eHabitat, a multi-purpose Web Processing Service for ecological modeling

G. Dubois[a], M. Schulz[a], J. Skøien[a], L. Bastin[c], S. Peedell[a]

[a] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via Fermi 2749, Ispra, 21027(VA), Italy – (gregoire.dubois, michael.schulz, jon.skoien, stephen.peedell)@jrc.ec.europa.eu
[c] School of Engineering and Applied Science, Aston University, Birmingham, B4 7ET, UK – l.bastin@aston.ac.uk

Abstract – The number of interoperable research infrastructures has increased significantly with the growing awareness of the efforts made by the Global Earth Observation System of Systems (GEOSS). One of the societal benefit areas that is benefiting most from GEOSS is biodiversity, given the costs of monitoring the environment and managing complex information, from space observations to species records including their genetic characteristics. But GEOSS goes beyond simple data sharing to encourage the publishing and combination of models, an approach which can ease the handling of complex multi-disciplinary questions. It is the purpose of this paper to illustrate these concepts by presenting eHabitat, a basic Web Processing Service (WPS) for computing the likelihood of finding ecosystems with equal properties to those specified by a user. Despite the availability of the agreed WPS standard for Web-based geospatial modeling, few practical implementations exist, making eHabitat a significant addition to the field. On the other hand, the wide uptake of Web access standards for geospatial data has led to a wealth of data sources within GEOSS which can be effectively combined using eHabitat. When chained with other services providing data on climate change, eHabitat can be used for ecological forecasting and becomes a useful tool for decision-makers assessing different strategies when selecting new areas to protect. eHabitat can use virtually any kind of thematic data that can be considered as useful when defining ecosystems and their future persistence under different climatic or development scenarios. The paper will present the architecture and illustrate the concepts through case studies which forecast the impact of climate change on protected areas or on the ecological niche of an African bird.

Keywords: ecological modeling, multi-disciplinary interoperability, SOA, Web processing services, Model Web, eHabitat.

1. Introduction

The use of distributed computing technology is revolutionizing the way we deal with information, and international initiatives, such as the Group on Earth Observations (GEO), are encouraging different communities to make their systems and applications interoperable. Biodiversity is one of the societal benefit areas that is likely to benefit most from this initiative because of the nature of the datasets required for environmental monitoring and strategy evaluation; they are huge in their spatio-temporal scope and dimensionality, while at the same time they are often documented and managed in a very fragmented and inconsistent manner.

When we consider ecological modeling, better results have traditionally been achieved either by improving existing models or by developing new ones. Chaining interoperable model components is now a third alternative that is particularly interesting because such a chain can potentially answer more questions than the individual models alone, allowing users to address complex questions in a variety of different contexts. Still, setting up a computing infrastructure where models can be easily plugged and played remains a challenge (Service, 2011).The "Model Web" proposed by Geller and Turner (2007) envisages such an environment of interacting models and encourages the practical development of a distributed, multidisciplinary network of independent, interoperating models and datastores communicating with each other using Web services. Beyond the simple sharing of information, the Model Web conceives increasing access to models and their outputs, and aims to facilitate greater model-model interaction, resulting in a web of interacting models, databases, and websites (Nativi et al., 2012).

A first effort in this direction has been described by Best et al. (2007) with OBIS-SEAMAP[1], the Ocean Biogeographic Information System - Spatial Ecological Analysis of Megavertebrate Populations, a spatially referenced online database, aggregating marine mammal, seabird and sea turtle observation data from across the globe. Another milestone was the setting up by Nativi et al. (2009) of an ecological niche modeling framework built around OpenModeller[2] (Muñoz et al., 2011), a popular tool for ecological niche modeling. The proposed modeling framework successfully employed a Service Oriented Architecture (SOA) even though at the time OpenModeller was still a stand-alone application, by making the modeling kernel accessible through external interfaces like SOAP.

Another example of an interoperable biodiversity information system where models are chained, is the Digital Observatory for Protected Areas (DOPA)[3] that is currently being developed at the Joint Research Centre of the European Commission in collaboration with other international organizations, including the Global Biodiversity Information Facility (GBIF), the UNEP-World Conservation Monitoring Centre (WCMC), Birdlife International and the Royal Society for the Protection of Birds (RSPB). DOPA is conceived as a set of distributed databases combined with open, interoperable Web services to provide end-users, from park managers to scientists and decision-makers, with the means to assess the state of protected areas at the global scale (Dubois et al., 2010). DOPA needs to easily exchange *information* with a number of reference spatial data infrastructures (SDIs) in order to compute the indicators involved in the assessments, but it must also rely on automated *services* for monitoring purposes. Ultimately, when used in conjunction

[1] http://seamap.env.duke.edu/
[2] http://openmodeller.sourceforge.net/
[3] http://dopa.jrc.ec.europa.eu/

with other environmental services which can supply information on phenomena such as predicted climate change, DOPA should be flexible enough to allow ecological forecasting and consideration of alternative future scenarios. This last objective has been partly achieved through the development of eHabitat[4], DOPA's core modeling service that is made available to the community by means of a Web Processing Service (WPS). It is the purpose of this paper to present eHabitat and discuss its use in an environment of interoperable data and model services.

The largest potential benefit from the Model Web is likely to be the practical and easy re-use of basic modeling components for different purposes. We believe that the granularity of the models expected to interact with each other is a critical factor in any operational version of the Model Web. A higher granularity is likely to generate more reusable elementary services, greater control for the users composing those services and thus, ultimately, more complex and useful modeling chains. Being a relatively simple modeling service for ecologists, eHabitat can be chained with other services and the reusability of its results is assured by wrapping the statistical modeling with the standardized OGC (Open Geospatial Consortium) WPS interface.

Version 1.0 of the WPS standard for exposing algorithms on the Web was agreed and published by the Open Geospatial Consortium (OGC) in 2007 (Schut, 2007). Five years on, take-up has been relatively slow and has been largely confined to a small number of academic and research institutions who publish discoverable models and algorithms using the standard (Lopez-Pellicer et al, 2012). The Web Coverage Service (WCS) and Web Feature Service (WFS) standards for raster and vector geospatial data (published in 2003 and 2005 respectively) have so far led to a far larger pool of interoperable data sources, though at present many of these are consumed for cartographic, rather than analytic or modeling purposes. The models exposed by the eHabitat WPS therefore represent a significant addition to the available suite of Web-based models which can be discovered and used to compose scientific workflows consuming data from many existing distributed sources across the scientific disciplines. Web-based clients developed for the service are publicly accessible, to allow straightforward interactive use and parameterization of the underlying models. Alternatively, the service can be called automatically as part of a workflow, and such experimental chaining of eHabitat with other Web-based models such as conservation planning algorithms and climate simulators is planned and ongoing.

In the following section, the reader will find an introduction to the use of Mahalanobis distances for modeling habitats before we describe in section 3 how we expose the model as a Web Processing Service (eHabitat). Two case studies are then proposed in section 4 where we illustrate the use of eHabitat when assessing climate change impact on a protected area, the UNESCO site of Tassili n'Ajjer, and on the ecological niche of an African bird, the Black Harrier (*Circus maurus*). A discussion will follow in section 5 on the use of eHabitat when chained with other web based modeling services before the general conclusions of section 6.

