Rapid Development of a Hybrid Web Application for Synthesis Science of *Symbiodinium* with Google Apps

Erik C. Franklin, Michael Stat, Xavier Pochon, Hollie M. Putnam, Ruth D. Gates
Hawaii Institute of Marine Biology, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa
erik.franklin@hawaii.edu, stat@hawaii.edu, pochon@hawaii.edu, hputnam@hawaii.edu, rgates@hawaii.edu

Abstract— **Data interoperability facilitates the integration, access, and delivery of information from a variety of sources to synthesize knowledge for scientific collaboration. Often the success of a workgroup-scale data integration project can be hindered by the insufficient computing expertise of the team, inadequate network resources, and limited funding to support cyberinfrastructure. We explore the utility of the free, cloud-based Google Apps to overcome these potential shortfalls and present a case study for the development of a hybrid web application, called GeoSymbio, that synthesizes global bioinformatic and ecoinformatic data of *Symbiodinium*, a group of uni-cellular, photosynthetic dinoflagellates that are found free-living or in symbiosis with a wide range of marine invertebrate hosts including scleractinian coral. Google Apps allowed our five member multidisciplinary group of biologists to develop a web-based tool to discover, explore, and visualize project data in a rapid, cost-effective, and engaging manner. Although the final product exceeded our expectations, there were certain limitations that we encountered including file data storage limits, the slow loading speed of some tools, and incomplete integration among applications. Traditionally, scientific data synthesis and integration has been presented as static journal review articles. Here, we demonstrate a path to develop a novel type of web-based, data-driven, and publically accessible review of scientific knowledge that allows the user to dynamically interact with the compiled information using Google Apps. GeoSymbio is located at https://sites.google.com/site/geosymbio/.**

Keywords—bioinformatics; data interoperability; ecoinformatics; GeoSymbio; Google Apps; Symbiodinium

I. Introduction

In March 2011, our group of five biologists was tasked with the compilation of global bioinformatic and ecoinformatic data on coral host-symbiont symbioses for analysis, synthesis and visualization as part of the "Tropical Coral Reefs of the Future" working group at the National Center for Ecological Analysis and Synthesis (NCEAS). Over the prior two years, we had already considered this issue and thus had created a data schema and populated a database with approximately 2500 records manually data-mined from GenBank and journal articles. Yet after the extensive early work on the project, the information only existed as a spreadsheet file circulated between our desktop computers. During the working group, we were challenged with the issue of how our small scientific research team could, in a rapid time frame for a low cost, integrate various data streams, expand the database, and then broadly share the synthesis results without having a computing support team (such as a network administrator, database administrator, or web programmer). Our proposed solution involved the adoption of the Google Apps software as the computing framework for data entry, management, and visualization of project information. By May 2011, we had a functional solution of web-based tools that exceeded our project requirements. In this article, we describe the development process relative to the compilation, integration, access, and delivery of information for the scientific synthesis of global symbiont data (of the genus *Symbiodinium*) with Google Apps.

II. Background

A. Symbiodinium Biology and Taxonomy

The genus *Symbiodinium* is a group of uni-cellular, photosynthetic dinoflagellates found either free-living or in symbiosis with a wide range of marine invertebrates including scleractinian corals. *Symbiodinium* encompasses nine divergent genetic lineages called clades [1] which each contain multiple subclade sequence types. The Internal Transcribed Spacer 2 region (ITS2) of the nuclear ribosomal array has been used extensively for genetic identification and taxonomic description of over 400 distinct *Symbiodinium* subclade types in invertebrate hosts sampled from a variety of marine habitats of tropical and subtropical waters [2, 3, 4, 5].

B. Global Symbiodinium Database Schema

Prior to the NCEAS working group, we had previously designed a database plan to reflect the bioinformatic and ecoinformatic information relevant to global *Symbiodinium*-host symbioses (Table I). The plan had 33 variables that described information based on *Symbiodinium* occurrences such as sequence identification, method of identification, host taxa, collection event, sampling location and citation reference details. The variables and their definitions were adapted from the Ocean Biogeographic Information System (OBIS) Schema v1.1 [6] which is an extension of the Darwin Core Version 2 standard. The detailed definitions of each of the data fields are available online at the GeoSymbio schema webpage [7].

