 shows a query for records of the genus *Strombus* in the region where the ocean surface temperature is between 17 and 28 °C. The results are shown on a Google Map interface, where the different colored squares represent the locations where the organism was reported (from OBIS), and the 17–28 °C temperature band around the equatorial region shows the temperature distribution in 1° bins (from WOA). The taxonomic tree structure shown on the lower left-hand side is stored as a specially indexed hierarchical data structure that lends

[9] Open database connectivity.
[10] Distributed Generic Information Retrieval; http://digir.sourceforge.net/.
[11] http://opendap.org/.
[12] http://www.esri.com/software/arcgis/extensions/spatialanalyst/about/mapalgebra.html.
[13] http://modis.gsfc.nasa.gov/.

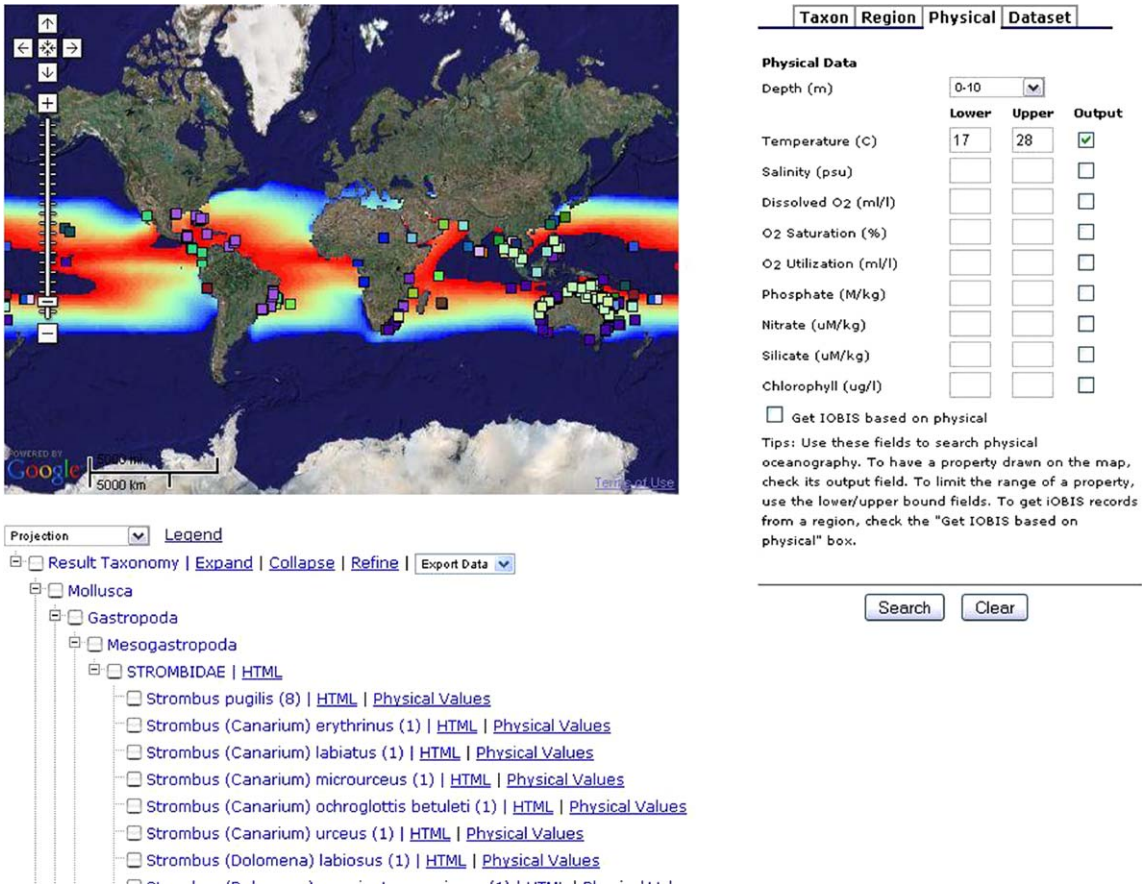

Fig. 1. Prototype portal for the integration of OBIS and WOA data, showing the response from a query asking for all records of species within the genus *Strombus* in the 0–10 m depth zone of waters with an annual mean temperature of 18–15 °C (the "Taxon" tab where the genus *Strombus* was specified is not visible).