2. A short introduction to similarity modeling using Mahalanobis distances

The main idea behind eHabitat is to provide a service allowing end-users to find areas that have similar ecological properties to a reference location. This approach is typically used for ecological niche modeling, in which a spatial prediction model for a given species is computed from a set of environmental parameters, or 'indicators' (see e.g. Clark et al., 1993, Knick and

[4] http://ehabitat.jrc.ec.europa.eu/

Dyer, 1997, Rotenberry et al., 2006). In this context, Geographic Information Systems (GIS) have proven to be very useful tools for conservation because of the ease of handling various thematic layers and using multi-criteria decision trees for extracting information. A very common format for thematic data such as temperature is the raster grid, consisting of discrete pixels, each with a measured or modeled value. The computations in eHabitat are performed using a set of such raster data. To compute similarity to a reference location for each pixel of the domain under study, one popular approach is based on the Mahalanobis distance (Mahalanobis, 1936). The method is mathematically simple and fairly easy to understand, performs relatively well compared with most other models (Tsoar et al., 2007) and is computationally fast compared with more complex methods such as MaxEnt (2006). This method has therefore been used in the examples below, although other methods can easily be added within our WPS setup.

Numerically, the covariances and the variances of the (ecological) variables at a set of reference pixels define how much the vector of variables at a pixel *i* can deviate from the average within these reference pixels and still have a high similarity. For a pixel *i* the Mahalanobis distance *D* is defined as:

$$D_i^2 = (\mathbf{X}_i - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{X}_i - \mathbf{m}) \quad [1]$$

where $\mathbf{X}_i$ is the vector of indicators from this pixel, $\mathbf{m}$ the vector of the mean values and $\mathbf{C}^{-1}$ the inverse covariance matrix of the indicator variables at the pixels of interest. The use of the inverse of the covariance matrix makes the Mahalanobis distance independent of the different scales and units of the measurements. Because of the use of the inverse covariance matrix, highly correlated indicators will have less individual effect on D_i than uncorrelated indicators which could be considered to be more salient in the characterization of the region of interest. When the indicators used to generate the mean vector and covariance matrix are normally distributed, then D_i is distributed approximately according to a χ^2 distribution with n-1 degrees of freedom, and so we can convert D_i into probability values (p-values) for each pixel, ranging from 0.0 representing no similarity to 1.0 for areas which are identical to the mean of the PA. This p-value can be seen as the probability that a pixel outside the investigated area has a similar set of indicators to those found in the selected area. If the indicators are not normally distributed, the conversion is still useful as it rescales the unbounded *D* values to a 0.0 to 1.0 range. Generally we cannot assume normality of the data without further testing, and therefore in the following explanation we will interpret this p-value as a metric of similarity between that pixel and the indicators in the reference area.

Figure 1 illustrates the use of Mahalanobis distances for identifying similar ecosystems based on 9 thematic maps for the Kafue national park in Zambia. The symbol $\mathbf{m}$ in Eq. 1 above refers to the mean values of the maps within the park boundaries, whereas $\mathbf{C}$ refers to the covariance of the same values. We can then compute the similarity between the Kafue national park and the surroundings, shown in Figure 1.

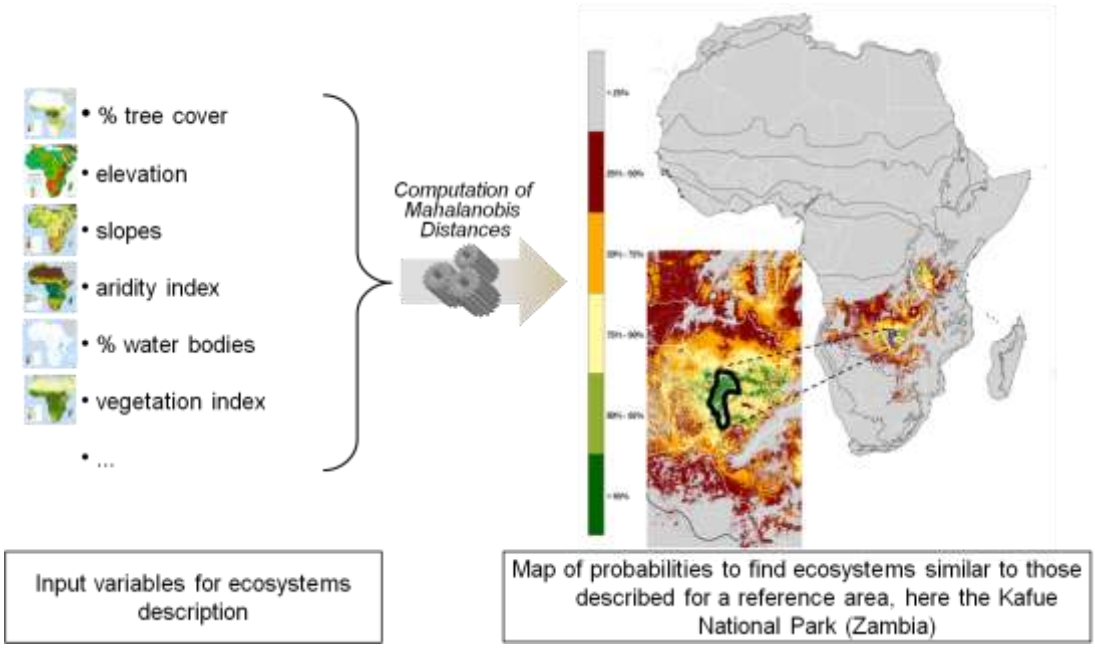

Figure 1: Use of Mahalanobis distances to compute probabilities of finding areas that are ecologically similar to a reference area, here a protected area in Zambia.

3. eHabitat as a Web Processing Service (WPS) for multi-purpose modeling

In a Model Web context, basic Web services exposing generic models can offer greater flexibility than complex modeling services. Because of the simplicity of the WPS, eHabitat can be reused for a wide variety of purposes; end-users can simply select their area of reference and identify their own data sets as input variables which characterize the phenomenon of interest. In the examples shown below, habitats defined by biophysical layers are considered, but there is no theoretical limitation to the number or type of variables that can be used for computing the similarity. The potential applications of the Mahalanobis distance model are therefore practically unlimited, provided that appropriate input data are available. There is a broad range of interdisciplinary possibilities, ranging from socio-economic modeling and ecological forecasting to the optimization of environmental monitoring networks. By the same token, the simpler the service, the easier it is to chain it with other services. We have therefore implemented the eHabitat WPS as a single and flexible service that is able to handle different types of input and perform different tasks depending on the requests. Figure 2 illustrates the main components of eHabitat which are further detailed in the next sections. The WCS provide input data which are processed by a statistical model (Mahalanobis distances) written in R and made accessible with a library (PyWPS) written in Python. Boundaries of the analyzed area can be either defined by the end-users or derived from a database of polygons representing protected areas.