TABLE I. DATA FIELDS OF GLOBAL *SYMBIODINIUM* DATASET.

Group	*Field*	*Data Type*
Symbiodinium	Clade	Text
	Subclade	Text
	Gene	Text
	Isolate	Text
	Redundancy of Sequence	Text
	Species	Text
	Methodology	Text
	Genbank	Text
	Genbank link	Hyperlink
Host Taxa	Host Phylum	Text
	Host Class	Text
	Host Order	Text
	Host Family	Text
	Host Genus	Text
	Host Species	Text
	Host Scientific Name	Text
	Host AphiaID[a]	Text
	Environment	Text
Collection Event	Ocean	Text
	Country	Text
	State Region	Text
	Sub Region	Text
	Locale	Text
	Latitude	Numeric
	Longitude	Numeric
	Coordinate Precision	Numeric
	Minimum Depth	Numeric
	Maximum Depth	Numeric
	Start Year Collect	Date
	End Year Collect	Date
Citation	Reference short	Text
	Reference full	Text
	Reference link	Hyperlink

[a] World Register of Marine Species unique taxonomic identifier (www.marinespecies.org)

C. Data-Mining GenBank and the Scientific Literature

The primary repository for *Symbiodinium* genetic sequence information is the US National Center for Biotechnology Information's GenBank. Sequence records are archived digitally, identified with an accession number, and accessible through a variety of online NCBI search tools. In 2009, we began querying GenBank for all *Symbiodinium* ITS2 sequences in order to populate the database. We quickly discovered that GenBank contained many redundant entries, records that were often incomplete, and that there was little quality control on the submitted ITS2 data. Furthermore, the missing or coarse resolution of geographic description often encountered in GenBank submissions severely limited our ability to automate the geographic mapping of genetic sequence data, an important requirement for our database. From the redundant sequences, we identified identical sequences (i.e., 100% residue similarity) with different accession numbers as synonyms with the first published record as the "parent" accession number. Then, we manually searched the source literature to confirm or ascertain the following descriptive characteristics for each sequence: host taxa, location, collection year, and laboratory methodology. The mapping of *Symbiodinium* occurrence locations often required reading the primary literature source identified in the GenBank accession record, with a cross-check of location in GEOnet Names and Google Earth. Although the process was time consuming, we had approximately 2500 records in our global *Symbiodinium* data table by March 2011.

III. DEVELOPMENT OF GEOSYMBIO

Building the capacity to examine the diversity, ecology and biogeography of *Symbiodinium*-host symbioses has global and societal implications and thus, the compilation and dissemination of this information was essential. One of the major barriers to progress was that the geographic, host taxa, and temporal details of the *Symbiodinium* occurrence records were not exposed and documented well in existing databases. This required manual examination of data records as well as extensive reading of the primary literature to extract useful ecological information to match with the genetic data. Our data-mining activities had already provided a good foundation for the dataset but we lacked a streamlined means to visualize and explore the data for research. To provide better access to this information, we determined that we required a system that provided four basic functions: (1) geospatial visualization, (2) text-based queries, (3) knowledge summaries, and (4) data products for further analyses. Given the time, personnel, and fiscal constraints, we required a simple, cost-effective (as in free), and robust solution for the system. After exploratory research of potential solutions, we began development of GeoSymbio using Google Apps in March 2011 at the NCEAS working group meeting.

A. Project Framework with Google Apps

Google Apps are a suite of cloud-based software that provides a variety of functionality for performing computing tasks. To meet our system functionality requirements, we utilized Google Sites to host the web application, Google Maps and Google Earth for geospatial visualization, Google Spreadsheets for data entry and management, Google Fusion Tables for data management and visualization, and Google Gadgets for data queries, knowledge summaries, and visualization (Fig. 1). Google APIs were also used to script minor components of the system to retrieve data from remote servers and share map data from Fusion Tables using Javascript, for example. Once the initial data was imported to a Spreadsheet, the project activities were primarily cloud-based.