itself to fast query response. It is used both to display data returns, and also as a control for refining queries—it can be used to select a subset of the taxonomic tree to refine the original query, or to view the summary environmental parameters of a taxon. As with most of the portal's features, the particular functionality, such as which taxonomic levels to show in the hierarchy, can be tailored by the portal developer. After the spatial component of the result is computed, it is sent to the map server to create a result image that is sent to the Google Maps interface. Fig. 2 shows the result of second query where the user asked for all occurrences of *Thunnus alalunga* inside the Namibian Exclusive Economic Zone. In this case, the system performed a spatial query: point data records come from OBIS and are filtered against the polygon in the PostGIS system. Searching can also be done with user-drawn polygons.

The user interface that exposes the application functionality was divided into two parts—the first part is a set of tables and API that can be retrieved only from a single source. These were part of the virtual schema, but were accessed from the application directly without going through the mediator. The second part needed union, intersection or join queries across multiple sources. These made use of one or more virtual tables directly or through a function call on top of virtual tables. This approach was found to be quite efficient in practice because it tends to reduce the number of cross-source join operations, and acts as an optimization scheme to facilitate faster query processing.

3.2. Seamount ecology testbed

To assess whether the mediator was able to query data from the variety of sources needed in real scientific applications, we developed an instance of the mediator to support the research of a seamount ecology postdoctoral researcher. Seamounts are common features on the ocean's floor, with an estimated 100,000 seamounts over 1 km in height (Wessel, 2001). Some seamounts have been reported to have high levels of endemics (defined here as species found only on one seamount or seamount chain) (Parin et al., 1997; Richer De Forges et al., 2000; Wilson and Kaufmann, 1987). How prevalent this endemism is across seamounts, and to what degree the apparent levels of endemism are artifacts of undersampling in the oceans or taxonomic inconsistency, is currently actively debated (Samadi et al., 2006; reviews in McClain, 2007; Stocks and Hart, 2007). It has been hypothesized, however, that at least some seamounts may be biologically isolated communities and, like terrestrial islands, have increased local speciation (MacArthur and Wilson, 1967; Richer De Forges et al., 2000; Whittaker, 1998).

Oceanographic retention (e.g., Taylor cones) is one mechanism that has been proposed as an isolating mechanism for seamounts—recirculating water traps larvae over the seamount and acts as a barrier to dispersal on or off the seamount (Mullineaux and Mills, 1997; Parker and Tunnicliffe, 1994). Here, the effect of retention on seamount communities was tested by using a model to predict retention potential over seamounts that have been sampled biologically. The biological response variable used was taxonomic distinctness—this measures the average phylogenetic path length between two species on a seamount (Warwick and Clarke, 2001; Webb et al., 2002). It was predicted that seamounts with more retention would have lower average taxonomic distinctness (due to reduced species breadth with isolation) and higher variation in taxonomic distinctness (due to clustering of species within higher ranks

