Discovery Informatics: AI Takes a Science-Centered View on Big Data
AAAI Technical Report FS-13-01

ELSEWeb Meets SADI: Supporting Data-to-Model Integration for Biodiversity Forecasting

Nicholas Del Rio[1], Natalia Villanueva-Rosales[1], Deana Pennington[1], Karl Benedict[2]

Aimee Stewart[3], C. J. Grady[3]

[1]Cyber-Share Center of Excellence, University of Texas at El Paso, 500 W University Ave, El Paso, TX 79968
[2]Earth Data Analysis Center, MSC01 1110, 1 University of New Mexico, Albuquerque, NM 87131
[3]University of Kansas Biodiversity Institute, 1345 Jayhawk Blvd., Lawrence, KS, 66045-7593

Abstract

In this paper, we describe the approach of the Earth, Life and Semantic Web (ELSEWeb) project that facilitates the discovery and transformation of Earth observation data sources for the creation of species distribution models (data-to-model) transformations. ELSEWeb automates the discovery and processing of voluminous, heterogeneous satellite imagery and other geospatial data available at the Earth Data Analysis Center to be included in Lifemapper Species Distribution models by using AI knowledge representation and reasoning techniques developed by the Semantic Web community. The realization of the ELSEWeb semantic infrastructure provides the possibility of combinatoric explosions of scientific results, automatically generated by orchestrations of data mash-ups and service composition. We report on the key elements that contributed to the ELSEWeb project and the role of automated reasoning in streamlining the Species Distribution Model generation and execution.

1 Introduction

Biodiversity scientists are grappling with understanding potential climate and human impacts on biodiversity (Barnosky et al. 2011). There is much uncertainty involved - in what changes are likely to occur, how those changes interact with species, and how species interact with each other. In recent years numerous scientific efforts around the world have generated data and models necessary for biodiversity analyses. Indeed, there is a plethora of data and models to choose from, each with unique characteristics. There have been concerted efforts to standardize data and models to achieve interoperability. In particular, the Global Biodiversity Information Facility (GBIF; http://www.gbif.org) was established by governments in 2001 to provide access to species observation data. Similarly, the Group on Earth Observations (GEO; http://www.earthobservations.org) is a partnership of governments and international organizations creating a System of Systems (GEOSS) that aims to connect Earth observation data and tools. The GEOSS Model Web is an envisioned infrastructure to facilitate easy integration of data and models (Geller and Melton 2008; Nativi, Mazzetti, and Geller 2012). Yet progress is slow, much of the legacy data remain difficult to discover, and the relevant environmental data are often voluminous, heterogeneous, and require specialized expertise to work with. For example, an investigation into the combined impact of climate change and population growth on plant biodiversity in a region would require integrated analysis of data from climate change models, population models, species distribution models, and water models (since water availability drives plant productivity and is impacted by climate and population change). However, each of these model types are created by independent scientific communities that typically produce many model variations, each with its own input requirements. It is difficult to know what models are even available in a domain, much less how to access and use them. Data requirements typically include geospatial and satellite imagery that require specialized expertise and tools to work with. When one considers a complex, integrated investigation across domains such as the above example, significant challenges arise in 1) discovering relevant models, and 2) efficiently connecting them through performing the necessary data transformations to fit output data from one model into the format required by another model. Therefore, scientists who desire to conduct a particular analysis still typically use the models they are already familiar with and invest much of their research time collecting or finding relevant data, preprocessing those data into forms that can be input into the models, and manually transforming output data into the required input form of the next model. Hence, the significant amount of work involved means that they commonly conduct their analyses using a specific set of assumptions, data, and parameterizations based on the requirements of a single model or set of models. However, given that there is no agreement in any of these scientific comains on a best model, or even a few best models, a better characterization of uncertainty could be achieved by iterating over many combinations of data, models, and parameterizations. The goal of the Earth, Life, and Semantic Web (ELSEWeb) project was to enable scientists to easily conduct these kinds of iterative analyses - creating an infrastructure for them to employ "If-Then-ELSE" mechanisms in their research (e.g., on-the-fly hypotheis testing and result comparison).