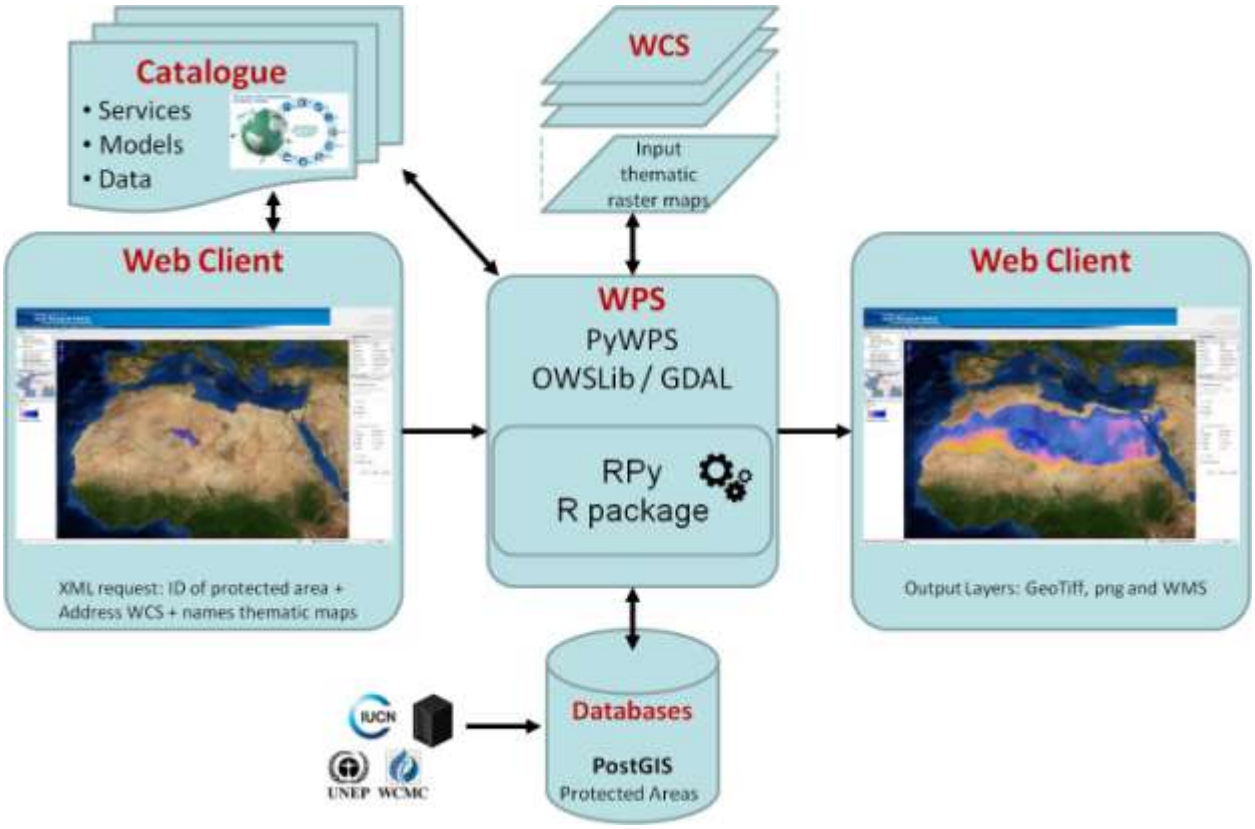

Figure 2. Design of the architecture of eHabitat 2.0.

3.1 Architecture of eHabitat WPS

The first version of eHabitat (eHabitat 1.0) was designed as a proof of concept to compute, for a given protected area, the probabilities to find elsewhere similar habitats using only three predefined thematic maps. However, the need to monitor ecosystems outside of protected areas, whether terrestrial or marine, is stronger than ever, if only to assess connectivity between protected areas and the external pressures caused by competition for land and water. Providing the scientific community with the means to compute habitat similarities anywhere on the globe, using their own thematic ingredients, is therefore an interesting option. The current version of eHabitat (eHabitat 2.0) therefore allows for an arbitrary number of input indicators along with the definition of an area of interest, which serves as a bounding box constraint for further processing (GEO AIP3, 2011). Technically, PyWPS (Cepicky and Becchi, 2007) was chosen as the WPS implementation. It is a lightweight Python based server that easily integrates with the Apache Web server, e.g. using the Common Gateway interface (CGI). The WPS serves the habitat modeling as one process which is also implemented using Python. The process expects several mandatory and some optional parameters (see Table 1). Using the Python-bindings for GDAL (Geospatial Data Abstraction Layer)[5] and OWSlib[6] the process can ingest and output a variety of different geospatial data formats (e.g. GeoTIFF, netCDF) as well as different OGC specifications, like WFS, WCS or Catalog Service for the Web (CSW). The adoption of

[5] http://www.gdal.org
[6] http://owslib.sourceforge.net/

interoperable standards for data access and process execution is actually a key prerequisite for model chaining.

Name	Card.	Description	Type[b]	Format
Mandatory parameters				
indicators	3..*	Multiple WCS/CSW URLs pointing to indicator coverages	ComplexData	Geotiff NetCDF
siteID[a]	0..1	WDPA site identification number, resulting in the reference geometry	LiteralData	Integer
sitePolygon[a]	0..1	Well-Known-Text representation of a user defined polygon	LiteralData	WKT
siteGeometryURL[a]	0..1	URL resulting in the reference geometry	ComplexData	WFS, KML, GeoJSON, GeoRSS, +
boundingbox	1	Bounding box defining the area of interest	BoundingBoxData	
Optional parameters				
forecast	0..1	Enable forecasting, default false (if true the forecasted indicators have to be provided as well)	LiteralData	Boolean
numRealisations	0..1	Number of realizations, to calculate uncertainty	LiteralData	Integer

[a] exactly one of these parameters has to be submitted
[b] these types refer to the InputFormChoice data structure (Table 2) as defined in Schut, 2007

Table 1. Mandatory and *optional* input parameters for the eHabitat process.

3.2 Model implementation

The computation of the Mahalanobis distances for the provided indicator datasets is done using the R statistical language (R Development Core Team, 2012) through an Rpy2[7] connection. This makes it easy to call the models from the python process and to take advantage of existing R implementations of methods such as the computation of the Mahalanobis distance from a mean vector and a covariance matrix, and the transformation from Mahalanobis distances to similarities through the χ^2 transform. The package has been written in a flexible way, so that the same function is used for computation of the current or forecasted similarities, and for different reference geometries such as polygons (protected areas) or points (classical niche modeling based on species occurrences) and with the possibility of weighting locations according to species density. It can also handle input data that present no spatial variability within the reference geometry as such a case will normally lead to a covariance matrix that cannot be converted) and categorical variables. This problem is often encountered with projected climate data which are usually computed on a low resolution grid. When used with high resolution data, no short scale spatial variability will be found and the computation of the covariance matrix will lead to numerical errors. The package is available on request, but we have not planned to upload it to the R package repository CRAN as it has been particularly developed for our purposes, and does not offer a substantial addition to other habitat modeling tools available under R.

[7] http://rpy.sourceforge.net/rpy2.html

3.3 Example operation of eHabitat WPS

The process is initiated by sending a WPS Execute request to the WPS server. This request describes all the required input parameters and desired outputs. Indicator datasets have to be referenced in the request using WCS *DescribeCoverage* or CSW *GetRecordById* URLs. The datasets are accessed using the provided area of interest and the default spatial resolution of the WCS layer with a *GetCoverage* request. All indicator layers that are requested must share similar geospatial properties (coordinate reference system (CRS), resolution). It is not in the scope of the habitat modeling to provide resampling or reprojection. The reference geometry from which the Mahalanobis distances will be computed (i.e. the boundary of a protected area or the point locations of species observations) is referenced using either a specific unique identifier for a park defined by the World Database on Protected Areas (WDPA), a WFS *GetFeature* URL or its Well-Known-Text (WKT) representation. These data are downloaded and transformed into R data structures before the computation of the Mahalanobis distances can be initiated with the package written in R.

