Semantic Bridges for Biodiversity Sciences

Natalia Villanueva-Rosales[1,2], Nicholas del Rio[3,*], Deana Pennington[1,4], Luis Garnica Chavira[1,2]

[1]Cyber-ShARE Center of Excellence, [2]Department of Computer Science, [4]Department of Geology, University of Texas at El Paso, El Paso, US.
{nvillanuevarosales,ddpenninngton,lagarnicachavira}@utep.edu
[3]Air Force Research Lab, Information Directorate, Rome, US.
nicholas.del_rio@us.af.mil

Abstract. Understanding the impact of climate change and humans on biodiversity requires the retrieval and integration of heterogeneous data sets for the generation of models that provide insights not possible with a single model. Scientists invest a significant amount of time collecting and manually preprocessing data for the generation of such models. The Earth Life and Semantic Web (ELSEWeb) project aims to create a semantic-based, open-source cyberinfrastructure to automate the ingestion of data by models. This paper describes the ontologies at the backbone of ELSEWeb that provide semantic bridges between environmental data sources and species distribution models.

Keywords: Ontology, data-to-model integration, model web, biodiversity, climate change.

1 Introduction

What will happen to native species in national parks under scenarios of climate change? When and where might we expect zoonotic infectious disease to spread? These questions and others can be addressed using species distribution models (SDMs) [1]. SDMs predict where animal or plant species might find suitable habitat given present conditions or under change scenarios. Species might be socially relevant because they are endangered or carry diseases. Conducting "what-if" analyses provides insights of changes in the environment as they occur – or before they occur. Scenario analysis is becoming a key tool for the biodiversity (and other) sciences [2] to understand human impacts on the environment coupled with climate change. SDMs require species occurrence data and environmental data. Species occurrence data contains the location of known occurrences of a species in a given time period. Species occurrence data is mostly available in museums, which have invested in digitization and the development of metadata standards and repositories through the Global Bio-

[*] Affiliated with the University of Texas at El Paso when producing this work.

diversity Information Facility (GBIF)[1]. Environmental data is heterogeneous and available from multiple sources, e.g., satellite imagery. Scientists spend considerable time deciding what data might be most relevant, where to find it, how to obtain it, and what to do with it since data manipulation may require the use of proprietary tools. In addition, each modeling algorithm has its own constraints, operational requirements, assumptions, parameter and data requirements, usage history, and advocates (or dissidents). Data are required to be further manipulated and assembled into the formats and scales consistent with the selected algorithm. The GEO Model Web initiative [3] aims to increase access to models and interoperability across models, databases and websites. The Model Web advocates the creation of infrastructure with four underlying principles: open access, minimal barriers to entry, service-driven, and scalability. The Earth, Life and Semantic Web (ELSEWeb) project aims to enable interoperability between data and model providers in a way that data can be automatically retrieved and ingested by a model seamlessly to the user. ELSEWeb follows the four principles of the Model Web. The current implementation of ELSEWeb integrates data sets from the University of New Mexico Earth Data Analysis Center (EDAC)[2] with the species distribution modeling service provider Lifemapper (Lifemapper)[3].

2 Semantic Descriptions of Data and Model Providers

Ontologies in ELSEWeb describe concepts needed to automate the retrieval and manipulation of data for the generation of SDMs, to advance our understanding of how to automatically chain models together – a requirement of the Model Web. These ontologies were created using an iterative, bottom-up approach driven by standards and best practices followed by EDAC and Lifemapper, such as those provided by the Open Geospatial Consortium (OGC)[4] and the Federal Geographic Data Committee (FGDC)[5]. Non-semantic resources were inspected, interpreted, and expressed as classes and properties using the Web Ontology Language (OWL)[6]. These resources included Web Services' metadata, arbitrary REST-full service descriptions in XML, shell scripts and Java code. ELSEWeb's initial semantic descriptions focused on technical requirements (e.g., format) to enable the seamless integration of data and models from EDAC and Lifemapper respectively [4]. These ontologies were iteratively refined with expertise knowledge from EDAC and Lifemapper partners and aligned to upper-level, and community accepted ontologies and vocabularies such as the Provenance Ontology (PROV-O)[7]. In this paper, ELSEWeb's ontology names, classes and properties are denoted in *italics*.