2 Earth, Life, and Semantic Web

The Earth, Life and Semantic Web (ELSEWeb) project was funded in 2012 through the NASA ACCESS program. The goal of ELSEWeb is to develop generalizable semantic approaches to data and model integration that enable scientists to more easily conduct integrated modeling investigations by automating the transformation of data to fit model input requirements. In addition, since the system and not the scientist will select and perform transformations, it is necessary to automatically capture provenance across all models and data transformations that are executed and provide an easily interpreted trace. Two existing environmental data and model providers are collaborating with semantic experts to provide a testbed for semantic approaches. The University of New Mexico Earth Data Analysis Center (EDAC; http://edac.unm.edu) specializes in satellite imagery and GIS data commonly used in modeling land surface and environmental change, and provides numerous web-service based data transformation tools. The University of Kansas Lifemapper project (http://lifemapper.org) provides web services for species distribution modeling (SDM). SDM are one approach to projecting the effect of climate change on biogeography, and are widely regarded as the best available tool for producing species specific information necessary in conservation planning (Hannah 2003). Species distribution models integrate a wide range of environmental data to predict potential habitat for a species based on conditions where it is known to presently occur (see (Franklin 2009) for detailed information on SDM). Lifemapper SDM (LmSDM) is a set of web services that project potential future species distributions, see Figure 1, from specimen occurrences and environmental data such as bioclimatic data derived from Worldclim (http://www.worldclim.org) and the International Panel on Climate Change (IPCC) climate models (e.g. temperature, precipitation). LmSDM lacked the ability to easily incorporate user-selected environmental datasets such as the geospatial and satellite imagery provided by EDAC (land cover, soil type, water depth) into the analysis. Users must locate relevant environmental datasets, prepare them in whatever way necessary for ingestion into Lifemapper, and manually upload them into LmSDM. A key feature of EDAC and Lifemapper is the use of open standards at both sites, creating the opportunity for automated data to model integration if the disparities between EDAC data and Lifemapper requirements could be identified and dealt with systematically.

2.1 Use Case

The use case currently supported by ELSEWeb is discovery and transformation of data at EDAC that 1) covers a region of interest by place name or geographic coordinates; 2) is derived from particular instruments; and 3) has a particular semantic data type. These semantic data types, programmatically inserted as thematically defined keywords into the metadata published by EDAC, are a key element for enabling ELSEWeb and are not currently specified as components in the community service or data metadata standards, for example Open Geospatial Consortium (OGC, http://www.opengeospatial.org/) or Federal Geographic

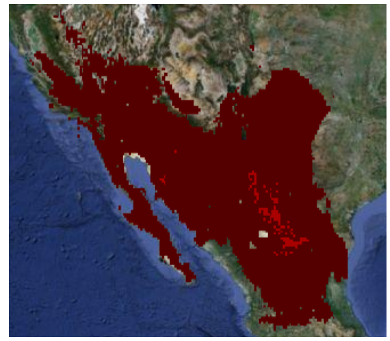

Figure 1: Projected distribution of Larrea tridentata (dark red) modeled from environmental characteristics at known occurrence points (orange) using Services available at Lifemapper

Data Committee (FGDC, http://www.fgdc.gov/metadata). Nevertheless, any data provider that reuses the semantic data types at EDAC or maps their own semantic data types to EDAC's could easily be integrated into the ELSEWeb system. More information about the semantic metadata extensions are described in the next section. At Lifemapper, ELSEWeb enables streamlined ingestion of a range of data types to supplement the bioclimatic change data already available. The bioclimatic data that Lifemapper uses has been preprocessed from raw climate change data to generate more biologically meaningful data. For instance, many species are constrained by the minimum nighttime temperatures. If a region is too cold at night during winter, species cannot survive. But the semantic meaning of "winter" is not based on a time of year; rather, it varies depending on global location. Hence, data discovery and integration based on the desired characteristics (e.g. coldest/warmest month, or wettest/driest month) rather than time of year requires the use of semantic descriptions.

2.2 Earth Data Analysis Center Data

EDAC's large collection (over 280,000 individual datasets comprising over 1 billion features) of environmental and geospatial data are made available as OGC Web Map, Web Feature and Web Coverage Services (WMS, WFS and WCS respectively), which are REST-like web services accessible using HTTP Get requests. The published WCS services are the focus of the data publication capabilities supported in the ELSEWeb project. In general, WCS services advertise capabilities in a GetCapabilities XML document that describes the data layers (coverages) available for download, the region encompassing the data layers and the different formats in which the data can be returned (e.g., PNG, JPEG, and TIFF). From the GetCapabilities XML, clients are able to create URLs that specify what data layer is being requested, the subregion that should be returned, any resampling or coordinate transformation that should be performed, and how the returned data should be encoded. Although GetCapabilities XML describes many of aspects of the data needed to support our use case, the information is not specified at a semantic level and therefore may not be easily integrated with other non-OGC specific datasets using different nomenclature. Upon submission of a WCS request URL, services return the specified data in a multipart MIME format, as per the specifications set forth by OGC. These multipart mime messages contain two parts: an XML metadata description of the data returned (e.g., size and encoding) and the actual data payload.

The GetCapabilities XML schema was designed to be extendable in order to allow publishers to describe additional metadata pertinent for specific domains. For example, service publishers can include information such as the semantic type of the data layers, the duration the data was collected as well as the sensor responsible for collecting the data if the data are remotely sensed. In particular, EDAC currently publishes metadata using the Federal Geographic Data Committee (FGDC) (Committee and others 1998) metadata standard to describe both the semantic type of the data as well as information about the collecting instrument (e.g., sensor) and provides links to these metadata from the GetCapabilities XML document. The FGDC metadata extensions are associated with the OGC Web Services XML element: ogc:Metadata. Semantic data type descriptions are expressed using the Climate Forecast (CF, http://cf-pcmdi.llnl.gov/) terminology and these semantic descriptions are referenced from within the FGDC metadata title element (Eaton et al. 2003). Therefore, EDAC employs a family of three metadata standards to describe their WCS services: OGC GetCapabilities XML, FGDC, and CF. Looking forward, EDAC has implemented support for the ISO 19115, 19110, and 19119 set of data and service metadata standards as a complement to the standards currently supported for the ELSEWeb project.