Name	Description	Type[a]	Format
MahalDist	Raw similarity data as computed by the chosen method	Reference	Geotiff NetCDF
layerMahalDist	Reference to an OGC-WMS layer serving the result (GetMap-Request)	Reference	OGC-WMS
PNGoutput	Rendered image of the result with country borders background, legend and scale	Reference	PNG
[a] these types refer to the OutputData data structure (Table 60) and DataType data structure (Table 46) as defined in Schut, 2007.			

Table 2. Output parameters for the eHabitat process.

The results returned from the R code are processed to generate different output formats depending on the requirements of the end-users (see Table 2). If the user wants to perform some further analyses on the results, a GeoTiff, NetCDF (Network Common Data Format) or OGC WCS reference can be requested. Should the output be only for visualization purposes, PNG (Portable Network Graphics) images may be sufficient. For visualization in Web mapping clients, the user may request the output as an OGC-WMS reference.

4. Use cases

In the following, we will show some examples on how different web clients using eHabitat as back-end modeling service can be designed to answer different research questions.

4.1 Ecological forecasting in the Tassili n'Ajjer

In this subsection, we illustrate how eHabitat can be used for ecological forecasting by mapping the similarity of the climatic conditions for different time intervals to those found today in the Tassili n'Ajjer. The approach used here is following a time-based model (El-Geresy et al., 2002) where snapshots of the state of a specific location are captured and predictions made for different times.

The Tassili n'Ajjer is a UNESCO World Heritage site covering an area of 72,000 km^2 located in the Sahara, in the south-east of Algeria at the borders of Libya, Niger and Mali (Figure 3). The

modeling is done with the help of the eHabitat web client, which provides easy access to a set of current and forecast climate variables. The three variables are:

- the bio-temperature (the annual average of the temperature after values below freezing are set to zero);
- the average total annual precipitation;
- the ratio between the annual Potential EvapoTranspiration (PET) and the total annual precipitation.

These three variables are actually those used by Holdridge (1947) to define life zones, i.e. areas with matching biological characteristics. Depending on the relative values of the three variables, a site can be approximately classified within one of 38 defined classes (e.g. tropical rain forest, boreal desert, warm temperate dry forest, etc.). For the case illustrated here, we derived the three climatic variables from the WorldClim [8] database (Hijmans et al., 2005) which provides gridded maps of current and future climate variables at different lat-long resolutions, i.e., 10 minutes, 5 minutes, 2.5 minutes and 30 arc seconds. The dataset for the current climate is produced by interpolating the records from climate stations with a spline interpolation method. The forecast data have been produced by adding the changes from the large scale global circulation models to high resolution maps of the current climate (Ramirez et al., 2010), the results also being available from the WorldClim database. Note that the PET was obtained using the equation of Thornthwaite (1948); the equation is simple and frequently used when dealing with large scale computations.

Figure 3 shows a screen capture of a web client designed for eHabitat where the selected UNESCO site of Tassili n'Ajjer appears in a dark blue polygon while other protected areas are shown in a lighter blue. The right menu of the web client shows options of the model, i.e. the type of climate change model, the environmental scenario considered, the forecasted dates (today, 2020, 2050 or 2080) and the resolution of the outputs.

[8] http://www.worldclim.org

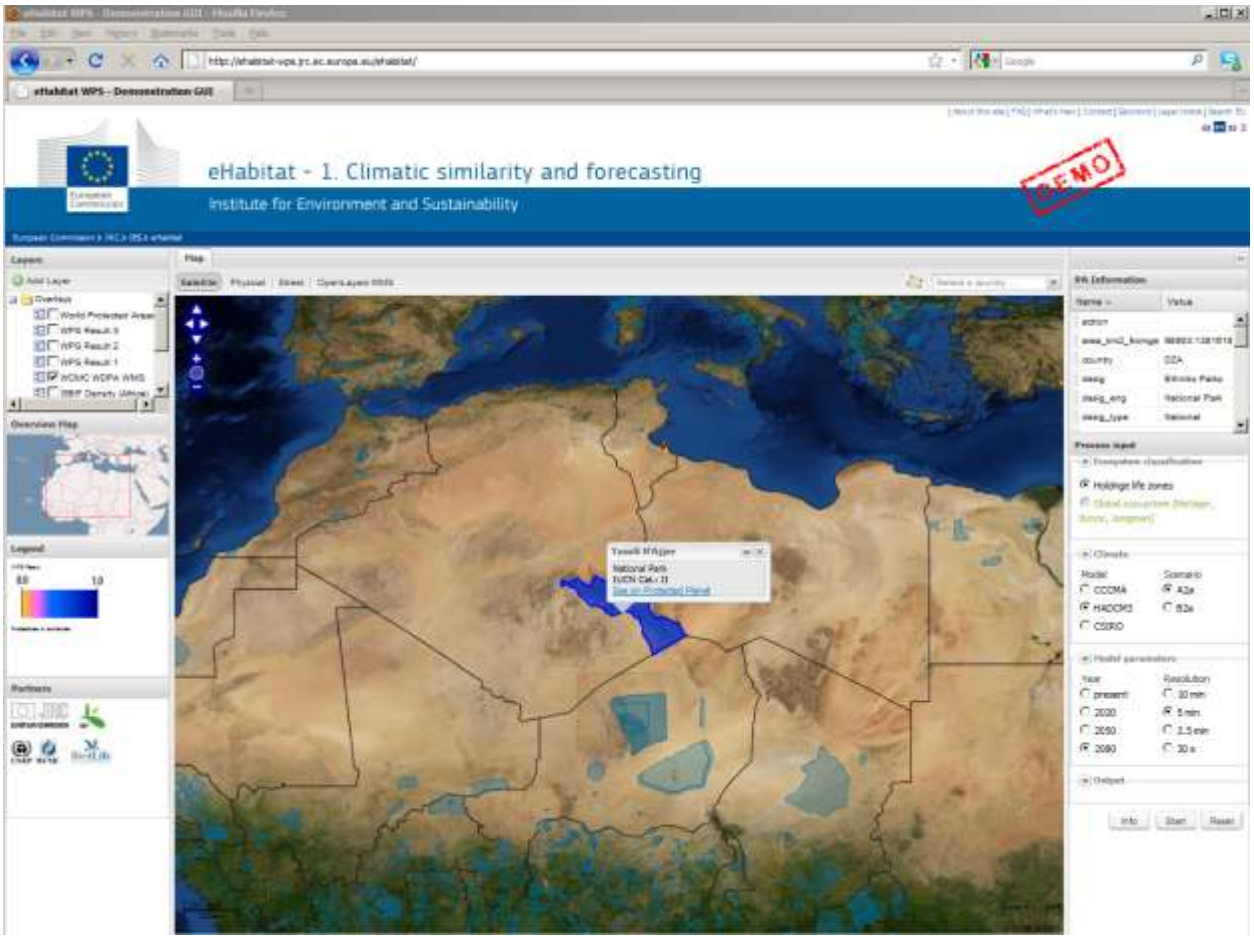

Figure 3: Screen capture of a web client designed for eHabitat showing the selected UNESCO site of Tassili n'Ajjer (dark blue polygon) and, in the right window, the modeling parameters to be selected by the end-user

By combining the park boundaries of Tassili n'Ajjer, as our reference area, with the three climatic input maps, one can compute the vector of means (**m**) in Eq. 1 of the climatic variables within the park boundaries for the current conditions (Figure 4) or for future dates (Figure 5). The covariance matrix (**C**) is computed from the same variables and $\mathbf{X}_i$ is defined here by the values of the climatic variables for a certain pixel for different time intervals.