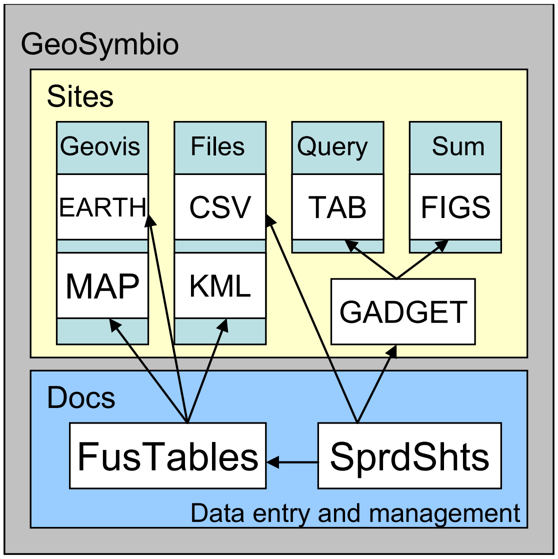


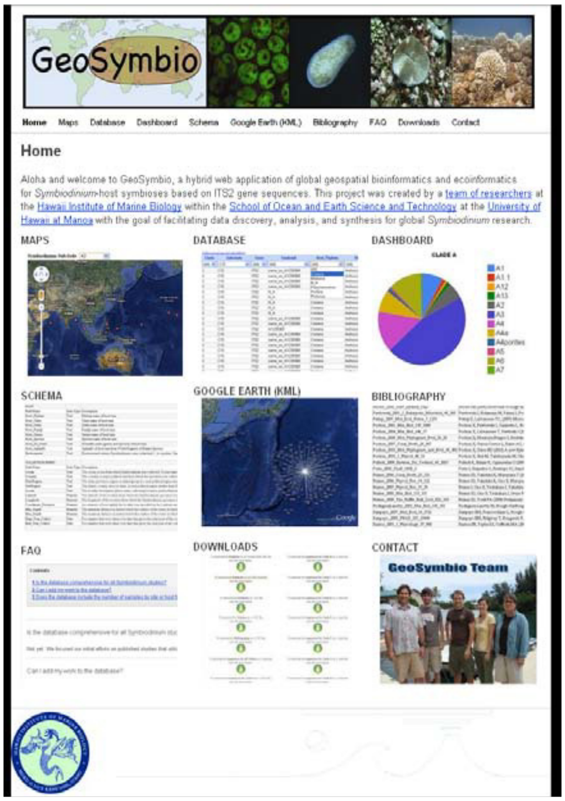
Figure 1. Schematic of the software components of the GeoSymbio web application. Components are white boxes, arrows are directional flow of data, and colored boxes are component groups based on function (green) or accessibility (yellow for public Google Site and blue for private Google Docs). Google Spreadsheets serves as the data entry point and primary management interface for the research team. Abbreviations in the figure are defined as Sites: Google Sites; Geovis: geovisualization, Files: files for download, Query: text-based tabular queries; Sum: knowledge summaries through dashboard figures; Earth: Google Earth; Map: Google Maps; CSV: a comma-separated tabular data file; KML: a keyhole markup language data file; TAB: table for text-based queries; FIGS: pie-chart figures of database element summaries; GADGET: Google Gadget; Docs: Google Docs; FusTables: Google Fusion Tables; and SprdShts: Google Spreadsheets.

B. *Data Entry and Management*

A Google Spreadsheets file provided the primary data entry and management interface for the *Symbiodinium* dataset. Previously, the dataset had been kept as a desktop spreadsheet file that was mailed to collaborators as new changes arose. This inefficient method of data management spawned multiple versions of the data file without a good means of tracking changes amongst the team. Furthermore, many additional *Symbiodinium* studies had been published which needed to be added to the database for the working group. Prior to the upload, we determined the most accurate version of the existing dataset for the project. Once in Spreadsheets, the data table allowed multiple simultaneous edits, versioning, and controlled vocabularies for data entry that greatly accelerated our ability to compile additional records in an efficient and robust manner. Several functions within Google Spreadsheets proved extremely useful for remote access of other data providers such as “ImportXML”. For example, this function allowed genetic sequence retrieval from NCBI through their Entrez Programming Utilities (E-utilities) programs with XPath expressions. In addition, the RESTful structure of the applications allows direct access to the entire or subsets of the data files through the Google Fusion Tables SQL API. Using these methods, we nearly doubled the number of records to over 4800 and included records from all published studies of ITS2 gene. Once completed, the Spreadsheets data table was manually copied to a Google Fusion Table, an action that is not yet automated. The two data files in Spreadsheets and Fusion Tables provided the foundation for the other components of the hybrid web application, GeoSymbio.