2.3 Lifemapper Experiment Requirements

Lifemapper modeling is also available as a RESTful service accessible using HTTP Post. The interface to Lifemapper is well defined using the Web Application Description Language (WADL), which describes the input and output schema of the XML HTTP payloads. The basic constructs of the Experiment.xml is composed of a "Scenario Layer Set", consisting of references to TIFF environmental data such as temperature or rainfall; passing by reference is a useful technique in order to keep message payloads from exploding because of base64 encoding of binary data. An experiment also specifies species occurrence data whose predicted distribution will be calculated from the environmental scenario specified. Finally, the specific modeling algorithm that will be used to calculate the predicted distribution, such as BIOCLIM (Busby 1991), is specified.

After successful submission of an experiment, Lifemapper generates a species distribution model (SDM) and predicted species distribution map and returns a URL referencing the newly generated SDM metadata and map. The model is processed with one set (observed climate), the projections (observed and future predicted climate) with 10 sets for totals of approximately 300GB and 3TB, which can be considered "Big Data". Outputs are 1.5GB each. The map can be returned from the Lifemapper website as an image, shown in Figure 1 using OGC Web Mapping Services (WMS), or imported into Quantum GIS (QGIS; http://www.qgis.org) and VisTrails (http://www.vistrails.org) systems using the Lifemapper plug-in support. QGIS supports scenarios such as when users want more sophisticated visualizations than the default map images, to examine the data, or perform additional operations on the LmSDM such as aggregations with other models and statistical analyses.

2.4 Challenges to Generate Lifemapper Models using EDAC Data

The disparity between the (1) forms of available EDAC data and (2) Lifemapper data ingestion requirements requires a more sophisticated process than data discovery alone. Note that the data returned from EDAC's WCS services cannot be directly referenced as a scenario layer set due to the multipart message format that is not supported by Lifemapper. The ELSEWeb integration, described in the next section, extends discovery and introduces data aggregation, sequencing, and format transformations which are operations necessary for appropriately structuring EDAC WCS service responses to satisfy Lifemapper data requirements.

Figure 2 highlights the necessary transformations required to structure EDAC gridded WCS data into scenario layer sequences required by Lifemapper:

1. Search through (thousands of) WCS services provided by EDAC and identify a relevant subset:
   (a) Read through XML if metadata is exposed in its raw form
   (b) Map states and geographical regions to latitude and longitude bounding boxes
   (c) Read through cryptic satellite/sensor labels
   (d) Read through date ranges, possibly confounded within the label of the service name itself
2. Generate a WCS calling sequence
3. Execute each selected WCS service
4. Extract data payloads from the multipart messages
5. Construct Lifemapper Experiment.xml
   (a) Select species occurrence set
   (b) Select algorithm
   (c) Embed reference to TIFF URL Sequence (i.e., scenario layers)
6. Invoke Lifemapper by requesting an SDM provided the Experiment.xml

Although these tasks could be hard-coded into a workflow, the resulting software may not be easily extendible to include other data sources or other model providers, the key target for the envisioned Model Web where scientists can easily mix and match disparate models. If the workflow was extended to include other data sources or models, the resulting specification may become complex and difficult to maintain due to the high number of various data formats and modeling capabilities currently published on the Web. One goal of ELSEWeb is to provide scalable solutions for other providers (e.g, data or models) that might wish to "plug-in" into ELSEWeb infrastructure. In order to develop an infrastructure that accommodates this flexibility, we turned to semantic web technologies that can be configured to automat-

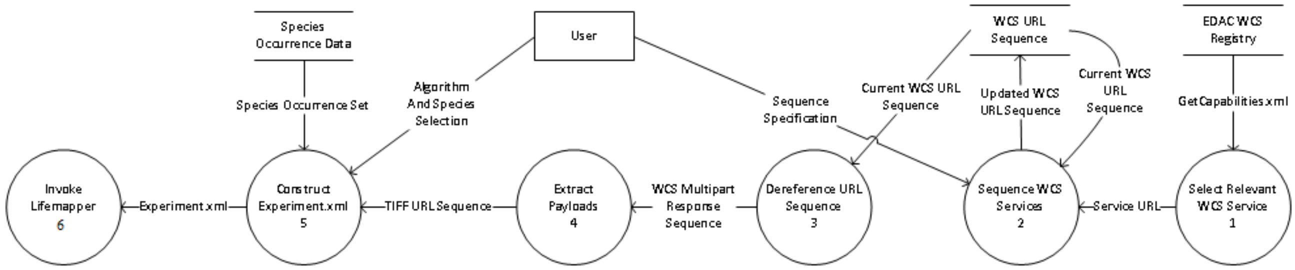

Figure 2: The workflow that streamlines EDAC data into Lifemapper modeling services

ically mitigate disparities between data providers and modeling services.