Figure 4 shows the screen capture of the same web client shown in Figure 3 with the outcome of the modeling step using the bioclimatic conditions found today in the Tassili n'Ajjer. Blue colors show areas with high similarities with the average conditions found currently in the UNESCO site, while red and yellow colors show, respectively, medium and low similarities with current conditions. The results obtained for the forecasted cases depicted in Figure 5 are showing the probabilities to find similar conditions to today for the year 2050 (top) and 2080 (bottom). The upper screen shot shows that the area of Tassili n'Ajjer will have already lost almost all of its current properties in 2050 and similar conditions to today's situation will be found mainly North-East of the protected area. The situation depicted for 2080 is even more dramatic as the forecasted habitats will further shrink in surface and be further fragmented. An obvious word of caution is needed here as the example selected is to illustrate the concepts and one should be careful with any scientific interpretation of the results.

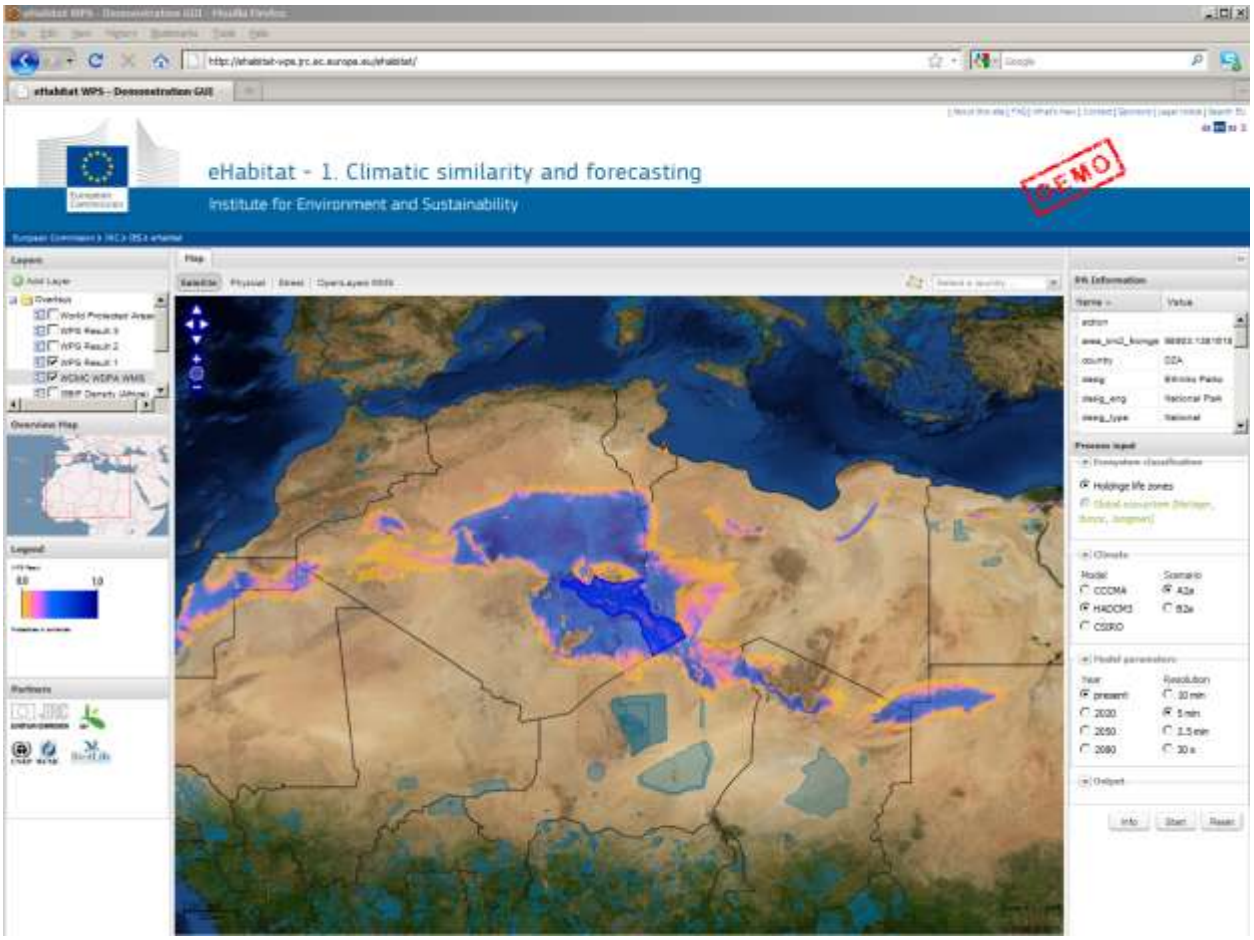

Figure 4: Screen capture of a web client designed for eHabitat showing the areas that are similar, from a climatic point of view, to the conditions found today in the Tassili n'Ajjer.