C. *Hybrid Web Application*

GeoSymbio is the first comprehensive effort to collate and visualize *Symbiodinium* ecology, diversity, and biogeography in an online web application that is freely accessible and searchable by the public. The application structure is a hybrid or compilation that draws functionality and information from a variety of visualization tools and digital data and reference sources, with the core of the application hosted remotely or “in the cloud” using Google Sites [8]. The interconnected components of the application are made up of Google Spreadsheets, Google Fusion Tables, Google Maps, Google Earth, and Google Gadgets (Fig. 1). Thus, project information is accessible through any web-browser with internet access, so the application is not specific to a computer platform. The application is comprised of a collection of 10 web pages which include database knowledge summaries (DASHBOARD), searchable text-based queries (DATABASE) and spatial-based maps (MAPS and GOOGLE EARTH), the database schema (SCHEMA), a bibliography (BIBLIOGRAPHY), frequently asked questions (FAQ), downloadable map and sequence data files (DOWNLOADS), and project team contact information (CONTACT) (Fig. 2). The following sections of the paper



Home

Aloha and welcome to GeoSymbio, a hybrid web application of global geospatial bioinformatics and ecoinformatics for *Symbiodinium*-host symbioses based on ITS2 gene sequences. This project was created by a team of researchers at the Hawaii Institute of Marine Biology within the School of Ocean and Earth Science and Technology at the University of Hawaii at Manoa with the goal of facilitating data discovery, analysis, and synthesis for global *Symbiodinium* research.

Figure 2. Screenshot of GeoSymbio hybrid web application home page.

detail the site functionality of geospatial visualization, text-based queries, knowledge summaries, and data downloads. The GeoSymbio URL is https://sites.google.com/site/geosymbio/.

D. Geospatial Visualization

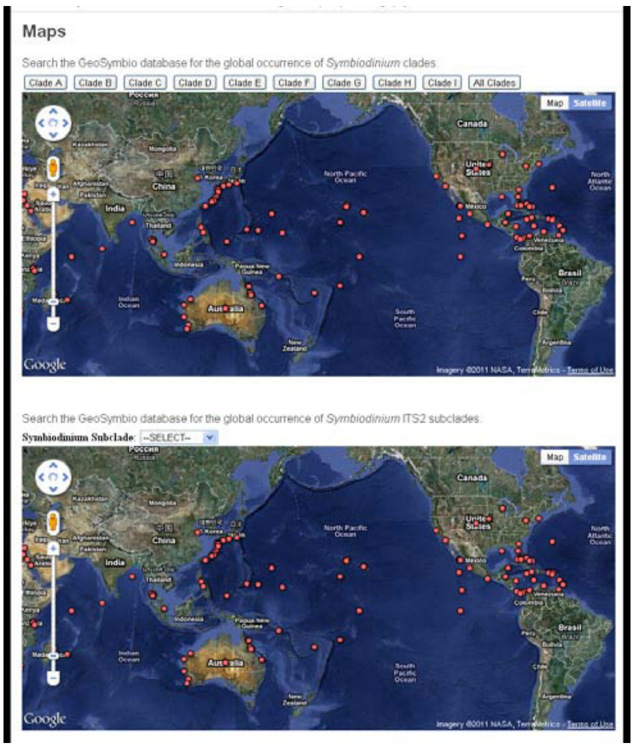
Within GeoSymbio, the maps and Google Earth pages provide geospatial searches and visualization of the dataset for *Symbiodinium* clades and subclade types. The data for both mapping methods are stored in a Google Fusion table. The map components access the data through AJAX using the unique numeric identifier associated with the Fusion Table. Building off of the basic tutorials for Fusion Tables, we customized the map page interface with Javascript to allow a user to select the clade or subclade with buttons or a drop-down menu, respectively. The KML data network link for Google Earth is a standard feature available for Fusion Tables and did not require customization. The network link was used for both the Google Earth embedded viewer in the web page and the creation of the KML download file. The GeoSymbio maps webpage allows searches for *Symbiodinium* clade and subclade type as determined by ITS2 sequence type (Fig. 3). The Google Earth (KML) page provides a dynamic globe embedded in the website with the attributes of the GeoSymbio database accessible for each location in pop-up info windows.