3 The ELSEWeb Approach

ELSEWeb enables LmSDM users to automatically integrate EDAC data into SDMs using the Semantic Automated Data Integration framework (SADI; http://sadiframework.org) (Wilkinson et al. 2011) to support the specific task of automatically transforming data from EDAC to fit input requirements of Lifemapper.

Users in ELSEWeb request for the generation of species distribution models by specifying SPARQL (PrudHommeaux, Seaborne, and others 2008) queries, which are satisfyied by the SHARE client (Vandervalk, McCarthy, and Wilkinson 2010) provided by the SADI framework. SHARE executes ELSEWeb SADI services and aggregates resulting RDF output to compose an Experiment knowledge base, which contains the information needed to answer a specific SPARQL query. SHARE relies on the ELSEWeb knowledge base that contains SADI service descriptions that wrap EDAC and Lifemapper services in order to formulate service execution plans that will generate the minimal subset of RDF needed to satisfy a specific query. Figure 3 presents a data flow perspective of ELSEWeb, where the interfaces between SADI/SHARE, Lifemapper, and EDAC are visualized. The use of the SADI framework is further detailed in the following subsections.

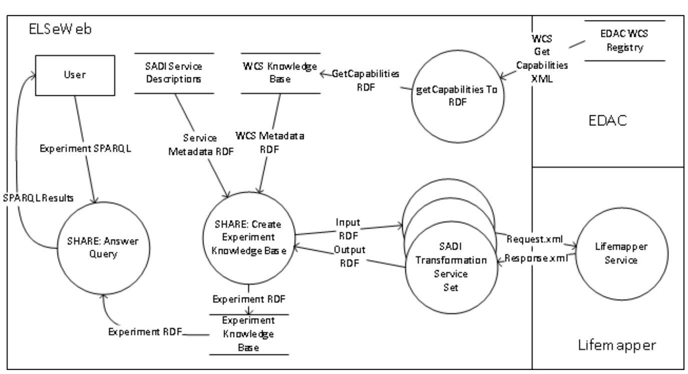

Figure 3: Data flow representation of the ELSEWeb infrastructure

3.1 Data Discovery and Mashup with the SADI Framework

The ELSEWeb project leveraged the SADI framework to expose services providing data and modeling services from EDAC and SDM services from Lifemapper. SADI uses standards-compliant Web languages and Semantic Web service patterns to exchange RDF resources (Wilkinson et al. 2011). SADI services are defined in terms of the input and output class descriptions using the Web Ontology Language (OWL) (McGuinness, Van Harmelen, and others 2004) corresponding to the instances they consume and produce respectively. Every service provides explicit relations between the output data and the input data through RDF properties. The SADI framework includes APIs and plug-in tools to facilitate the generation of these services by non-programmers, e.g., the SADI Taverna plug-in (Withers et al. 2010) and the Protege plug-in (Wilkinson et al. 2010). SADI has been used in the biomedical domain, in particular, to enable data integration for personalized medicine (Vandervalk, McCarthy, and Wilkinson 2010; Vandervalk et al. 2013) and the integration of biological data (Riazanov et al. 2012; Callahan et al. 2013).

To exemplify our use of SADI in ELSEWeb, consider the Extract Payloads activity that extracts TIFF payloads from EDAC WCS service responses. In our workflow diagram, the input to payload extraction is a sequence of WCS service request URLs and the output is a sequence of TIFF URLs. The SADI WCSPayload Extractor ingests `WCSCoverage Sequences`, which are composed of (e.g., hasWCSCoverage) a sequence of maximum ten ordered `WCSCoverages`. The `WCSCoverage` class is elaborated on in Figure 5. Provided input, WCSPayload Extractor iterates through the `WCSCoverage Sequence` and dereferences each WCS URL specified. The service then extracts the returned payload from the WCS responses and constructs the output `WCSPayload Sequence`, which references ten URLs of the TIFF payload data that was extracted from the WCS responses.