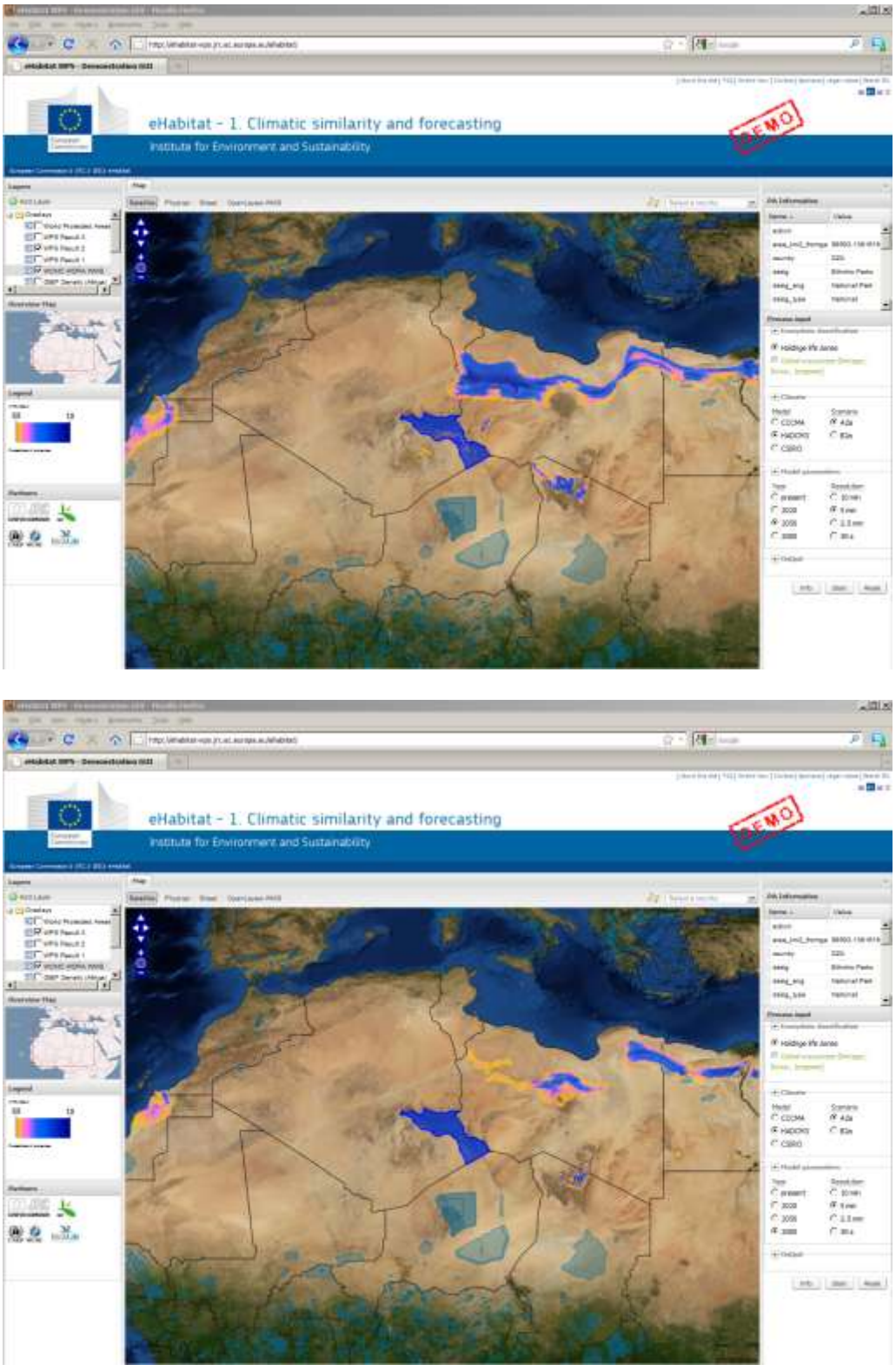

Figure 5: Screen captures of a web client designed for eHabitat showing the areas that are similar in 2050 (top) and 2080 (bottom), from a climatic point of view, to the conditions found today in the Tassili n'Ajjer.

4.2 Ecological forecasting of birds ranges

The example in section 4.1 used a polygon representing a protected area as the sampling support from which to derive the covariance matrix. However, one can equally well use a set of point data such as georeferenced species observations. In the following example, we used the eHabitat WPS in conjunction with a Web client similar to the one described in the previous section, but which has been enhanced with the option to query spatial occurences of a species. To obtain locations at which a specific species has been observed, we used services provided by the Global Biodiversity Information Facility (GBIF) which enables free access to biodiversity data via the Internet. As of May 2012, its Data Portal[9] provides unified access to over 320 million records from some 9 000 datasets supplied by hundreds of data publishers. Several REST-based Web services of GBIF provide means to construct and submit complex queries from machine to machine. In the case described here, the Occurrence service[10] is accessed via our web client to return records for a taxon occurring within a particular geographic bounding box. Output formats for these taxon records include the international KML (Keyhole Markup Language).

Figure 6 displays the enhanced web client and a use case summarizing the possible impact of climate change on an African bird, the Black Harrier (*Circus maurus*). Concentrated in the Western Cape (its core range) in South Africa, the total population is estimated to be around 1,000-1,500 individuals and the species is classified as vulnerable on the red list of endangered species. Computing Mahalanobis distances using the Holdridge data at the 96 locations where the GBIF reported the bird species, one gets a map of habitat similarity that can be interpreted as the theoretical climatic niche for this bird. Looking at the results from the forecasted Holdridge data for 2080 (Figure 6, bottom), one sees a dramatic loss of today's climatic niche, which becomes more restricted to coastal areas. This has particular conservation significance since coastal areas are usually under high pressure in the competitition for land.

5. Theoretical limitations of eHabitat

The modeling approach presented in the above use cases has a number of limitations. Ecologically, when monitored areas present a complex set of highly variable environments, such as a mountain near a lake, or a coastal area, computing Mahalanobis distances from such heterogeneous environments gives results which do not make much sense. There are a number of ways to circumvent these problems, for example by carefully stratifying the area into more homogeneous environments before launching a set of separate computations for each environment. While such a stratification step has not been implemented in the WPS, end-users are still getting means to detect with eHabitat such heterogeneous areas because the variability in the ecological parameters within the analyzed protected area is displayed. Figure 4 shows, for example, some variability in the bioclimatic conditions within the Tassili n'Ajjer UNESCO site. Environments which exhibit particularly low variability within the assessed region may also create some numerical challenges for the interpretation of the results. Nonetheless, these obstacles are intrinsic to the algorithmic implementation of eHabitat and do not jeopardize the broader idea of the modeling service as an elementary component designed to be used and re-used for a variety of use cases. It can also be seen that the approach can be easily extended to deal with 3D datasets, increasing the potential for multidisciplinary use. This step would allow

[9] http://www.gbif.org
[10] http://data.gbif.org/ws/rest/occurrence

marine experts, for example, to gain access to simple web based applications for modeling marine environments.