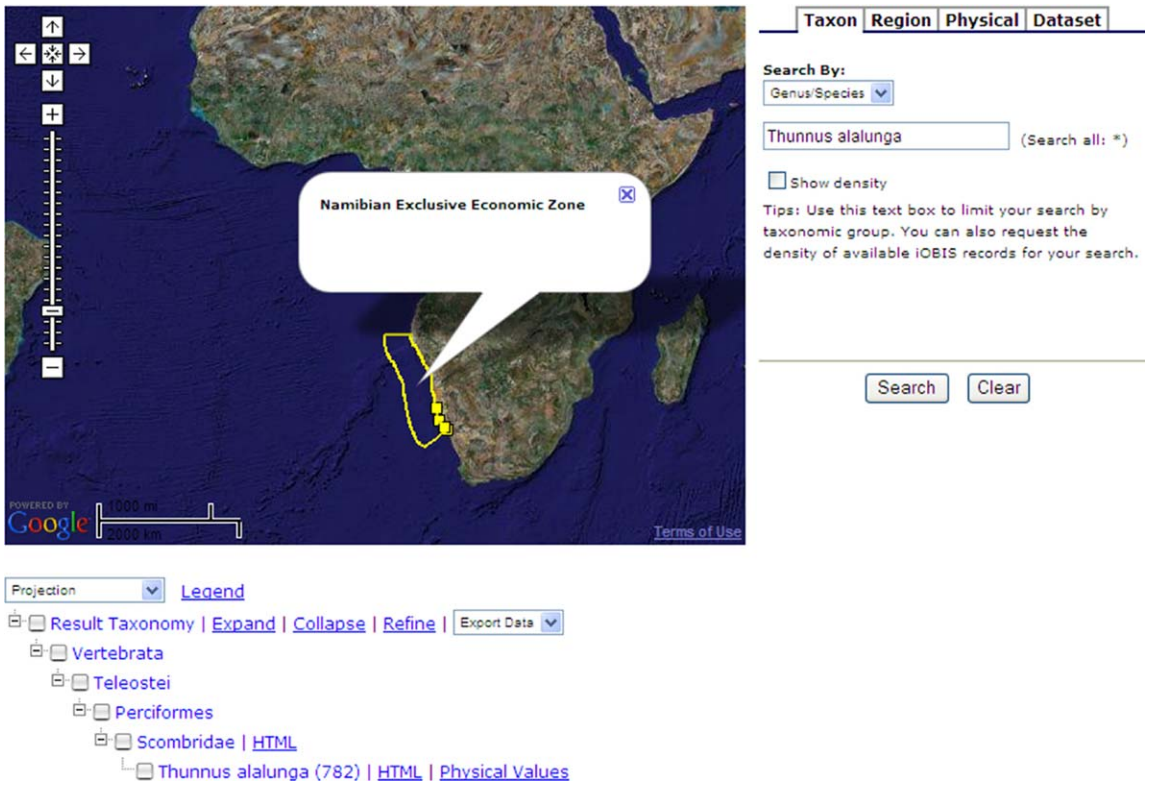

Fig. 2. Prototype portal for the integration of OBIS and WOA data, showing the data returned from a search on the species *Thunnus alalunga* within the Namibian Exclusive Economic Zone.

with increased isolation). Gastropods and bivalves were the two groups assessed.

The data sources on which the mediator operated for this testbed were the SeamountsOnline PostgreSQL database of species records from seamounts (Stocks, 2005) and a suite of resources needed for the retention prediction model: the World Ocean Atlas 2006 (Levitus, 2006) physically residing in a local Oracle database; ETOPO2v2 bathymetry (US Department of Commerce, 2006) held as a GeoTiff; output from the SODA-POP ocean model (Carton et al., 2005) taken from an OPeNDAP server and held as a GeoTiff; and a delimited text file of predicted seamount locations and heights (Wessel, 2001) converted to PostGIS (Fig. 3). For this research, the needed data sources were registered to the mediator, a virtual schema was defined that included the needed variables, and queries were run to extract integrated data maps and tables. The data tables were then used to (1) create a retention potential estimate for each seamount; (2) calculate the average and variance taxonomic distinctness for each taxon on each seamount, and (3) use an analysis of variance (ANOVA) to test for a relationship between predicted retention potential for a seamount and taxonomic distinctness (both average and variability) of each taxon. The taxonomic distinctness calculations and the ANOVA were programmed in *R*.[14]

4. Results and discussion

4.1. OBIS testbed results

The purpose of the OBIS testbed was to evaluate performance time—i.e. is the system able to return data from key queries in a reasonable time? Data searches that required data from just one data source performed well. For example, returning all records for a taxonomic group of interest took from 0.1 s (for a taxon with few records) to 5.8 s (for a taxon with >50,000 records). Performance with respect to the three more complex target query types varied. Selecting OBIS data from within a polygon, such as a country's Exclusive Economic Zone performed well, averaging 3.9 s. The other two searches did not perform as well. Returning average temperature and salinity for a taxonomic group of interest varied with respect to the number of physical variables being averaged and the number of records involved. It averaged 125 s for returning means of two environmental variables. Searching OBIS records based on physical conditions, such as returning all records from waters colder than 1 °C, often did not complete. They could only be made to work when the interval of the physical parameter was artificially small: returning all records where the ocean temperature is between 8.99 and 9.02 °C averaged 50 s. The limitation on response times for these large queries was network speed—they require large amounts of data to be passed from the source data sets.