The Lifemapper SADI service, presented in Figure 4, wraps the Lifemapper RESTful application. Lifemapper SADI service input is constrainted by the definition of the OWL class `FullySpecifiedExperiment`, which specifies an algorithm name and an ID of a species occurrence set. The Lifemapper SADI service extracts the required information from the input RDF and constructs the

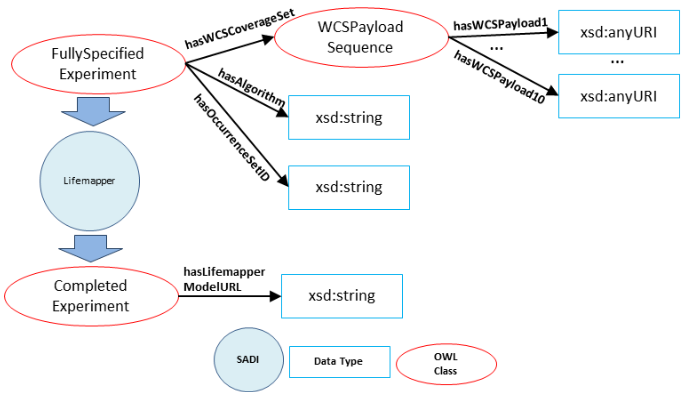

Figure 4: Input and Output Interface of the SADI Lifemapper Service

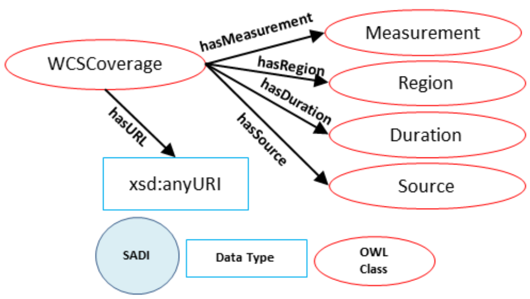

Figure 5: A depiction of the WCSCoverage OWL specification of EDAC SADI service outputs

XML experiment request for the Lifemapper RESTfull service. The SADI service also intercepts the Lifemapper response, which is a URL to a generated species distribution models, and creates the output RDF which in this case is an Experiment Stage 4, that simply references the SDM URL. Therefore, the SADI Lifemapper service encompasses both step 5 and 6 of the ELSEWeb workflow in Figure 2.

Other SADI services composing ELSEWeb are those that wrap the EDAC WCS services. These SADI services generate output RDF of type WCSCoverage, shown in Figure 5. WCSCoverages specify what source or sensor recorded the data, the measurement type (i.e., semantic type) of the data, the bounding box region containing the data and the date the data was collected or recorded. Additionally, ELSEWeb contains a sequencing service that constrcutes the WCS Coverage Sequence required by the payload extraction service. The input and output descriptions of our ELSEWeb SADI services comprise our WCS and SADI service knowledge bases presented in Figure 3

3.2 Service Orchestration using SHARE

The SHARE client is a prototype tool provided by the SADI framework, which consumes SPARQL queries and maps each predicate (or relationship) involved in these queries to a SADI service that can provide such relation (Wilkinson et al. 2011). Each SPARQL query in ELSEWeb executes services for generating, aggregating and transforming data, creating models and explicitly asserting the relations between them in an ad-hoc manner via service orchestration. The SHARE client relies on a description logics reasoner for the automated service orchestration by accessing the ELSEWeb SADI registry. Although the SADI framework allows the registration of services into the global SADI registry, we opted for creating an ELSEWeb specific registry that currently contains only services supporting the required aggregations and transformations required by ELSEWeb and avoids unnecessarily searching through hundreds of SADI services from the biomedical domain registered in the global registry.

Below is an example ELSEWeb specific SPARQL query that requests for the generation of a Lifemapper model provided a set of experimental constraints. The experiment constraints define what WCSCoverages should be bound to each scenario layer, the algorithm that should be employed in Lifemapper, and the species occurrence ID. These constraints are defined in OWL and referenced in the query using the FROM clause (line 2). An interesting aspect of the SPARQL query is that users can remain devoid of any knowledge about the required transformations between EDAC and Lifemapper; they need only be concerned with *what* they want not, which in this case is a $?modelURL$. `CompletedExperiment` (line 8) denotes an experiment that has been processed by Lifemapper and therefore has a property `hasModelURL` that points to the URL of the resulting SDM.

```
1 SELECT ?modelURL
2 FROM <http://../experiment-1.owl>
3 where
4 {
5 ?experiment hasModel ?model.
6 ?model a Model.
7 ?model hasModelURL ?modelURL.
8 ?experiment a CompletedExperiment.
9 }
```

The experimental constraints referenced by the SPARQL query are specified in the OWL ontology specified in the FROM clause (experiment-1.owl)[1] and specify what kind of data should be bound to what specific layer (1 through 10). For example, a user can use the properties of the EDAC data ontology[2], shown in Figure 5, to specify that data for some environmental layer should be sourced from the MODIS sensor, reside within the `Western United States` and recorded within some `Duration1`. Below is an example of a rule that defines the relevant characteristics of some data to occupy the first element of a WCSCoverage Sequence. The rules is specified in OWL and encoded using the Manchester Syntax (Horridge et al. 2006). The paper will continue to use Manchester Syntax for any rule encodings.

```
Class: WCSCoverage1
  EquivalentTo:
    (hasDuration some Duration1)
```

[1]http://ontology.cybershare.utep.edu/ELSEWeb/experiments/experiment-1.owl

[2]http://ontology.cybershare.utep.edu/ELSEWeb/edac.owl

```
    and (hasRegion some WesternUnitedStates)
    and (hasSource value MODIS)
```