---

[1] `http://www.gbif.org`
[2] `http://edac.unm.edu`
[3] `http://lifemapper.org/`
[4] `http://www.opengeospatial.org/standards`
[5] `https://www.fgdc.gov/`
[6] `http://www.w3.org/2002/07/owl#`
[7] `http://www.w3.org/ns/prov-o`

2.1 ELSEWeb's Environmental Data Ontology

The *elseweb-data* ontology provides concepts to describe datasets with characteristics relevant to SDMs such as spatial/temporal dimensions. The *elseweb-data* ontology covers remote sensing environmental data, spatial in nature, geo-referenced, and measured from some instrument such as a MODIS sensor[8]. Environmental data represent any phenomenon that might be important for providing required habitat for a given species or for constraining the ability of a species to survive such as vegetation type. These biophysical data are usually combined with climate data that often establish thresholds of survival. Fig. 1 illustrates the core classes of the *elseweb-data* ontology with blue nodes and the ontologies it extends. The *elseweb-data* ontology extends the Data Catalog Vocabulary DCAT[9] to describe the temporal and spatial coverage of data (e.g., *Geographic Region)*, the corresponding *Theme* (e.g., *Vegetation*), and how the data can be accessed (e.g. a *File Manifestation* that describes the format of a file and where it can be downloaded).

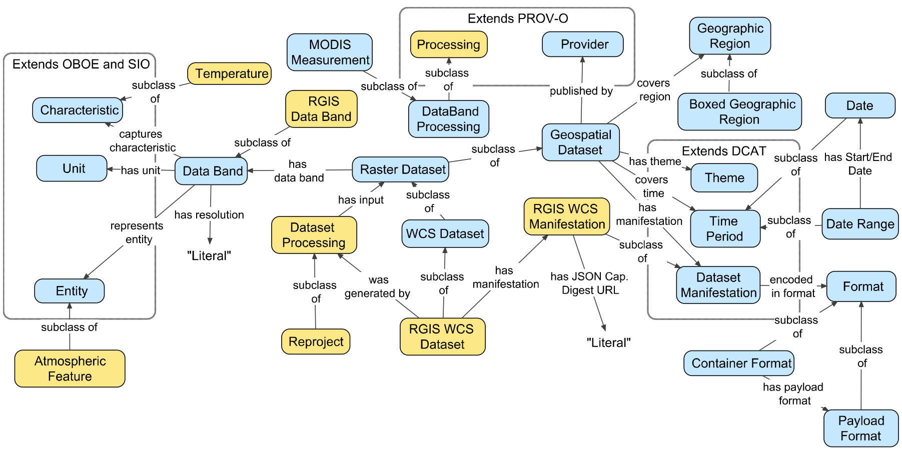

Fig. 1. The *elseweb-data* ontology (blue nodes) provides concepts for describing characteristics of a dataset in a biodiversity scenario: spatial/temporal dimensions, how it can be accessed and the focal entity described. The *elseweb-edac* ontology (yellow nodes), extends and instantiates the *elseweb-data* ontology to describe data sets published by EDAC. Dashed boxes outline concepts of the ontology that extend existing ontologies and vocabularies.

With respect to spatial coverage, ELSEWeb describes a *Boxed Geographic Region* specified by left longitude, lower latitude, right longitude and upper latitude. This notion follows OGC standards and best practices for Web Services. More generally, OGC Web Coverage Services (WCS) metadata describes how to invoke services for data retrieval, their corresponding spatial-temporal coverage, and the different kinds of formats available for the retrieved file (e.g., PNG or TIFF). Additionally, the OGC metadata schema defines extension points which can be used to reference additional

[8] `http://modis.gsfc.nasa.gov/`
[9] `http://www.w3.org/ns/dcat`

metadata. ELSEWeb's data provider, EDAC, relies on these extension points to include FGDC metadata in their services. The design of the *elseweb-data* ontology was largely inspired by this conglomerate-service metadata employed by EDAC. The most specific concepts of *elseweb-data* ontology describe different kinds of Geospatial data sets. For example, a *Raster Dataset* consists of a set of *DataBand*s. A *DataBand* captures a *Characteristic* (e.g., *Temperature*) of an *Entity* (e.g., *Air*), with a specific *Unit* (e.g., Fahrenheit) and *Resolution* (e.g. 250 m). By extending the Extensible Observation Ontology (OBOE)[10], *elseweb-data* can answer questions about measurements contained in specific *DataBands* or the observed focal entity. By extending the PROV-O, *elseweb-data* can answer questions about data providers and how the data was generated.