In response, the system is currently being improved to include the ability to cache and index data, which should improve response times significantly. Initial tests using bitmap indexing on raster data sources indicates that response times are improved from 10 to 160 s with test queries on un-indexed data to <10 s in all test queries on indexed data. However, the response time results highlights that the mediator is not an appropriate tool for all systems—when it is practical to warehouse data centrally, and optimize the data for integration and retrieval, response times will be much faster.

4.2. Seamount testbed results

The ecological results of the test of retention effects on taxonomic distinctness on seamounts are discussed in a separate

[14] http://www.r-project.org/.

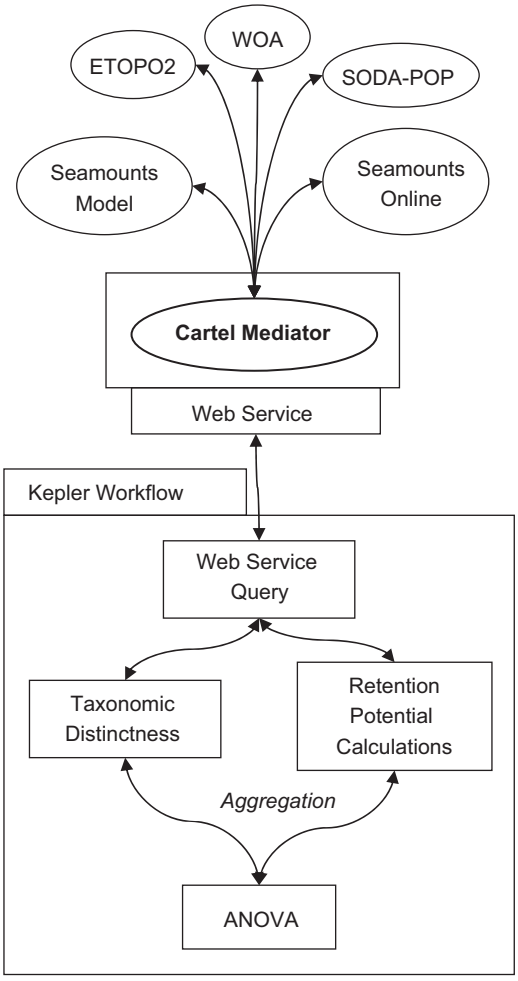

Fig. 3. Diagram of the seamount ecology workflow, showing the data sources used (acronyms are explained in the text) and analytical steps.

publication (Brewin et al., 2009). Here, we evaluate the performance of the data integration engine in supporting this research.

The system was successful in supporting the basic data searching and extraction needs of the research. It could interact with the relational database behind SeamountsOnline to provide records for the taxonomic groups of interest, and return the full taxonomic hierarchy. It could also efficiently query the physical data sets in geospatial data formats. Response times averaged less than 4 s.

However, the testbed process highlighted additional features that the researchers wanted to have in this system. First, they wanted to be able to automate repeated steps, such as extracting the latitude, local current velocity, and seamount height for a set of 50 seamounts. They did not want to initiate 50 separate searches for this. Second, they wanted to be able to move data from queries directly into analysis packages and have the system reformat the data as needed to meet the requirements of the analysis program. For example, they wanted to retrieve all gastropod records for a set of seamounts, then feed those data directly into *R* routines to calculate taxonomic distinctness, in the format the routine requires, then take the output of this, plus the output of the retention potential calculation, and feed it into their statistical package to run an ANOVA, again without having to reformat the data in any way (Fig. 3). In the information technology world, this is called a scientific workflow. These requests are beyond the data integration goals of the mediator itself, but to meet them we expanded the mediator so that it could be an "actor" in a Kepler workflow. Kepler is an open-source scientific workflow application.[15] A researcher using Kepler can now include calls to the mediator within their pipeline of data extraction and analysis.