Note that the `WCSCoverage1` class definition is defined in terms of a duration (i.e., `Duration1`) and a region (i.e., `WesternUnitedStates`) that effectively specify the time and space requirements for the data layer. `Duration1` is defined in terms of a `hasStartDate` and `hasEndDate` (not shown in this paper), while `WesternUnitedStates` is defined in terms of a bounding box by using necessary and sufficient restrictions through an equivalent OWL class definition below.

```
Class: WesternUnitedStates
  EquivalentTo:
  (hasLeftLon some double[>= -130.0])
    and (hasLowerLat some double[>= 20.0])
    and (hasRightLon some double[<= -100.0])
    and (hasUpperLat some double[<= 50.0])
    SubClassOf:
      UnitedStatesRegion
```

The SHARE client will search through the ELSEWeb SADI registry and identify any EDAC services that match these constraints. SHARE will infer that any data offered by EDAC services that matches the selection criteria to be of type `WCSCoverageForLayer1` and therefore considered as an element of some `WCSCoverage Sequence`. The process is identical for specifying the other nine layers.

3.3 Service Orchestration

After SHARE identifies the relevant data to bind to each layer and composes a `WCSCoverageSequence`, the set of identified layers will be structured into a payload sequence. Focusing on the two SADI services presented earlier, it will be illustrated how the output of the payload extractor service can be used to create the `FullySpecifiedExperiment` that is required by the Lifemapper SADI service.

Consider the following rules (1) and (2) specified in our ontology that describes our service intputs and outputs:

```
Rule1
Class: UnderspecifiedExperiment
  EquivalentTo:
    hasWCSCoverageSet some WCSCoverageSequence

Rule2
Class: FullySpecifiedExperiment
  EquivalentTo:
    hasWCSCoverageSet some WCSPayloadSequence
    SubClassOf:
      UnderspecifiedExperiment
```

Rule 1 states that an individual that has a `WCSCoverageSequence` is considered to be of type `UnderspecifiedExperiment`, meaning it is not ready for ingestion by Lifemapper. Rule 2 states that any individual that has a `WCSCoveragePayload` is ready for Lifemapper processing and classified as a `FullySpecifiedExperiment`. Assume that in the ELSEWeb knowledge base there exists an individual of type `UnderspecifiedExperiment` (generated by other SADI service not discussed here for simplicity). Considering our example SPARQL query that requests for a `CompletedExperiment`, SHARE must determine what sequence of services to execute in order to generate such an individual. SHARE knows about:

1. The SADI service input/output descriptions for PayloadExtractor and Lifemapper
2. The Service ontology that contains Rule 1 and Rule2
3. The target `CompletedExperiment`, specified by the SPARQL query

Based on these resources, SHARE can determine that it must execute the Lifemapper service to obtain as an output a `CompletedExperiment` individual. However, at this stage of the process, SHARE is only aware of an `UnderspecifiedExperiment` individual. Based on Rule 2, it can determine that the difference between an `UnderspecifedExperiment` and a `FullySpecifiedExperiment` is the range of the `hasLayers` property. In order to infer an `FullySpecifiedExperiment`, SHARE must invoke the WCSPayload Extractor service, since it transforms WCSCoverageSequences to WCSCoveragePayloads. Once the WCSPayload Extractor has executed, the returned WCSPayload Sequence RDF will trigger RULE 2 and reclassify the `UnderspecifiedExperiment` as a `FullySpecifiedExperiment` and thusly pass this experiment individual to the Lifemapper SADI service.

3.4 Provenance in ELSEWeb

Provenance is a trace of all of the data sources and analytical methods that were used in a scientific analysis or model. Provenance is of particular relevance here to trace the automated process carried out. ELSEWeb provenance enables users to visualize, query, and understand data sources at EDAC and Lifemapper; analytical methods used at EDAC to generate derived products; and parameters and analytical methods used within a Lifemapper computational experiment. ELSEWeb services extended the original SADI design to include provenance information about the service that generated the data and the parameters used (if any) using the PROV data model (Moreau and Missier 2012). ELSEWeb SADI services use the PROV Ontology (PROV-O) (Lebo et al. 2013), which is a W3C recommendation to represent the PROV Data model using OWL. ELSEWeb provenance, presented in Figure 6, promotes transparency and confidence in model outputs derived. Transparency is in particular importance in automated systems such as ELSEWeb where processes are not defined *a priori*.