The *elseweb-data* ontology is extended and populated with data provided by EDAC in the *elseweb-edac* ontology, illustrated by yellow nodes in Fig. 1. For example, information about data sets published by EDAC instantiates the class *RGIS Data Band*, a subclass of the generic *Data Band*. Fig. 1 illustrates that EDAC also offers data processing services, such as *Reproject*, used by EDAC when publishing a *Raster* dataset. Provenance information indicates, for example, that EDAC's data is retrieved from the US National Aeronautics and Space Administration (NASA) and transformed following EDAC's quality assurance guidelines. Once the dataset is published as a service, it becomes an *RGIS WCS Service* that has a corresponding *RGIS WCS Manifestation* with information on how to retrieve the data. Appropriate classes and properties of *elseweb-edac* extend PROV-O classes such as PROV-O Activity. EDAC's geospatial data is exposed through OGC Web Map, Web Feature and Web Coverage Services. EDAC's Web Coverage Services' metadata is leveraged by ELSEWeb's harvester to automatically populate *elseweb-edac*.

2.2 ELSEWeb's Model Service Ontology

The *elseweb-model* ontology, also in OWL, describes biodiversity modeling services as implemented algorithms, parameters, and data inputs/outputs. Fig. 2 illustrates the core classes of this ontology with blue nodes. Concepts in *elseweb-model* are used to describe a modeling algorithm (e.g. *Species Modeling Algorithm*) and corresponding parameters (e.g., *Species Modeling Parameter*) extend the Semantic Science Integrated Ontology (SIO)[11]. Concepts describing services, inputs, and outputs are inherited from the Semantic Automated Discovery and Integration (SADI) framework[12] and PROV-O. For example, in the *elseweb-model* ontology, a SADI Service Agent is also a PROV-O Agent. This model allows ELSEWeb to describe a computational process with an *Activit*y associated with a SADI *Service*, which is also a PROV-O Agent. The combination of PROV-O and SADI provides a unified view to describe discovery, run time, and provenance metadata.

---

[10] `http://ecoinformatics.org/oboe/oboe.1.0/oboe-core.owl`
[11] `http://semanticscience.org/ontology/sio.owl`
[12] `http://sadiframework.org/`

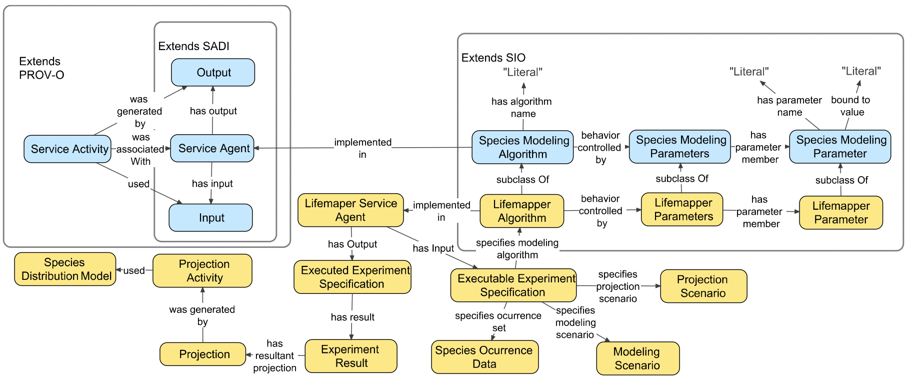

Fig. 2. The *elseweb-model* ontology (blue nodes) provides concepts for describing biodiversity modeling algorithms, their inputs, outputs, parameters, and services providing these algorithms. The *elseweb-lifemapper* ontology (yellow nodes), extends and instantiates the *elseweb-model* ontology to describe SDM services provided by Lifemapper. Dashed boxes outline concepts of the ontology that extend existing ontologies.

The *elseweb-lifemapper* ontology, illustrated by yellow nodes in Fig. 2, extends and instantiates the *elseweb-model* ontology with information about the modeling services provided by Lifemapper. *Elseweb-lifemapper* describes the specific inputs and outputs generated by a modeling service. For example, a *Lifemapper Service Agent* requires an *Executable Experiment Specification* as an input. An *Executable Experiment Specification* is composed of *Species Occurrence Data*, *Modeling Scenario* (e.g., a stack of environmental data used to generate an SDM), *Projection Scenario* (e.g. a stack of environmental data used to generate a projection of the species distribution by applying the SDM to changed conditions), and an *Algorithm Specification* (e.g., *Maximum Entropy)*. The *elseweb-lifemapper* ontology also contains concepts that describe how a projection was generated by including the *Projection* and *Species Distribution Model* generation activities. ELSEWeb provides users with provenance to understand how their specific modeling and projection scenarios were used to create a specific SDM. Lifemapper's Web Services expose XML files with metadata about available algorithms and their parameters. The *elseweb-lifemapper* ontology is automatically populated using a Java-based harvester that leverages this metadata.

