Contribution ID: **117** Type: **Presentations**

Cloud Computing for Ecological Modeling in the D4Science Infrastructure

Tuesday, 9 April 2013 11:00 (20 minutes)

Impact

Species distribution models aiming at estimating the presence of a species in a given area are essential instruments in the development of strategies and policies for the management and the sustainable and equitable use of living resources . These models rely on available occurrence records to investigate the relationships between observed species presence and the underlying environmental parameters that –either directly or indirectly –determine a species distribution in a known area and use this information to predict the probability of a species occurrence in other areas.
Despite their importance in the development of informed conservation and management strategies, the generation of such models is nowadays still limited to few specific cases. There are two main factors that prevent the vast majority of researchers from exploiting such approaches on a large scale. First, the generation of these models requires very often large computing capabilities and appropriate modeling tools. These are often not available in the research centers where scientists operate. Second, meaningful models can only be developed when both a sufficient amount of good quality occurrence point datasets and suitable environmental datasets are available. A lot has been done in the last years to collect such datasets. However, they are still scattered in many different heterogeneous databases. This makes very difficult and time consuming to use them on a large scale in an integrated way. The requirements behind the generation of predictivemodels are common to many other scientific areas where modern data-intensive science approaches are used . In these contexts data come in all scales and shapes and innovative technological solutions are continuously introduced that promote new ways of acquiring, storing, manipulating and transmitting vast data volumes, as well as stimulating new habits of communication and collaboration amongst scientists".
Among the new technological solutions a primary role is played by e-Infrastructures, and in particular by Hybrid Data Infrastructure (HDI) . These aim at supporting large-scale resource sharing –where resources range from hardware to data and software –and at providing scientists with resource-as-a-service. This last paradigm extends the notions of applications as services (a.k.a. "SaaS") and hardware and software systems as services (a.k.a. "IaaS"and "PaaS") as introduced in Cloud computing . HDIs offer a rich array of data and data management services by leveraging other infrastructures (including Cloud). Their goal is to enable a data-management capability delivery model where computing, storage, data and software are made available "as-a-Service"by the infrastructure. By means of these capacities, an HDI supports the creation and operation of virtual research environments , i.e., web-based cooperation environments equipped with all the resources needed to accomplish a scientific investigation. This Hybrid Based Infrastructure model implemented by D4Science , has been developed to serve the needs of biodiversity scientists. In particular this infrastructure offers the capabilities for facilitating species distribution modeling. These consist both of large scale data processing facilities, obtained through the use of services compliant to the cloud paradigms, and of services supporting users in accessing environmental data spread among distributed data providers.

Summary

Species distribution modeling is a process aiming at computationally predicting the distribution of species in geographic areas on the basis of environmental parameters including climate data. In order to further promote the diffusion of such an approach it is fundamental to develop a flexible, comprehensive, and robust environment enabling practitioners to produce species distribution models more efficiently. A promising way to build such an environment is offered by modern infrastructures promoting the sharing of resources, including hardware, software, data and services. We describe an approach to species distribution modeling based on the D4Science Infrastructure that can offer a rich array of data and data management services by leveraging other infrastructures (including Cloud), by discussing the services needed to support the phases of such a complex process.

URL

http://i-marine.eu/

Description

Species distribution modeling refers to the process of using computer algorithms to predict the distribution of species in geographic space on the basis of a mathematical representation of their known distribution in the environmental space. It is a complex and iterative process which include at least the following key steps : (i) identification of relevant data; (ii) modeling, i.e., deciding how to deal with the correlated prediction variables, selecting the appropriate algorithm, training the model, assessing the model; and (iii) mapping predictions to geographic space.
The D4Science Infrastructure, deployed and maintained by several EC projects, starting from D4Science, D4Science II, iMarine and EUBrazilOpenBio, can support such a complex process by offering both services oriented to simplify data discovery by users, when datasets are scattered among heterogeneous and distributed repositories, and services support the processing of such datasets for various purposes. In order to better illustrate the challenges involved in implementing such kind of workflows and the needs of the scientists in performing them, the talk first focus on the description of a typical and large-scale modeling scenario in the marine biology domain is given in the next section. This description is followed by a presentation of the basic facilities characterising the D4Science Infrastructure as a modern computing platform is discussed . In addition to the basic facilities, D4Science provides practitioners involved in species distribution modeling with (a) facilities specifically conceived to support environmental data discovery and access when data are federated from a number of data providers and (b) facilities supporting a cloud oriented computation of species distribution modeling. Finally, a detailed description of how such facilities have been exploited to support species distribution modeling phases and the resulting benefits are discussed.

Primary authors: Dr CASTELLI, Donatella (ISTI-CNR); Mr SINIBALDI, Fabio (ISTI-CNR); Dr CORO, Gianpaolo (ISTI-CNR); Dr CANDELA, Leonardo (ISTI-CNR); PAGANO, Pasquale (CNR - ISTI)

Presenter: PAGANO, Pasquale (CNR - ISTI)

Session Classification: Cloud Platforms

Track Classification: Community Platforms (Track Lead: P Solagna and M Drescher)