4 Related Work

The environmental Model Web is founded on principles which dictate approachable methods for interaction with minimal barriers (Geller and Melton 2008). Therefore, ELSEWeb can be considered to implement a subset of the Model Web, in particular the Data-to-Model relationship, because users can quickly run experiments without any regard for how the data is sourced, aggregated, transformed

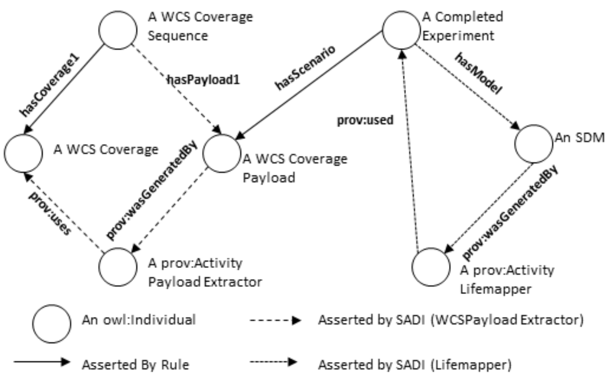

Figure 6: Prov Trace from Lifemapper

and served to models. A similar contribution was made by the authors of eHabitat (Dubois et al. 2011), where a Habitat Irreplaceability Index (HRI) service was published as an OGC Web Processing Service (WPS) and integrated with WCS data, similar to EDAC. The eHabitat environment however lacks the semantic capabilities of ELSEWeb and so adding non-OGC compliant data sources or models may require manual reconfiguration of the workflows rather than delegating the transformational process to agents such as SHARE. The iPlant Semantic Web Platform uses SSWAP (Simple Semantic Web Architecture and Protocol) for the discovery of services and the construction and execution of workflows (Gessler et al. 2013). iPlant uses OWL ontologies and Resource Description Graphs (RDGs) to describe services, their inputs and their outputs. A difference between SADI and iPlant is that SADI describes services in terms of their input and output only, without creating a hierarchy of types of services. While SADI automatically orchestrates and executes service pipelines using a description logics reasoner, iPlant allows users to create pipelines through a graphical interface and uses a reasoner to suggest the next service in the pipeline according to the service descriptions. Similar to SADI, iPlant contains a knowledge base of services and their semantic annotations for the discovery and invocation of services that are executed by service providers over the web.

5 Discussion

ELSEWeb leverages SADI best practices and lessons learned from other scientific communities (Vandervalk, McCarthy, and Wilkinson 2010; Vandervalk et al. 2013; Callahan et al. 2013; Riazanov et al. 2012) and enables scientist to employ the "If-Then-ELSE" mechanism in their research. Different combinations of source data and models can be more easily tested enabling a better characterization of uncertainty due to data and model selection. These more complex computational experiments are highly desirable. In the EDAC/Lifemapper testbed being able to easily explore the impact of selecting different input datasets and SDM algorithms on projected species distribution enables scientists to assess where the outputs agree and disagree. This can enable decisions about where future data collection efforts or investigations could be focused to resolve these discrepancies. It can enable decisions about where field monitoring investigations should occur in order to most quickly identify change on the ground when it occurs. And it can enable better decisions by environmental managers regarding how to best adapt to projected changes. More broadly, ELSEWeb is a generic approach that can lead in the future to more easily conducting integrated modeling investigations across domains. For instance, output data from Lifemapper could be automatically transformed to become input data for an investigation of how changing climate and plant species distributions may impact water availability in the future, through changes in evapotranspiration and surface runoff. A key element in reusing and repurposing EDAC data and Lifemapper is the use of a family of metadata standards to describe their services. Service metadata is programatically created and can be automatically discovered and processed. ELSEWeb has been constructed in a generic way that allows for extensibility to other data and models as long as data and models are provided as services and defined following a standard-compliant format and language.

SADI is agnostic to any particular ontology, as long as the RDF and OWL languages are used, reducing the knowledge negotiation process to an ontology mapping problem.

6 Future Work

Graphical Interface. SPARQL can be seen as a workflow schema that will be used by the SADI framework to execute the services needed to obtain and transform data and models required by our user. We are in the process of creating a graphical interface that enables scientist to use the SHARE client without having to learn SPARQL or OWL to specify the experimental constraint considering two key variables for ELSEWeb: location and time.

Expanding ELSEWeb services. The vision of ELSEWeb includes the addition of services that provide additional data and models relevant to biodiversity forecasting. Ontology mapping will be used to align service descriptions and metadata. As a first step towards expanding ELSEWeb knowledgebase, services provided by the GEOSS registry (http://geossregistries.info/) will be analyzed for their addition through the SADI framework.

Visualization. ELSEWeb provides the possibility of combinatoric explosions of scientific results, automatically generated by orchestrations of data integrations and service composition yielding many different SDMs. An equally automated mechanism for helping users analyze the wealth of scientific products is needed and therefore we intend delegate the orchestration of visualization services to a reasoning engine tailored for pipeline composition (Del Rio and da Silva 2012), similarly to how transformation orchestration is automated by SHARE.

Efforts similar to ELSEWeb to enable the automated publishing, discovery, and orchestration of services providing data and models will enable scientist to reuse data and test hypothesis on-the-fly using the "If-Then-ELSE" mechanism.

Acknowledgements ELSEWeb is 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 would like to thank Soren Scott and Bill Hudspeth from EDAC for their discussions about service metadata and the SADI development team for their steadfast support through their mailing list.