2.3 Semantic Bridges from Data to Model Providers

Scientists often write custom scripts in order to move, augment, and transform source data into alternative forms that suit the needs of target analytical tools [5]. These custom scripts can often lack explicit documentation about data ingestion, transformation processes, and generated outputs, making these scripts difficult to reuse. To facilitate reuse of data transformation services, ELSEWeb creates semantic bridges (i.e., alignments) between the *elseweb-edac* ontology and the *elseweb-lifemapper* ontology. These bridges, specified in the *elseweb-mappings* ontology, provide a declarative, formal specification that can automate the data transformations required to generate an SDM. The *elseweb-mappings* ontology was specified by ELSEWeb team using an iterative and bottom-up, two-stage process. First, an ad-hoc workflow composed of shell script and Java widgets that performed the necessary transformations was created. Then, this workflow was promoted from the procedural program level to the declarative level using OWL formalisms. For example, an *RGIS WCS Dataset* published by EDAC can be automatically classified as part of an *Executable Experiment Specification*, which in turn is described as the input of a *Lifemapper Service Agent*.

3 Results

ELSEWeb's ontologies are stored in an instance of Virtuoso triple store[13] with over 350,000 triples and available through an SPARQL endpoint[14]. Additional resources can be found at the project's website[15]. The linked data section of the website includes SPARQL sample queries to navigate ELSEWeb's knowledge base and answer competency questions. The ontology section provides links to ontology files and additional diagrams. Ontologies can also be retrieved under their designated namespace[16]. ELSEWeb currently enables the integration of over 6,650 environmental data sets, 1000 species occurrence data sets and 11 algorithms for the generation of SDMs.

Users can generate SDMs by creating experiments through the Graphical User Interface (GUI)[17]. Interested readers can test ELSEWeb by accessing the demo section. After submitting an experiment, SDMs can be retrieved at Lifemapper with the credentials user:elseweb2, password:elsewebtwo. ELSEWeb's GUI also allows the manual submission of experiments through a JSON specification. The service-oriented architecture used in ELSEWeb's infrastructure, GUI design, and a discussion of the technical challenges addressed by the ELSEWeb framework can be found in [6].

Existing semantic web service frameworks such as SADI can leverage semantic bridges and orchestrate the execution of Web Services to dynamically generate the output required. In ELSEWeb, SADI services are used to retrieve and transform data and generate SDMs along with a provenance trace. SADI services semantically describe EDAC's Web Services to retrieve data requested by the user as a measured characteristic with time and spatial constraints. SADI services semantically describe the inputs and outputs of services at Lifemapper that generate SDMs. ELSEWeb uses the SHARE client[18] provided by the SADI framework, to automatically orchestrate the execution of SADI services. Interested readers may refer to [4] for an in-depth description of ELSEWeb's SADI services and the service orchestration process.

---

13 `http://www.openlinksw.com/`
14 `http://visko.cybershare.utep.edu/sparql`
15 `http://elseweb.cybershare.utep.edu/`
16 `http://ontology.cybershare.utep.edu/ELSEWeb/`
17 `http://elseweb.cybershare.utep.edu/experiments`
18 `http://biordf.net/cardioSHARE/`

4 Related Work

Efforts towards the integration of heterogeneous data for the creation of biodiversity models include eHabitat [7] and iPlant Collaborative[19]. eHabitat implements the Model Web principles using OGC Web Processing Services for multi-purpose modeling, including ecological forecasting under climatic or development scenarios. To the best of our knowledge, eHabitat does not make use of ontologies or generic reasoners for service orchestration. Instead, eHabitat provides web clients where users can manually orchestrate service execution. The iPlant Semantic Web Platform enables the semantic discovery of services and service orchestration using the Simple Semantic Web Architecture Protocol (SSWAP) and Resource Description Graphs [8]. In contrast to ELSEWeb, service orchestration in iPlant is a manual task, facilitated by a GUI where a reasoner leverages service descriptions, encoded in ontologies, to suggest the next service in the pipeline. iPlant Collaborative efforts also include the development of ontologies to support biodiversity knowledge discovery [9]. These ontologies provide complementary concepts that can be used to further inform the models generated at ELSEWeb. In particular, the Biological Collections Ontology (BCO) describes specimen collections and sampling processes, the Environmental Ontology (ENVO) describes habitats, environmental features and materials, and the Population and Community Ontology (PCO) describes communities of biological entities and their qualities and interactions, among other concepts.