The third additional feature that the biological oceanographers requested was a better ability to browse and "prune" the data. In real research, it is rare that a scientist approaches a complex set of data and knows exactly what piece they want to extract and analyze. It is more common to want first to explore the data in a less structured way, asking questions like what taxa have been studied widely enough to be representative? If I throw away any seamounts with fewer than 20 records, how many are left? Can I exclude a few seamounts by hand that I know are "different"—too close to land to count as a seamount? Adding this kind of functionality does not require advances in the capability of the mediator itself, but does need thoughtful development of browsing capabilities in the interface to allow more flexible searching, and easier roll-back and repeating of steps.

4.3. Additional considerations

Implementing Cartel within a separate project, the CAMERA metagenomics project,[16] has raised an additional functionality request that we expect to be relevant to biological oceanography. Researchers requested the ability to store and query derived data products, like diversity statistics and taxon accumulation curves, and to have these products searchable in the same system as the other data sources. In some cases, researchers holding data are not willing to share the full data, but are willing to provide certain derived or summary statistics or graphs, so these must be captured by the system directly. This represents a new data type for the system, and we are currently working on adding this capacity to the data integration engine.

5. Conclusions

Cartel is an open-source, extensible data mediation engine that has been tailored to biological oceanography applications, especially to biodiversity studies. It allows a user to integrate data from multiple data sources, without needing to know the underlying data schema and computational capabilities of each data source. It has the ability to query data sources in their native format using the capabilities of the native format, and to allow joins between data sources in different formats (e.g. to use the data in one source to refine a query on another source). It is best suited to situations where central data warehousing is not desirable, such as when the base data sources are updated frequently, or when the data sources have different native formats (such as relational vs. geospatial data), and when joins between data sources are required. Warehousing and mediation approaches can also be combined, with centralization and standardization applied as far as practical and mediation used for sources that must be distributed or held in heterogeneous formats.

The system, once modified to allow it to operate within a scientific workflow, performed well in a testbed study of seamount ecology. It had unacceptably slow response time when applied to large data sets (OBIS and World Ocean Atlas), which is currently being addressed through new indexing and caching approaches. The authors invite contacts from groups interested accessing the mediator code.

[15] http://kepler-project.org/.
[16] http://camera.calit2.net/.

Acknowledgements

We thank the Gordon and Betty Moore Foundation for providing the support for this research. The seamount ecology testbed also benefited from funding from NSF Grant 0340839. We thank Fred Grassle and the OBIS portal team for their assistance in accessing the OBIS database.

References

Astakhov, V., Gupta, A., Grethe, J.S., Ross, E., Little, D., Yilmaz, A., Qian, X., Santini, S., Martone, M., Ellisman, M., 2006. Semantically based data integration environment for biomedical research. In: 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06), pp. 171–176.

Brewin, P.E., Stocks, K.I., Haidvogel, D.B., Condit, C., Gupta, A., 2009. Effects of oceanographic retention on decapod and gastropod community diversity on seamounts. Mar. Ecol. Prog. Ser. 383, 225–237.

Carton, J.A., Giese, B.S., Grodsky, S.A., 2005. Sea level rise and the warming of the oceans in the SODA ocean reanalysis. J. Geophys. Res. 110.

Gupta, A., Ludäscher, B., Martone, M.E., Rajasekar, A., Ross, E., Qian, X., Santini, S., He, H., Zaslavsky, I., 2007. BIRN-M: a semantic mediator for solving real-world neuroscience problems. In: Proceedings of the 2003 ACM Sigmod International Conference on Management of Data (Demonstration). ACM, New York, USA, p. 678.

Gupta, A., Marciano, R., Zaslavsky, I., Baru, C., 1999. Integrating GIS and imagery through XML-based information mediation. In: Agouris, P., Stefanidis, A. (Eds.), Integrated Spatial Databases: Digital Images and GIS, Lecture Notes in Computer Science, vol. 1737. Springer, Berlin, pp. 211–234.

Jones, M.B., Schildhauer, M.P., Reichman, O.J., Bowers, S., 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Ann. Rev. Ecol. 37, 9–544.

Levitus, S. (Ed.), 2006. World Ocean Atlas 2005. NOAA Atlas NESDIS 61-64. US Government Printing Office, Washington, DC.