References

Barnosky, A. D.; Matzke, N.; Tomiya, S.; Wogan, G. O.; Swartz, B.; Quental, T. B.; Marshall, C.; McGuire, J. L.; Lindsey, E. L.; Maguire, K. C.; et al. 2011. Has the earth/'s sixth mass extinction already arrived? *Nature* 471(7336):51–57.

Busby, J. 1991. Bioclim-a bioclimate analysis and prediction system. *Plant Protection Quarterly* 6.

Callahan, A.; Cruz-Toledo, J.; Dumontier, M.; et al. 2013. Ontology-based querying with bio2rdfs linked open data. *Journal of Biomedical Semantics* 4(Suppl 1):S1.

Committee, F. G. D., et al. 1998. Fgdc-std-001-1998. *Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, DC*.

Del Rio, N., and da Silva, P. P. 2012. Capturing and using knowledge about the use of visualization toolkits. In *2012 AAAI Fall Symposium Series*.

Dubois, G.; Skøien, J.; De Jesus, J.; Peedell, S.; Hartley, A.; Nativi, S.; Santoro, M.; and Geller, G. 2011. ehabitat:" a contribution to the model web for habitat assessments and ecological forecasting. In *Proceedings of the 34th International Symposium on Remote Sensing of Environment. April*, 10–15.

Eaton, B.; Gregory, J.; Drach, B.; Taylor, K.; Hankin, S.; Caron, J.; Signell, R.; Bentley, P.; Rappa, G.; Höck, H.; et al. 2003. Netcdf climate and forecast (cf) metadata conventions.

Franklin, J. 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Geller, G. N., and Melton, F. 2008. Looking forward: Applying an ecological model web to assess impacts of climate change. *Biodiversity* 9(3-4):79–83.

Gessler, D. D.; Bulka, B.; Sirin, E.; Vasquez-Gross, H.; Yu, J.; and Wegrzyn, J. 2013. iplant sswap (simple semantic web architecture and protocol) enables semantic pipelines for biodiversity. 979.

Hannah, L. 2003. Regional biodiversity impact assessments for climate change: A guide for protected area managers. *A Users Manual for Building Resistance and Resilience to Climate Change in Natural Systems* 235.

Horridge, M.; Drummond, N.; Goodwin, J.; Rector, A.; Stevens, R.; and Wang, H. H. 2006. The manchester owl syntax. *OWL: Experiences and Directions* 10–11.

Lebo, T.; Sahoo, S.; McGuinness, D.; Belhajjame, K.; Cheney, J.; Corsar, D.; Garijo, D.; Soiland-Reyes, S.; Zednik, S.; and Zhao, J. 2013. Prov-o: The prov ontology. *W3C Recommendation, http://www. w3. org/TR/prov-o/(accessed 30 Apr 2013)*.

McGuinness, D. L.; Van Harmelen, F.; et al. 2004. Owl web ontology language overview. *W3C recommendation* 10(2004-03):10.

Moreau, L., and Missier, P. 2012. Prov-dm: The prov data model. *World Wide Web Consortium, Fourth Public Working Draft*.

Nativi, S.; Mazzetti, P.; and Geller, G. N. 2012. Environmental model access and interoperability: The geo model web initiative. *Environmental Modelling & Software*.

PrudHommeaux, E.; Seaborne, A.; et al. 2008. Sparql query language for rdf. *W3C recommendation* 15.

Riazanov, A.; Hindle, M. M.; Goudreau, E. S.; Martyniuk, C. J.; and Baker, C. J. 2012. Ecotoxicology data federation with sadi semantic web services. In *CEUR Workshop Proceedings*, volume 952.

Vandervalk, B.; McCarthy, E. L.; Cruz-Toledo, J.; Klein, A.; Baker, C. J.; Dumontier, M.; and Wilkinson, M. D. 2013. The sadi personal health lens: A web browser-based system for identifying personally relevant drug interactions. *JMIR Research Protocols* 2(1).

Vandervalk, B.; McCarthy, L.; and Wilkinson, M. 2010. Share & the semantic web-this time its personal! In *Proceedings of OWLED*.

Wilkinson, M. D.; McCarthy, L.; Vandervalk, B.; Withers, D.; Kawas, E.; and Samadian, S. 2010. Sadi, share, and the in silico scientific method. *BMC bioinformatics* 11(Suppl 12):S7.

Wilkinson, M. D.; Vandervalk, B.; McCarthy, L.; et al. 2011. The semantic automated discovery and integration (sadi) web service design-pattern, api and reference implementation. *Journal of biomedical semantics* 2(1):8.

Withers, D.; Kawas, E.; McCarthy, L.; Vandervalk, B.; and Wilkinson, M. 2010. Semantically-guided workflow construction in taverna: the sadi and biomoby plug-ins. In *Leveraging Applications of Formal Methods, Verification, and Validation*. Springer. 301–312.