E. Text-Based Data Queries

The GeoSymbio database page provides a dynamic data table with text filtering and grouping functions, which provide extremely flexible means to query for data. This functionality is provided by a Google Table Gadget that draws data from the primary data table in Spreadsheets. Filtering the database allows a simple yet powerful method to examine combinations of single filters for each attribute column. For example, a researcher interested in the occurrence of a subclade type within a particular host could filter dynamically to view records that meet the criteria. The grouping method of the database lends an even greater capacity to summarizing data with hierarchical relationships among the database attributes. To continue the previous example, a hierarchical grouping of host and clade with a count by subclade would dynamically update the table to show the selected criteria with subtotal record counts by group elements.

F. Knowledge Summaries

The knowledge summaries represent a quick view of the information contained in the dataset. The dashboard page presents a set of interactive pie charts that visualize the number and proportion of data records by *Symbiodinium* clade, ITS2 subclade sequence type, taxonomic order of the host, collection year, and location. These pie charts are Google Gadgets that pull data from the summarized subsets of information in the dataset. The visual nature of the charts presents a powerful means to rapidly convey relationships between fields in the dataset. The charts are dynamically updated as data is added to the Spreadsheets dataset. The charts can also be dynamically queried for the count and proportion of records for each category. In addition, the database schema and bibliography are both dynamically linked to the database and displayed as embedded table Google Gadgets on their respective webpages.



Maps

Search the GeoSymbio database for the global occurrence of *Symbiodinium* clades

Search the GeoSymbio database for the global occurrence of *Symbiodinium* ITS2 subclades

Symbiodinium Subclade: --SELECT--

Figure 3. Screenshot of GeoSymbio maps webpage for geographic visualization of *Symbiodinium* clades (upper map) and ITS2 subclade sequence types (lower map) in embedded Google Maps and data from Google Fusion Tables. Search critieria can be subset by user in both maps with buttons and menus.

G. Data Downloads

The data download page includes links to download the map files (as .kml and .shp) to view the data in map programs such as Google Earth or ESRI ArcGIS. In addition, a set of genetic sequence alignment files (.fasta) can also be downloaded. Each of the nine sequence alignment files (including all sequences from one of each of nine existing *Symbiodinium* clades) was subjected to the following three steps. First, the sequence alignment file was imported into the alignment software BioEdit v7.0.9 [8] where it was subjected to automatic alignment using ClustalW [9] and further improved manually. Second, the aligned sequence file was run in 'DNA to haplotype collapser and converter' freely available at the online FaBox [10]. Except for the sequence and shapefile files, all other files were created through Google tools linked to the dataset.

H. Limitations of Google Apps

The overall development process with Google Apps was a strong success since we met our project requirements in a rapid and cost-effective manner. Nonetheless, there were elements about Google Apps that were not ideal including data storage limits, slow page loads, and less than seamless integration between applications. For example, the file data storage limit for a table is currently 400,000 cells. With the 33 variables in our data schema, we will be limited to approximately 12,000 records in the data table. At that point, we would possibly need to reconfigure the structure to multiple tables. Fusion Tables offers more data storage but not the same data editing and management functionality as Spreadsheets. Further, there is no current way to automate the association between a Spreadsheet and a Fusion Table which necessitates a manual update between the two. Also the slow load speed of several of the Google Gadgets and, in particular, the embedded Google Earth viewer requires 30 seconds to several minutes of wait time. These load times seem to improve after the first page loaded but may detract the casual user from using the tools by clicking away from the page during the delay. Furthermore, the high rate of change of infrastructure and functionality of Google Apps represents both an advantage and a disadvantage for this type of web solution. The advantage is that desired features may be implemented much more quickly than in a commercial off-the-shelf package, but the disadvantage is that the way things work can change without notice. Optimal performance was noted with the following browsers: Google Chrome v10, Microsoft Internet Explorer v8, and Apple Safari v4. These concerns in sum suggest that Google Apps may be optimal for smaller datasets and workgroup or smaller project teams. It is unclear if the free tools are scalable for larger projects.

