Lifemapper: Infrastructure and Services for Biodiversity Science

Aimee Stewart[1], James Beach[1], C. J. Grady[1], Jeffrey Cavner[1]
[1] University of Kansas, Biodiversity Institute
astewart@ku.edu, beach@ku.edu, cjgrady@ku.edu, jcavner@ku.edu

***Abstract*—Lifemapper is an archive of species and environmental data, predicted habitat maps and a suite of data and analysis web services based on these data and the computational processes used to create them. Behind the scenes, Lifemapper relies on open source software libraries, modular code design, and a collaborative development process. As a community resource, Lifemapper is committed to standard data formats and Internet access protocols and is increasingly focused on data transparency and repeatability through cataloging and documenting metadata and provenance.**

Keywords—biodiversity; geospatial; species distribution modeling; macroecology; metadata; standards; infrastructure; web services

I. Introduction

Lifemapper (www.lifemapper.org) is a computational infrastructure project funded by the National Science Foundation (NSF) that combines open source geospatial and biodiversity informatics tools to: enable biogeographical analyses of current and future distributions of species, demonstrate the biological impacts of climate change to junior and senior high school students, and increase the research utilization of the data associated with biological specimens housed in museums around the world. Lifemapper (LM) is organized around two primary components: 1) an archive of predicted current and future species distribution maps and, 2) a set of software tools and services that enable biological researchers to predict and analyze single- and multi-species, multi-scale patterns of species distribution. Lifemapper's software architecture includes a data pipeline that moves researcher requested modeling experiments to a 64-node cluster for computation, and then retrieves the results. Lifemapper then catalogs resulting model outputs, datasets, statistics and metadata for retrieval through standardized web services defined by Open Geospatial Consortium (OGC, http://www.opengeospatial.org/) standards and simple Representational State Transfer (REST) [1] architectural style.

II. Archive

The first Lifemapper component is an extensive archive of predicted species habitat maps. LM's species distribution modeling (SDM) data pipeline automatically assembles experiments with available species occurrence data and with current and future scenario climate data. The input species occurrence data used by LM are aggregated from biological museums, collections and observation databases by the Global Biodiversity Information Facility (GBIF, http://data.gbif.org/). LM calculates SDM experiments from GBIF specimen data and climate data using openModeller (http://openmodeller.sourceforge.net) [2], an open source species modeling framework, which supports a number of ecological niche modeling algorithms as plug-ins, including the most widely-used methods: GARP with Best Subsets [3], Bioclimatic Envelopes [4,5] and Maxent [6]. Climate data includes bioclimatic variables from Worldclim (http://www.worldclim.org) and Global Climate Model (GCM) outputs distributed by the UK Met Office Hadley Centre (http://www.metoffice.gov.uk/climate-change/resources/hadley/) and the National Institute for Environmental Studies, Japan based on International Panel on Climate Change (IPCC) defined scenarios for the Third Assessment Report (TAR, http://www.ipcc-data.org/gcm/monthly/SRES_TAR/index.html) and Fourth Assessment Report (AR4, http://www.ipcc-data.org/gcm/monthly/SRES_AR4/index.html). LM maintains an archive of automatically generated niche model maps, as well as the input species occurrence and climate data used in their creation, for public exploration and retrieval through the Lifemapper web site and web services.

The Lifemapper SDM Pipeline connects the data archive and the computational processes to monitor the system for user-requested experiments and updated specimen data from GBIF, which trigger initial or re- calculation of affected experiments. Worker threads simultaneously update experiment status and inputs, submit experiments to and retrieve results from a 64-node compute cluster. Once results have been written to storage, and metadata cataloged in the system, they are immediately available through LM web services.

III. Web Services

Lifemapper provides the second component, a set of geospatial data and analysis capabilities for use with the LM archive or user data, as web services. All Lifemapper web services are available as web applications at http://www.lifemapper.org, but also can be accessed programmatically using simple Uniform Resource Locator

(URL) construction to identify the web service and appropriate parameters. LM data web services serve specimen occurrences, environmental datasets and predicted habitat maps, as well as metadata for all these data layers.

