Interoperability and Sharing of Biodiversity Data on a National Network in Italy

Corrado Iannucci
IPTSAT Srl
Rome, Italy
info@iptsat.it

Valter Sambucini
ISPRA Istituto Superiore per la Protezione e la Ricerca Ambientale (Italian environment agency)
Rome, Italy
valter.sambucini@isprambiente.it

Abstract—Biodiversity protection requires to access and to process many sources of information. The Italian Ministry of Environment has entrusted ISPRA (Italian Institute for Environmental Protection and Research) to manage a program aimed to support interoperability and harmonization of the information stored in different data bases. Some aspects of such program, including the National Network for Biodiversity (NNB), are highlighted and commented, with reference also to other relevant initiatives. NNB has already been implemented, relying also on relevant systems as EC-CHM and EUNIS. However, the current availability of interoperability standards (as for biodiversity related themes in the Annexes of INSPIRE) require to update NNB, also taking into account the global development of the ITC architecture of the public sector.

Keywords—clearinghouse mechanism; heterogeneous data bases; interoperability; biodiversity data and networks

I. INTRODUCTION

Collecting data, processing information and sharing knowledge are essential features of any program meant to ensure the sustainability of the environment by co-ordinated actions of citizens, business entities and public administrations. The availability of network infrastructures and protocols allows to implement the sensible approach based on leaving the data where they are produced and supporting the users to access them whenever they are. However, such approach requires a thoroughly planned effort in order to overcome the barriers that prevent or slow down the circulation of data.

As shown inter alia by the implementation of the INSPIRE Directive [1] (whose related initiatives are among the more important ones aiming to cross-link disconnected silos of data), such barriers are cultural, organizational and technological at the same time. On the other hand, supporting changes of cultural legacies, of organizational frameworks and of operational tools is the main mission of ICT. This is specifically true in the domain of environment-related activities, where huge progresses have been made since the Seventies of the past century: the availability of remotely sensed imagery, of telematic networks and of large data bases have supported new approaches of studying and managing the territorial processes.

Such new approaches include the distribution of digitized data and metadata about the collections of physical items of environmental interest, as those hosted in archives and museums. Similarly to what was happening in other analogous sectors (e.g. the libraries, storing paper documents as books or maps), the user community has felt it necessary to efficiently find information about such collections even if remotely located or also without the need of prior knowledge of such locations.

On such basis, many activities have been carried out at regional as well as national levels. In Italy, the relevant initiatives have to harmonize within the context set by the architectural choices made for the global ITC infrastructures.

II. THE ITALIAN PROGRAMME FOR BIODIVERSITY

ISPRA (the Italian Institute for Environmental Protection and Research) cooperates with the European Environment Agency (EEA), in its capacity of being the Italian national focal point of the EIONET network and also a member of both the European Topic Centre for Biodiversity Diversity (ETC/BD) and the European Topic Centre for Spatial information and Analysis (ETC/SIA).

The Institute (www.isprambiente.it) acts under the vigilance and policy guidance of the Italian Ministry for the Environment and the Protection of Land and Sea (Ministero dell’Ambiente e della Tutela del Territorio e del Mare - MATTM). ISPRA carries out (involving partners from academia, industry and public administration as necessary) its programs on the basis of national laws and of international obligations (e.g. the Copernicus program establishing a European capacity for Earth Observation).

ISPRA has been entrusted with the management of a program aimed to improve information access and sharing in the biodiversity community. Such information is stored in different heterogeneous data bases, therefore technical and organizational problems have to be dealt with. The related initiatives have been carried out with a strong involvement of the relevant stakeholders, as a means to reach their goals in an effective way. The technical solutions have been based upon applicable experiences in Italy, Europe and abroad.
A. The National Strategy for Biodiversity (NSB)

As an obligation derived from art. 6 of the Convention on Biological Diversity (www.cbd.int), MATTM has established the National Strategy for Biodiversity (NSB) [2], in cooperation with other stakeholders (ministries, regional authorities, citizens etc.) and taking into account the relevant EU Directives. Such Strategy (currently spanning from 2010 up to 2020) focuses on three main themes: biodiversity and ecosystem services, biodiversity and climate change, biodiversity and economic policies.