IV. Conclusions

The need for a tool like GeoSymbio arises from the difficulties of integrating multiple data sources and information to perform bioinformatic and ecoinformatic data synthesis particularly in a geospatial context. These tasks can be challenging to execute without an interdisciplinary skill set of highly specialized scientific knowledge and a strong computing background, thus creating a barrier for progress among researchers. We demonstrate the rapid, cost-effective, and successful implementation of a hybrid web application for synthesis science developed with Google Apps. The web application provides four primary functions: (1) geospatial visualization, (2) text-based queries, (3) knowledge summary, and (4) data products. Starting with an existing data schema, the web application was developed and fully functional over a 5-week period from March to May 2011. Some disadvantages of using Google Apps include file data storage limits, the slow loading speed of some tools, and incomplete integration among applications. The rapid pace of development of Google Apps presents the benefit of an expanding suite of functionality but the potential for unwanted change with limited notice. Although we have expressed some caveats regarding the tools, we strongly endorse Google Apps for workgroup-scale projects that seek interoperability between various datasets and a set of web-based tools for dynamic exploration, synthesis, and sharing of knowledge on a scientific topic.

Acknowledgment

The National Marine Sanctuary Program (memorandum of agreement 2005-008/66882), the US Environmental Protection Agency Science To Achieve Results (STAR) PhD Fellowships (FP917096 and FP917199), the US National Science Foundation (NSF) grants through Biological Oceanography (OCE-0752604, OCE-1041673) and the Long Term Ecological Research (LTER) program (NSF 04-17412) provided financial support for this research. This work was conducted as a part of the "Tropical coral reefs of the future: Modeling ecological outcomes from the analyses of current and historical trends" Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California. This is HIMB contribution # 1460 and SOEST contribution #8370.

References

[1] Pochon, X., and Gates, R.D., 2010. A new *Symbiodinium* clade (Dinophyceae) from soritid foraminifera in Hawaii. Molecular Phylogenetics and Evolution 56:492-497.

[2] Lajeunesse, T.C., 2005. “Species” radiations of symbiotic dinoflagellates in the Atlantic and Indo-Pacific since the Miocene-Pliocene transition. Molecular Biology and Evolution 22: 570-581.

[3] Stat, M., Carter, D., and Hoegh-Guldberg, O., 2006. The evolutionary history of *Symbiodinium* and scleractinian hosts—Symbiosis, diversity, and the effect of climate change. Perspectives in Plant Ecology, Evolution and Systematics 8:23-43.

[4] Correa, A.M.S., and Baker, A.C., 2009. Understanding diversity in coral-algal symbiosis: a cluster-based approach to interpreting fine-scale genetic variation in the genus *Symbiodinium*. Coral Reefs 28:81-93.

[5] Silverstein, R.N., Correa, A.M.S., LaJeunesse, T.C., and Baker, A.C., 2011. Novel algal symbiont (*Symbiodinium* spp.) diversity in reef corals of Western Australia. Marine Ecology Progress Series 422:63-75.

[6] Vanden Berghe, E., 2007. The Ocean Biogeographic Information System: web pages. Available on http://www.iobis.org. Consulted on [6 June 2011].

[7] Franklin, E.C., Stat, M., Pochon, X., Putnam, H.M., and Gates, R.D., 2011. GeoSymbio database schema webpage. Accessed 28 July 2011 <https://sites.google.com/site/geosymbio/schema>

[8] Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98

[9] Thompson, J.D., Higgins, D.G., and Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680.

[10] Villesen, P., 2007. FaBox: an online toolbox for FASTA sequences. Molecular Ecology Resources 7:965-968.