A. *Species Distribution Modeling*

Analysis tools include Species Distribution Modeling (LmSDM) services available through a REST and OGC Web Processing Service (WPS) interfaces. LmSDM services can be requested using either user-supplied or LM-provided data, and offer model calculations using openModeller and the algorithms implemented within that framework.

As part of the Kansas-Oklahoma NSF EPSCoR project "A Cybercommons for the Great Plains" effort, Lifemapper developed plug-ins for VisTrails scientific workflow software (http://www.vistrails.org), developed by the Scientific Computing and Imaging Institute at the University of Utah, to simplify LmSDM access. This plug-in integration between LM web services and the VisTrails workflow environment enables climate change scientists to assemble complex computational pipelines consisting of sequential tasks connected through an intuitive drag-and-drop programming user interface on the desktop. The LM-VisTrails plug-in enables users to design species distribution modeling experiments using LM data and LmSDM web services to run a species distribution modeling experiment. As additional web services move to production, the LM-VisTrails plugins will include those services as well.

B. *Range and Diversity*

In collaboration with the University of Connecticut (R. Colwell, T. Rangel) in the NSF project "Extending Lifemapper to Enable Macroecological Research", Lifemapper: Range and Diversity (LmRAD) explores the biogeography of species and biodiversity of regions. LmRAD focuses on two fundamental units of biogeography: species range and species diversity. It creates species Presence-Absence Matrices (PAMs), an approach for linking patterns of range size and of species richness at biogeographical scales [7]. The PAM is a gridded data format, where the x-axis represents species and the y-axis represents geographic sites. Each matrix element is coded for the presence (1) or absence (0) of each of hundreds or thousands of species at a given site, by intersecting species range data layers with a grid representing the area of interest. PAMs are the starting points for multiple methods used to test ecological and evolutionary hypotheses about the spatial patterns of biological diversity on continental and global scales.

Arita et al. [8] have shown there are correlations between: a) the species diversity of site (marginal total of diversity) and the mean range size of all species within that site, and b) between the range size of a species (marginal total of occupancy) and the mean species diversity within the range of that species. The correlations are mirror images of the same pattern, reflecting fundamental mathematical and biological relationships represented by the PAM. Range-diversity scatter plots depict these relationships graphically by-species and by-site. After computing indices, the grid is randomized and the process repeated to assess the significance of results.

Lifemapper is concurrently developing plug-ins to Quantum-GIS (QGIS, http://qgis.osgeo.org), a versatile open source Geographic Information System (GIS) desktop application, to simplify access to the LmRAD modules and visualize experiment inputs and results in a full-featured GIS application. By using the multi-platform QGIS as a client to the LmRAD services, Lifemapper brings a powerful set of macroecological analysis tools to a wide variety of users, regardless of the computational power or operating system of their desktop computer. All outputs are provided in standard formats, to simplify further analysis in other software applications.

IV. GUIDING PRINCIPLES

A. *Research and Education*

Students and educators are the main focus of Lifemapper archive creation. The LM archive presents overall picture of predicted distributions for species with adequate digital data. In the NSF Education-funded collaborative project "Change Thinking for Global Science" with the University of Michigan, we are building progressive learning sets with curricula using targeted species in the LM archive to teach middle school students complex concepts of science and ecology. In these learning sets, we have created online worksheets that present material about weather, climate, species, and allow exploration of species distribution maps predicted for current day, and three time steps in the future. Online worksheets guide students through the material to build upon knowledge gained in previous exercises.

Undergraduate students, graduate students, and researchers are the intended audience for data and analysis services, and client tools created to access them. Graduate and post-graduate researchers may use the client applications to easily create a suite of experiments comparing results between different datasets, parameters, and geographic scale. As our data and metadata publishing system goes into production, the metadata available for datasets and provenance information available for experiments will allow researchers to reference and publish input data or an entire experiment with parameters and explanatory annotations referenced in a peer-reviewed publication.

B. *Standards facilitate interoperability*

Running through all aspects of the Lifemapper project is a commitment to using data and communication standards. LM services adhere to well-defined standards giving developers a clear framework to work within, and providing LM users a service where issues and solutions are well documented. Metadata web services are based on the REST service model.