To implement the Strategy, three strategic objectives (all of them implying an effective information sharing) have been identified:

1. “By 2020, ensure the conservation of biodiversity, or the variety of living organisms, their genetic diversity and the ecological complexes of which they are part, and ensure the protection and restoration of ecosystem services in order to guarantee their key role for life on Earth and human well-being”;

2. “By 2020, substantially reduce the nationwide impact of climate change on biodiversity, by defining the appropriate measures to adapt to climate changes and mitigate their effects and increasing the resilience of natural and semi-natural ecosystems and habitats”;

3. “By 2020, integrate biodiversity conservation into economic and sectoral policies, also as potential for new employment opportunities and social development, while improving the understanding of the benefits from ecosystem services derived from biodiversity and the awareness of the costs of losing them”.

The implementation of NSB has been harmonized with the indications of the EU biodiversity strategy to 2020 [3].

B. The “Sistema Ambiente 2010” project

The preparatory studies of NSB pinpointed that the biodiversity information was scattered in a number of heterogeneous data base, making its exploitation difficult or even unreliable. Therefore, MATTM has financed the “Sistema Ambiente 2010” project, aimed to find a technical and organization solution to manage the biodiversity data in a way to meet the needs of NSB.

C. The National Network for Biodiversity (NNB)

In the framework of “Sistema Ambiente 2010”, ISPRA has implemented the National Network for Biodiversity (NNB) in 2012 [4]; afterwards, ISPRA is committed to maintain and to improve NNB, in order to extend the number of connected organizations in the biodiversity domain. NNB is meant to support the collection, validation and re-use of the thematic knowledge.

NNB, together with the dedicated “Portale NaturalItalia” web portal (www.naturaitalia.it), provides the services of the Clearinghouse Mechanism (CHM) for Biodiversity, fulfilling the role of Italian node in the system of the CHMs required by the CBD.

From the organizational point of view, a specific technical and scientific committee (CTS) has been entrusted to assess the quality (in terms of accuracy, relevance, completeness and reliability) of the information distributed by the network and to assist all the network entities in reaching and maintaining the necessary quality levels. The CTS members are appointed on the basis of their professional skills and are mainly selected from the academia.

The NNB entities pertain to three classes of users:

1. Contributors: this class includes network members providing data on a ad hoc basis;

2. Focal points (FP): public bodies, local administrations, museums and research centers that provide data and moreover are able to implement network nodes using their HW/SW infrastructure, in order to support the generic Contributors;

3. Centers of Excellence (CdE): university institutes and research laboratories, with a high level of specific expertise; they are committed to actively manage the scientific content of the network, both cooperating in the assessment of data provided by the network entities and carrying out large programs addressing the data collection and classification.

FPs and CdEs are recognized institutions at national and international levels and provide reliability to the whole network. Contributors, FPs and CdEs are the actual owners of their data and are in charge of their administration. Specifically, NNB in itself support the access and re-use of information, overcoming the heterogeneity of the data sources.

III. TECHNICAL ISSUES

The global requirements for the implementing NNB, as briefly described in the above, include the features of distribution, heterogeneity and autonomy. Such features, as defined in a seminal paper [5], imply ITC solutions based on a federated architecture.

Such ITC solutions support the sharing of information in a collections of loosely coupled data sources, on the basis of an agreed set of rules. Each data source stores its data according to its own specific model (not necessarily structured in a database, as shown in [6]; the set of rules allows to map each specific model to a agreed model, common to all the components of the federation i.e. to all the systems hosting the different data sources.