MacArthur, R.H., Wilson, E.O., 1967. The Theory of Island Biogeography. Princeton University Press, Princeton, USA.

McClain, C.R., 2007. Seamounts: identity crisis or split personality? J. Biogeogr. 34, 2001–2008.

Mullineaux, L.S., Mills, S.W., 1997. A test of the larval retention hypothesis in seamount-generated flows. Deep-Sea Res. I 44, 745–770.

Parker, T., Tunnicliffe, V., 1994. Dispersal strategies of the biota on an oceanic seamount: implications for ecology and biogeography. Biol. Bull. 187, 336–345.

Parin, N.V., Mironov, A.N., Nesis, K.N., 1997. Biology of the Nazca and Sala y gomez submarine ridges, an outpost of the indo-west pacific fauna in the eastern Pacific ocean: composition and distribution of the fauna, its communities and history. Adv. Mar. Biol. 32, 145–242.

Richer De Forges, B., Koslow, J.A., Poore, G.C.B., 2000. Diversity and endemism of the benthic seamount macrofauna in the southwest Pacific. Nature 405, 944–947.

Rees, T., Zhang, Y., 2007. Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System. In: Proceedings of 'Ocean Biodiversity Informatics': An International Conference on Marine Biodiversity Data Management Hamburg, Germany, 29 November–1 December, 2004, UNESCO/IOC, VLIZ, BSH, Paris, pp. 167–176.

Samadi, S., Bottan, L., Macpherson, E., De Forges, B.R., Boisselier, M.C., 2006. Seamount endemism questioned by the geographic distribution and population genetic structure of marine invertebrates. Mar. Biol. 149, 1463–1475.

Stocks, K., 2005. SeamountsOnline: An Online Information System for Seamount Biology. Version 2006-1, World Wide Web electronic publication, 〈http://seamounts.sdsc.edu/〉.

Stocks, K.I., Hart, P.J.B., 2007. Biogeography and biodiversity of seamounts. In: Pitcher, T.J., Morato, T., Hart, P.J.B., Clark, M.R., Haggan, N., Santos, R.S. (Eds.), Seamounts: Ecology, Conservation and Management. Fish and Aquatic Resources Series 12. Blackwell, Oxford, UK, pp. 255–281.

US Department of Commerce, National Oceanic and Atmospheric Administration, National Geophysical Data Center, 2006. 2-minute Gridded Global Relief Data (ETOPO2v2) 〈http://www.ngdc.noaa.gov/mgg/fliers/06mgg01.html〉.

Voisard, A., Juergens, M., 1999. Geographic information extraction: querying or quarrying? In: Goodchild, M., Egenhofer, M., Fegeas, R., Kottman, C. (Eds.), Interoperating Geographic Information Systems. Kluwer Academic Publishers, New York, USA, pp. 165–180.

Warwick, R.M., Clarke, K.R., 2001. Practical measures of marine biodiversity based on relatedness of species. Oceanogr. Mar. Biol. Ann. Rev. 39, 207–231.

Webb, C.O., Ackerly, D.D., Mcpeek, M.A., Donohue, M.J., 2002. Phylogenies and community ecology. Annu. Rev. Ecol. Syst. 33, 475–505.

Wessel, P., 2001. Global distribution of seamounts inferred from gridded Geosat/ERS-1 altimetry. J. Geophys. Res. 106 (B9), 19,431–19,441.

Whittaker, R.J., 1998. Island Biogeography: Ecology, Evolution, and Conservation. Oxford University Press, Oxford, UK.

Wilson, R.R., Kaufmann, R.S., 1987. Seamount biota and biogeography. Am. Geophys. Union Geophys. Mon. 43, 355–378.

Wiederhold, G., 1992. Mediators in the architecture of future information systems. IEEE Computer 25 (3), 38–49.

Zaslavsky, I., Ludäscher, B., Gupta A., Marciano, R., 2000. Accuracy mediation in a spatial wrapper mediator system. In: First Geographic Information Science Conference, Savannah, Georgia.

Zhang, Y., Grassle, J.F., 2003. A portal for the ocean biogeographic information system. Oceanol. Acta 25, 193–197.