Lifemapper implements four OGC standards. Web Processing Service (WPS) is a standard that defines an interface for publishing geospatial processing services, defines how a client may request those services, and standardizes requests and responses. Web Mapping Service (WMS), allows

simple rendering of one or more spatial datasets. Two data services, Web Feature Service (WFS) and Web Coverage Service (WCS) return XML formatted vector data and raster datasets respectively. All of these OGC services interact with geospatial data in standard formats supported by the GDAL/OGR (http://www.gdal.org) geospatial library.

C. Metadata empowers data

An important principle underlying Lifemapper data and services is that consistent metadata should be available concurrently with LM-associated data and analyses. Accurate metadata is the cornerstone of data discovery and re-use. All static LM data will be publicly cataloged and LM web services will allow users to catalog metadata for LM-generated data and experiments with varying degrees of public access. Metadata can currently be requested for any LM-generated data in Ecological Metadata Language (EML, http://knb.ecoinformatics.org/software/eml), a format ideal for a wide range of ecological datasets [9], with plans to offer other relevant formats in the near future.

To provide a more detailed description of the procedures performed in an LM experiment, Lifemapper is extending the process module of EML. This extended EML moves Lifemapper closer to the goal of creating full provenance documents containing a history for any research experiment. The LM EML Reader module then enables re-execution with the same or modified inputs and parameters to replicate or produce variations on the documented experiment. The metadata can be published with journal articles, linking the research to the inputs and software, code or web services used to perform the processing. The LM-VisTrails and LM-QGIS plugins contain the EML Reader allowing experiments to be recreated in those software applications. Lifemapper is also expanding the EML Reader to transform LM experiment metadata into narratives, suitable for different audiences. As these EML extensions are refined, Lifemapper will submit them to the EML working group to consider for inclusion in the standard.

V. Moving Forward

A. Lessons Learned

As the Lifemapper project has matured and expanded, the importance of a flexible codebase has become increasingly apparent. The Lifemapper project follows the object-oriented programming paradigm, with particular emphasis on modularity, inheritance, and data abstraction. All code is written in Python (http://www.python.org), an open-source, cross-platform, high-level language that facilitates rapid development and easy debugging.

As the project has expanded to encompass additional data and services, we have discovered areas of the code that were overly specific. As we encounter modules that are difficult to extend, we revisit the design of the module and refactor, often creating a more complex object hierarchy, or following an accepted software design pattern [10].

Similarly, after switching to a heterogeneous cluster environment, we generalized the scheduling code that distributes analysis jobs among cluster nodes. The new design enables us to distribute different types of jobs to a variety of compute engines, both local and remote.

As a team, we have increased our cohesiveness and adaptability and clarified our shared vision by adopting a modified Scrum [11] approach (http://www.scrum.org/) to Agile software development (http://agilemanifesto.org/), which emphasizes iterative and incremental software development. We use a Trac (http://trac.edgewall.org/) wiki and issue tracking system with plug-ins integrating a Subversion code repository and Agilo for trac (http://www.agilofortrac.com/), to set goals, document decisions, establish milestones, determine the tasks and subtasks required to reach those milestones, and track timelines and progress. This system has increased accountability, while giving all team members a clear vision of the road ahead.

B. Onward

As a core component of the NSF Experimental Program to Stimulate Competitive Research (EPSCoR) Cybercommons project, LM is committed to becoming a contributing node of the NSF Data Observation Network for Earth (DataONE, http://www.dataone.org). DataONE is a $20M, 10-year collaboration among several universities (including KU, UNM, Oak Ridge National Labs, and the National Center for Atmospheric Research) whose mission is to build sustainable, long-term infrastructure for storage, indexing, discovery and access to earth observation data. Data sets cataloged within the DataONE system will be available through a set of well-defined application programming interfaces (APIs) for analytical research client packages. By implementing the DataONE APIs for data and metadata, LM will connect to a community-standards based distributed repository which will archive LM-facilitated research and modeling outputs and promote wide interoperability and integration within the computational earth science community.

