Building an e-Infrastructure for Capture Statistics and Species Distribution Modeling: the D4Science Approach

Anton Ellenbroek, *FAO of the UN, FIE, anton.ellenbroek@fao.org (Italy)*
Leonardo Candela, CNR, Italy, leonardo.candela@isti.cnr.it,
Donatella Castelli, CNR, Italy, donatella.castelli@isti.cnr.it,
Pasquale Pagano, CNR, Italy, pasquale.pagano@isti.cnr.it,
Marc Taconet, FAO, Italy, marc.taconet@fao.org

Data analysts and environmental scientists expect data to be available at ever shorter intervals and in ever greater detail. Their activities require collaborations across parties that are widely dispersed and autonomous. Moreover, these collaborations are often cross-discipline and require innovative research supporting environments that integrate data, processing and work-flows to produce new knowledge. They also increase the demand for interoperability, and some collaborative products and initiatives are already emerging, and are brought together in D4Science, e.g. :

- The Environmental community, where ESA provides GPOD interoperability, to share on demand geospatialdata such Sea Surface Temperature maps.
- The Biodiversity community, where AquaMaps enables biologists to create species prediction maps, integrating biological and environmental data.
- The Fishery management community, where extraction of information from statistical data sets combined with geospatial and environmental can improve catch estimates.

These collaborations imply an environment with the ability to load and archive various sources of data, to harmonize datasets by mapping correspondences, and to query across these datasets.

In the case presented here, Fisheries Resources Management has an ever increasing demand for capture data with ever finer geographical resolution, and an ever greater number of environmental variables to take into account. The management information process requires access to relevant documentation and source data. There is a need for innovative approaches that support the production of dynamic reports that integrate data from heterogeneous sources. These may be the result of complex aggregation, processing, analysis and editing of continuously evolving data. There also is the need to generate model-based, large-scale predictions of occurrence of aquatic species taking into account various environmental variables.

To support such demanding scenarios a promising approach is based on e-Infrastructures. By definition, an e-Infrastructure is a framework enabling secure, cost-effective and on-demand resource sharing across domain boundaries. A resource here can be “physical” (e.g. storage and computing resources) or “digital” (e.g. software, processes, data). It can be shared and interact with other resources to provide functions to its clients, that can be human or an application. Thus, an e-Infrastructure is the “mediator” in a market of resources and accommodates the needs of resource providers and consumers. An important feature of this mediator is the variety of solutions put in place to remove barriers related to the resources heterogeneity. thus facilitating their “smooth” consumption.
The infrastructure layer supports: (i) resource providers, in “selling” their resources; (ii) resource consumers, in “buying” and organizing resources to build their applications. Furthermore, it provides organizations with logistic and technical aids for application building, maintenance, and monitoring. A well-known example of such an e-Infrastructure is represented by the Grid, where a service-based paradigm is adopted to share and reuse low-level physical resources. Application-specific e-Infrastructures are in their turn inspired by the generic e-Infrastructure framework and bring this vision into specific application domains by enriching the infrastructural resource model with specific service resources, i.e. software units that deliver functionality or content by exploiting available physical resources.

The potentially unlimited availability of resources allows a new development paradigm based on the notion of Virtual Research Environment (VRE). A VRE is an integrated environment that provides seamless access to resources and offers facilities for communication, collaboration and interaction among scientists and researchers. This is built by aggregating the needed constituents after hiring them through the e-Infrastructure. The resulting research environments are organized 'views' built atop the pool of available assets, ranging from computers and servers to collections and services.

This presentation focuses on the implementation of this innovative approach in the context of D4Science (www.d4science.eu). D4Science is an EU funded project that started in 2008 (and has its roots in the DILIGENT project that started in 2004) aiming at developing a production-level e-Infrastructure capable of providing scientific communities (including the Fishery and Aquaculture Resources Management) with dedicated VREs to serve the needs of various challenging application scenarios. In October 2009, D4Science entered its second phase aiming at reinforcing the D4Science e-Infrastructure by adding interoperability facilities. These aim to develop data discovery, processing and sharing facilities bridging diverse autonomous e-Infrastructures. This will result ,in e-Infrastructure Ecosystems that potentially serve a significant set of communities dealing with multidisciplinary challenges whose solution is currently beyond reach.