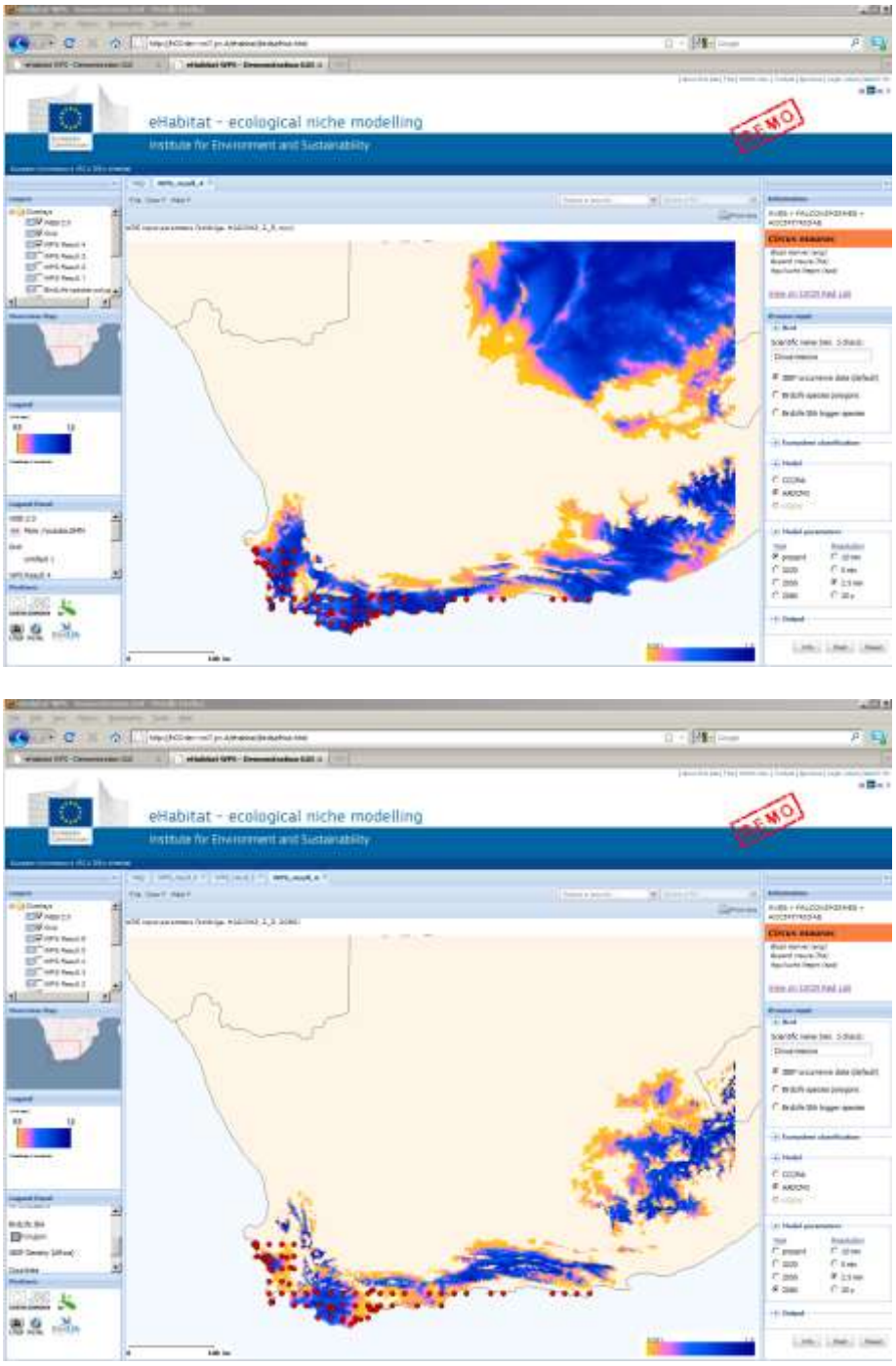

Figure 6: Screen captures of a web client designed for ecological niche modeling. Red dots, displaying the spatial distribution of the Black Harrier as reported by the GBIF, are overlaid on the output of the bioclimatic map of similarities derived from the observations. The upper figure shows the current conditions, the one below shows the probabilities of finding similar conditions in 2080.

6. eHabitat and the Model Web

The relative simplicity of the eHabitat WPS allows its reuse by other modeling services. This simplicity, however, increases the granularity of the fundamental elements in an environment based on the Model Web and, consequently, the difficulty of choosing the right components to construct a complex modeling workflow. If there is a proliferation of modular interoperable models and data services, then each must be very clearly documented so that a user discovering these resources can compare the available services and evaluate their fitness for the intended purpose. Such a framework within which end-users can select their own "ingredients" has been successfully prototyped in the context of the GEOSS AIP (Architecture Implementation Pilot) initiative and is briefly described in the next section.

6.1 Enhancing eHabitat WPS with a brokering approach

The Group on Earth Observations (or GEO) is coordinating international efforts to build a Global Earth Observation System of Systems (GEOSS) (GEO, 2009). The aim of GEOSS is to build a public infrastructure to *link together existing and planned observing systems around the world and support the development of new systems where gaps currently exist*[11]. The infrastructure that coordinates access to the systems, applications, models, and products is the GEOSS Common Infrastructure (GCI). To demonstrate the added-value of GEOSS and enhance the GCI functionalities, the GEO Architecture and Data Committee (ADC) launched the GEOSS AIP Initiative[12]. In December of 2010, the third phase of GEOSS AIP (AIP-3) was concluded: it developed scientific scenarios for several of the societal benefit areas recognized by GEOSS; cross-disciplinary pilots were also considered, including the Biodiversity and Climate Changes domain. Due to the multidisciplinary nature of this domain, the pilots required a multidisciplinary infrastructure to be set up. For GEOSS AIP-3 Biodiversity & Climate Change pilots, one of the objectives was to continue the successful experimentations developed by GEOSS in the framework of AIP-3 and the two previous AIP phases (Nativi et al, 2009). The EC-funded EuroGEOSS[13] (Pearlman et al., 2011) and GENESIS[14] projects developed a scenario in which the eHabitat model utilized a distributed discovery service (i.e. a Discovery Broker) to access Biodiversity and Climate Change datasets. In this way, end-users can select the 'ingredients' available on the Internet to model habitat similarities; an obvious enhancement to the existing modeling capacity of eHabitat.

The scenario architecture of eHabitat in AIP-3 included the following advanced components developed by the two EC-funded projects:

1) the EuroGEOSS Discovery & Access Broker services;

2) the EuroGEOSS/GENESIS Semantic Discovery Broker which extends the Discovery Broker by underpinning semantically-enabled queries;

3) WorldClim data served through an OGC WCS interface (developed by EuroGEOSS) to allow assessment of the impact of changing climatic variables in protected areas;

[11] http://www.earthobservations.org/geoss.shtml
[12] http://www.earthobservations.org/geoss_call_aip.shtml
[13] http://www.eurogeoss.eu/
[14] http://www.genesis-fp7.eu/

The “eHabitat” use scenario is fully detailed in the engineering report and accessible through the GEO Portal (GEO-AIP3, 2011).

The eHabitat scenario architecture benefits from the SOA brokering approach by implementing the "Catalogue" service through a Discovery Broker which is further coupled with another pair of effective components, the Access Broker and the Semantic Discovery Broker. The broker implements an extended version of the SOA where support for service composition and management, service orchestration and transaction are provided. This allows eHabitat to further interact with a plethora of heterogeneous services and data models characterizing multi-disciplinary scenarios. The broker also serves to lower the present GCI entry-barrier by providing users with a homogeneous discovery framework to heterogeneous resources (biodiversity, climate change, etc.) through the addition of “expert” brokering services which hide the heterogeneity of the underlying systems. This solution prevents the eHabitat user from having to “learn” and implement a diversity of information technologies which are sometimes immature and sparsely documented.

7. Conclusions and further considerations

Multi-disciplinary information integration is recognized by the scientific community as essential for the understanding of complex issues such as the response of biodiversity to global changes. This calls for the further development of flexible and scalable systems allowing integration with existing (and heterogeneous) services and data systems. The Digital Observatory for Protected Areas (DOPA), of which eHabitat is a component, is an example of such a platform where observations and models relating to trends in the world's ecosystems and species can be integrated. Relying on the dynamic model infrastructure envisioned in the Model Web, DOPA’s many benefits include improved means to discover, access, reuse and chain models and datasets for multiple purposes. The eHabitat WPS described here should illustrate these benefits: different web clients designed for different end-users and use-cases can be easily built on the top of a fundamental modeling service. The versatility of eHabitat allows it to be used within different contexts and workflows. At the same time, the benefits of being able to select from a large pool of fundamental modeling and data services, like the famous Lego blocks used to construct different toys, calls for well orchestrated and documented workflows and chains of analytical steps. These must apply international and disciplinary standards for achieving interoperability across different disciplinary systems and resources (i.e. data, services and models). The adoption of an extended SOA approach (i.e. Brokered SOA) realizes the necessary scalability and flexibility which should allow interoperability with a set of other services and data systems. If the adoption of standard Web services to publish eHabitat WPS should encourage its use by other communities, its reuse will largely depend on the development of new services allowing semantic interoperability.

Another downside of an environment based on numerous interacting model services is the potential use of a broad range of data types from uncontrolled sources. eHabitat in the Model Web would be exposed to many different types and levels of uncertainties and, when chained to other services, eHabitat itself becomes an additional component which further propagates uncertainties from a potentially long chain of model services. This integration of complex resources, such as data and models brings ever increasing challenges in dealing with uncertainty. For future developments, we are building on the lessons learnt from the UncertWeb

(www.uncertweb.org) project which promotes and develops tools and standards for quantifying and communicating uncertainty in a distributed, interoperable Model Web (Cornford et al., 2010, Bastin et al., 2012). eHabitat will adopt open source implementations of encoding standards, service interface profiles, discovery and chaining mechanisms developed in UncertWeb. Our first observations have been presented in Skøien et al. (2011a,b).

Acknowledgements

This work is partly supported by the European Commission, under the 7th Framework Programme, by the EuroGEOSS project funded by the DG RTD and by the UncertWEB project funded by the DG INFSO. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

We also acknowledge the comments of the reviewers who helped us to substantially improve the structure and contents of this paper.

More information about eHabitat and the DOPA can be found on the Internet, see http://dopa.jrc.ec.europa.eu/ and http://ehabitat.jrc.ec.europa.eu/, respectively.