5 Conclusions and Future Work

Easier and more efficient approaches for harnessing the vast array of scientific data are essential for the generation of biodiversity models. The ontologies developed in ELSEWeb aim to facilitate the use of data and data transformation and modeling services following the principles of the Model Web. ELSEWeb's ontologies were created by an interdisciplinary group of researchers and developers following an iterative, bottom-up approach driven by community standards and best practices. Semantically described data and models exposed on the Semantic Web enable frameworks like SADI to automatically find, retrieve, and manipulate data to generate SDMs. These capabilities, however, do not address the issues involved in determining whether data and model integration is sensible scientifically even if it can be accomplished technically. In part this is due to the sheer enormity of the task, given the wide range and diversity of concepts employed in any given biodiversity problem. Related efforts towards creating ontologies for biodiversity knowledge discovery include the development of BCO, ENVO and PCO. Aligning ELSEWeb's ontologies with such ontologies may lead to a more robust integration of domain and technical concepts, enable the automated verification of models, and provide a more comprehensive provenance trace. Future work of ELSEWeb includes the integration of additional data sets and services for the generation of water models. Water models will provide the opportuni-

19 http://www.iplantcollaborative.org/

ty to test ELSEWeb with members of academia, industry, society and government to facilitate interdisciplinary collaborations.

Acknowledgements. ELSEWeb was funded by NASA ACCESS grants NNX12AF49A (UTEP), NNX12AF52A (UNM), and NNX12AF45A (KU). This work used resources from Cyber-ShARE Center of Excellence supported by NSF grant HRD-0734825 and HRD-1242122. The authors thank Soren Scott and Karl Benedict (EDAC), and CJ Grady and Aimee Stewart (Lifemapper) for their contributions to the design of ELSEWeb's ontologies and the anonymous reviewers for their insightful comments on this manuscript.

References

1. Krause, C., Pennington, D.: Strategic decisions in conservation: Using species distribution modeling to match ecological requirements to available habitat. In: Maschinski, J. and Haskins, K. (eds.) Plant Reintroduction in a Changing Climate: Promises and Perils. p. 432. Island Press, Washington, D.C. (2012).
2. Settele, J., Carter, T.R., KüHN, I., Spangenberg, J.H., Sykes, M.T.: Scenarios as a tool for large-scale ecological research: experiences and legacy of the ALARM project. Glob. Ecol. Biogeogr. 21, 1–4 (2012).
3. Nativi, S., Mazzetti, P., Geller, G.N.: Environmental model access and interoperability: The GEO Model Web initiative. Environ. Model. Softw. 39, 214–228 (2013).
4. Del Rio, N., Villanueva-Rosales, N., Pennington, D., Benedict, K., Stewart, A., Grady, C.: ELSEWeb meets SADI: Supporting data-to-model integration for biodiversity forecasting. In: AAAI Fall Symposium on Discovery Informatics (2013).
5. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise data analysis and visualization: An interview study. Vis. Comput. Graph. IEEE Trans. On. 18, 2917–2926 (2012).
6. Villanueva-Rosales, N., Garnica Chavira, L., Del Rio, N., Pennington, D.: eScience through the Integration of Data and Models: A Biodiversity Scenario. In: 11th IEEE International Conference on eScience (2015).
7. Dubois, G., Schulz, M., Skøien, J., Bastin, L., Peedell, S.: eHabitat, a multi-purpose Web Processing Service for ecological modeling. Environ. Model. Softw. 41, 123–133 (2013).
8. Gessler, D.D., Bulka, B., Sirin, E., Vasquez-Gross, H., Yu, J., Wegrzyn, J.: iPlant SSWAP (Simple Semantic Web Architecture and Protocol) enables semantic pipelines for biodiversity. Semant. Biodivers. S4BioDiv 2013. 101 (2013).
9. Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L., Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., Tuama, É.Ó., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wieczorek, J., Whitacre, J., Wooley, J.: Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE. 9, e89606 (2014).