The Establishment of an operational e-Infrastructure that supports VREs capable to satisfy the needs of the user communities calls for a paradigm shift in the way the communities operate. This involves at least three major areas: (i) the technology, (ii) the organizational model of the resources, and (iii) the human processes. These three areas are not independent, but any choice in one of them strongly influences and constrains the others. Progress in one area stimulates modifications and adaptations in the others, and also to start possible further changes in that area itself.

In this presentation the introduction of VREs is described starting from the technological perspective. Then the possibilities to collect, analyze, and organize data and the collaboration features are described.

From the technological point of view, the development of the software for e-Infrastructure-based VREs applications was complex and demanding in terms of effort and required resources. The applications are expected to provide on-demand, powerful and easy to use functionality for supporting scientific collaboration and in-silico experiments. The VREs now include services for storage, processing and presentation of potentially huge and heterogeneous datasets, but they also allow semi-automated transformations across different data schemas and formats. The entire array of standards and interoperability-oriented approaches underlying the D4Science technology will be presented discussed.

D4Science implements many standards and innovative technologies that serve the networked and distributed features. The majority of these are not directly perceived by end-users but they are distinguishing features of the the gCube system, i.e. the technology behind the D4Science infrastructure. The gCube technology was designed in accordance with second-generation Web Services standards including Web Services Resource Framework (WSRF), Web Services Addressing (WS-Addressing) and Web Services Security (WS-Security). For information management, the system is equipped with features to handle compound information objects consisting of multiple parts and alternative manifestations. These objects can have single or multiple metadata records in various schemas, e.g. Dublin Core, ISO 19115, AgMES.

The D4Science e-Infrastructure is in production mode since June 2008. Since then it has been populated with widely different resources, and the information resources published so far are quite heterogeneous. They range from multidisciplinary fisheries data sources, such as Fishery Country Profiles, National Aquaculture Legislation Overviews, Capture Time series graphs, species distribution maps, to very different Earth Observation products. The standards and interoperability-oriented approaches to guarantee a seamless consumption of these resources will be presented.
Four VREs have been created that serve the needs of different scenarios in the Environmental Monitoring and Fishery and Aquaculture Resources Management domains. These specific services will be presented to demonstrate the power and flexibility of the proposed approach.

D4Science-II will work toward the strengthening of the D4Science e-Infrastructure by making it capable to consume resources coming from other e-Infrastructures established in other areas as well as sharing part of its resources with them. The resulting e-Infrastructures ecosystem will be a very rich space of resources that can be easily consumed in a seamless way thanks to the D4Science e-Infrastructure services. The D4Science pool of standards and technologies will be enriched with state-of-the-art data exchange capabilities. These will include the Statistical Data and Metadata eXchange (SDMX), Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Open Archives Initiative Object Reuse and Exchange (OAI-ORE).

As a consequence of this, the Virtual Research Environments that bring the power of the constituent resources in data processing and storage to users in an easy to understand and intuitive format, with minimal requirements on bandwidth and processing power.

Finally, the sustainability of the approach will be discussed. D4Science-II will introduce the new notion of “knowledge ecosystem”. This ecosystem is a community of e-Infrastructures that provide services. In the knowledge ecosystem vision the ”health” of a data e-Infrastructure is influenced by the “health” of the others since each of them affects, and is affected by, the modification and updates on the services provided by other e-Infrastructures. This may open the way to the adoption of alternative sustainability solutions. For example, the fact that each ecosystem e-Infrastructure has its own community that uses and enhances the ecosystem capabilities might imply that all these communities could federate to provide the support to maintain the entire ecosystem operation.