References

Bastin, L., Cornford,D., Jones, R, Heuvelink, G.B., Pebesma, E., Stasch, C. Nativi, S. Mazetti, P. and Williams, M. Managing Uncertainty in Integrated Environmental Modelling Frameworks. *Environmental Modelling and Software*, 2012. (In Press)
http://dx.doi.org/10.1016/j.envsoft.2012.02.008

Best, B. D., P. N. Halpin, E. Fujioka, A. J. Read, S. S. Qian, L. J. Hazen, R. S. Schick (2007). Geospatial Web services within a scientific workflow: Predicting marine mammal habitats in a dynamic environment. *Ecological Informatics*. **2**(3): 210-223

Cepicky, J. and L. Becchi (2007). Geospatial processing via Internet on remote servers – PyWPS, *OSGeo Journal* 1 (May 2007): 5p

Clark, J. D., J. E. Dunn, and K. G. Smith (1993), A multivariate model of female black bear habitat use for a geographical information system. *Journal of Wildlife Management*, **57**: 519-526.

Cornford, D., Jones, R., Bastin, L., Williams, M., Pebesma, E. and Nativi, S. (2010) UncertWeb: chaining web services accounting for uncertainty. *Geophysical Research Abstracts*, Vol. 12, EGU2010-9052,

Dubois, G., M. Clerici, S. Peedell, P. Mayaux, J.-M. Grégoire and E. Bartholomé (2010), A Digital Observatory for Protected Areas - DOPA, a GEO-BON contribution to the monitoring of African biodiversity, In: "*Proceedings of Map Africa 2010*", 23-25 November 2010, Cape Town, South Africa.

El-Geresy, B. A., A. I. Abdelmoty and C. B. Jones (2002). Spatio-Temporal Geographic Information Systems: A Causal Perspective. In "*Proceedings of the 6th East European Conference on Advances in Databases and Information Systems* (ADBIS '02)", Y. Manolopoulos and P. Nàvrat (Eds.). Springer-Verlag, London, UK, UK, pp. 191-203.

Geller, G. N. and W. Turner (2007). The model Web: a concept for ecological forecasting, *Geoscience and Remote Sensing Symposium*, 2007. IGARSS 2007. IEEE International, 2469 – 2472, 23-28 July 2007.

GEO (2009). Group on Earth Observations, GEO 2009-2011 Work Plan. December 2009. Available at: http://www.geosec.org/documents/work%20plan/geo_wp0911_rev2_091210.pdf (accessed 3 October 2012)

GEOSS AIP-2 (2009). GEOSS AIP-2 Engineering Report "*The Impact of Climate Change on Pikas Regional Distribution*", Climate Change and Biodiversity WG Use Scenario, available at http://www.ogcnetwork.net/system/files/FINAL-pikas_AIP_SBA_ER.pdf (accessed 3 October 2012)

GEOSS AIP-3 (2011). GEOSS AIP-3 Engineering Report, "*eHabitat*", Climate Change and Biodiversity WG Use Scenario, available at http://www.ogcnetwork.net/pub/ogcnetwork/GEOSS/AIP3/documents/CCBio-eHabitat-ER-v2.0-FINAL.pdf (accessed 3 October 2012)

Hijmans, R., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas, *International Journal of Climatology*, **25:** 1965-1978.

Holdridge, L.R. (1947). Determination of world plant formations from simple climatic data. *Science*, **105**: 367--368.

ISO (2003). International Organization for Standardization. Geographic Information – Metadata. ISO 19115:2003(E)

Knick, S. T., and D. L. Dyer (1997). Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho, *Journal of Wildlife Management*, **61**(1): 75-85.

Lopez-Pellicer, F. J., W. Rentería-Agualimpia, R. Béjar, P. R. Muro-Medrano and F. Javier Zarazaga-Soria (2012). Availability of the OGC geoprocessing standard: March 2011 reality check. *Computers & Geosciences*, **47**:13-19

Mahalanobis, P. C. (1936). On the generalised distance in statistics, for the classification problem. *Proceedings of the National Institute of Sciences of India*, **2**(1): 49-55.

Muñoz, M.E.S., R. Giovanni, M.F. Siqueira, T. Sutton, P. Brewer, R.S. Pereira, D.A.L. Canhos and V.P. Canhos (2011) openModeller: a generic approach to species' potential distribution modelling. *GeoInformatica*, **15**: 111–135

Nativi, S. and L. Bigagli (2009). Discovery, Mediation, and Access Services for Earth Observation Data, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, **2**(4): 233-240.

Nativi, S., P. Mazzetti, Geller G. N. (2012). Environmental Model Access and Interoperability: the GEO Model Web Initiative. *Environmental Modelling & Software*. (In press) http://dx.doi.org/10.1016/j.envsoft.2012.03.007

Nativi, S., P. Mazzetti, H. Saarenmaa, J. Kerr, E. O. Tuama (2009). Biodiversity and climate change use scenarios framework for the GEOSS interoperability pilot process, *Ecological Informatics*, **4**: 23-33.

Phillips, S. J., R. P. Anderson, R. E. Schapire (2006). Maximum entropy modelling of species geographic distributions, *Ecological modelling*, **190**: 231-259.

Pearlman, J., Craglia, M., Bertrand, F., Nativi, S., Gaigalas, G., Dubois. G., Niemeyer, S., Fritz., S. (2010) EuroGEOSS: an interdisciplinary approach to research and applications for forestry, biodiversity and drought. In: "*Proceedings of the 34th International Symposium on Remote Sensing of Environment*", April 10-15, 2011, Sydney, Australia

R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ramirez, J. and A. Jarvis. (2010) Disaggregation of global circulation model outputs. International Center for Tropical Agriculture, CIAT, Cali, Colombia.

Rotenberry, J.T., K. L. Preston and S. T. Knick (2006). GIS-based niche modeling for mapping species' habitat, *Ecology*, **87**, 1458-64.

Santoro, M., Mazzetti, P., Nativi, N., Fugazza, C., Granell, C., Diaz, L., (2011). Methodologies for Augmented Discovery of Geospatial Resources, In "*Discovery of Geospatial Resources: Methodologies, Technologies and, Emergent Applications*", IGI Global, in press

Schut, P., (2007). OpenGIS Web Processing Service. OGC Document 05-007r7. URL: http://portal.opengeospatial.org/files/?artifact_id=24151 (accessed 03 October 2012).

Service, R.F. (2011). Coming soon to a lab near you: drag-and-drop virtual Worlds. *Science*, 331(6018):669-671Skøien, J., Truong P., G. Dubois, D. Cornford, G. Heuvelink, G. Geller, (2011). Uncertainty propagation in the Model Web: a case study with eHabitat. In: "*Proceedings of the 34th International Symposium on Remote Sensing of Environment*", April 10-15, 2011, Sydney, Australia

Skøien, J., M. Schulz, G. Dubois, R. Jones, G.B.M. Heuvelink, D. Cornford (2011). Uncertainty propagation in chained web based modeling services: the case of eHabitat. In: "Innovation in sharing environmental observations and information". Proceedings of EnviroInfo 2011, 25th International Conference Environmental Informatics". W. Pillmann, S. Schade and P. Smits (Eds), pp: 46-58, 5-7 October 2011, Ispra, Italy

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, **38**(1): 55–94.

Tsoar, A., O. Allouche, O. Steinitz, D. Rotem, R. Kadmon, (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, **13**: 397-405.