As part of the ChangeThinking and LmRAD grants, our vision is to expand our educational resources to target graduate researchers as well as high school students. Our website will include guided documentation explaining and documenting previous research in SDM, algorithm strengths and weaknesses, the effect of various input parameters, limiting environmental factors, macroecological indices, species attributes affecting dispersal limits, and more. References to publications relevant to Lifemapper resources will be cited and provide a primer for students new to the field.

Our next collaboration expands the environmental data we provide for LmSDM and LmRAD to include NASA Earth observational data through a partnership with University of New Mexico (UNM) Earth Data Analysis Center (EDAC) and University of Texas at El Paso (UTEP) Cyber-ShARE Center. Cyber-ShARE provides an instrumental approach for collecting provenance information, the CI-Miner Method [12], developed at University of Texas at El Paso (UTEP). This project will instrument both EDAC and LM services to

capture end-to-end provenance within and across these two platforms.

C. Conclusion

Lifemapper's contribution to the biodiversity science infrastructure began with a simple vision of computing species distribution maps for available digital specimen data. It has grown to provide analysis and data web services to middle school students, researchers, and external applications. Lifemapper will continue to expand offerings of geospatial biodiversity data, computational resources, metadata, and research documentation in standard formats through community portals and well-publicized APIs to make data and research created with Lifemapper tools more accessible, reliable, and trustworthy.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Numbers EPS 0919443, DRL 0918590, BIO/EF 0851290, OCI 0753336. The authors are grateful for intellectual discussions with Jorge Soberón, Professor, Ecology & Evolutionary Biology, University of Kansas, Andres Lira and Narayani Barve, Graduate Students, Ecology & Evolutionary Biology, University of Kansas.

REFERENCES

[1] R.T. Fielding. "Architectural styles and the design of network-based software architectures," Doctoral dissertation, University of California, Irvine, 2000.

[2] V.P. Canhos, S. Souza, R. de Giovanni and D.A.L. Canhos. 2004. "Global biodiversity informatics: Setting the scene for a "New World" of ecological modeling," Biodiversity Informatics 1: 1-13.

[3] R. P. Anderson, D. Lew, and A. T. Peterson. "Evaluating predictive models of species' distributions: criteria for selecting optimal models," Ecological Modelling, vol. 162, pp. 211-232, 2003.

[4] H.A. Nix, "A biogeographic analysis of Australian Elapid snakes," Atlas of Elapid Snakes of Austrailia, vol. 8, R. Longmore, Ed., pp. 4-15, 1986.

[5] J. R. Busby, "BIOCLIM – A bioclimatic analysis and prediction system," Nature Conservation: Cost Effective Biological Surveys and Data Analysis, C.R. Margules and M.P. Austin, Eds., Canberra: CSIRO, 1991, pp. 64-68.

[6] S.J. Phillips, R.P. Anderson and R.E. Schapire, "Maximum entropy modeling of species geographic distributions," Ecological Modelling, vol 190, pp. 231-259, 2006.

[7] McCoy, E. D., and K. L. Heck, Jr. 1987. "Some observations on the use of taxonomic similarity in large-scale biogeography," Journal of Biogeography 14:79–87.

[8] H.T. Arita, J.A. Christen, P. Rodríguez, and J. Soberón, 2008. "Species diversity and distribution in presence-absence matrices: mathematical relationships and biological implications," The American Naturalist 172: 519-532

[9] M.B. Jones, C. Berkley, J. Bojilova and M. Schildhauer, "Managing Scientific Metadata," IEEE Internet Computing, vol. 5, no. 5, pp. 59-68, 2001.

[10] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, 1995. Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley. ISBN 0-201-63361-2.

[11] K. Schwaber, M. Beedle, 2002. Agile software development with Scrum. Prentice Hall. ISBN 0130676349.

[12] P. Pinheiro da Silva, L. Salayandia, A. Gandara, A.Q. Gates, "CI-Miner: semantically enhancing scientific processes," Earth Science Informatics vol. 2, no. 4, pp. 249-269, 2009.

.