When a federated system needs to get data from the other systems, it launches its query with reference to the common data model; suitable software tools (a.k.a mediators or brokers) translate such query to the specific models of every data sources as mapped to the common data model; other software tools (a.k.a wrappers) provide the mapping of the specific
models to the common model. Obviously, the more the specific model are similar (i.e. are harmonized to the common model), the less transformation effort is needed.

A more advanced approach extends the set of agreed rules to include the exchange of messages among the federated systems, in order to get more generic services, other than data retrieval [7]. Accordingly, the systems achieve various levels of interoperability, defined by [1] as “the possibility for … data sets to be combined, and for services to interact, without repetitive manual intervention, in such a way that the result is coherent and the added value of the data sets and services is enhanced”.

Actually, federated systems have been widely exploited in bioinformatics. Generic requirements and possible solutions are provided by [8] and [9]. The federation of data sources and the availability of web services is compared in [10]; accessing non-relational data sources on the web is discussed in [6].

Specific requirements for biodiversity data, with reference to the taxonomic frameworks, have been exposed by [11]. Biodiversity data requirements are linked to the wider domain of bioinformatics; the use of organism names as a universal metadata element connecting biodiversity data from separate data sources is pinpointed in [12].

Arguing about the adoption of taxonomic names as identifiers in federated systems, [13] reviews possible alternate candidates and suggests to rely upon Digital Object Identifiers (DOIs) and Life Science Identifiers (LSIDs); such themes are revised and extended in a more recent review paper [14]. The challenges of the data formats of different sources (e.g. natural history collections, survey reports, scientific literature) to the discovery of global biodiversity patterns and processes (i.e. within a “big data” approach) are discussed in [15].

NNB has been implemented with reference to such context. Specifically, the NNB design has initially been based on the architecture of the biodiversity clearinghouse (EC-CHM) implemented by the European Environment Agency at the end of past century, as an important experience of promoting and facilitating technical cooperation, knowledge sharing and information exchange amongst national CHMs of the EU member states. In order to duly ensure coherence with other systems of MATTM, some components of the technical infrastructure have been modified. Afterwards, BISE - Biodiversity Information System for Europe (biodiversity.europa.eu/chm-network) has been taken into account as the evolution of EC-CHM itself. The general approach has been to provide a similar set of functionalities and, as far as possible, a similar user interface even when the technical tools have been modified (as in the case previously mentioned).

As a federated system, NNB relies on:

- a central node supporting the queries on the accessible data sets;
- common protocols for describing and accessing the shared data.

The central node, on top of the usual functionalities (from user authentication to query routing and load balancing), has been completed by a replica of the EEA’s EUNIS database (eunis.eea.europa.eu). Such database (built at the beginning of this century) holds non-redundant, integrated information about species, habitats and sites of interest for EU; therefore, it provides a reliable context for the data provided by the entities participating to NNB.

The software infrastructure has been derived mainly by the relevant experiences of GBIF - Global Biodiversity Information Facility (www.gbif.org/). Specifically:

- data are mapped to ABCD (Access to Biological Collection Data) schema;
- BioCASE (Biological Collection Access Service) Provider Software is deployed as a middleware in order to ensure an abstraction layer between the single data source and the network.

The core component of BioCASE Provider Software is PyWrapper that provides a an XML/CGI interface towards an increasing list of data sets with different structures; a client application therefore exchanges messages with this PyWrapper interface (not directly with the targeted data source), on the basis of the rules made available by the BioCASE Protocol.

It should be noted that PyWrapper has been designed to deal only with SQL-compliant data bases; however, the last version, available since January 2015, is able to map data also from Excel spreadsheets (in order to publish data sources of limited extent, usually provided as spreadsheets). Some other similar extensions should be made available in the future, granting the due generality and flexibility to PyWrapper.

BioCASE Protocol can be seen as a (not back-compatible) evolution of the previous DiGIR Protocol; a new TDWG Access Protocol for Information Retrieval (TAPIR) is meant to reconcile both of them. Moreover, also the ABCD schema is evolving: e.g. a version denominated ABCDEFG (ABCD Extended for Geosciences) has been produced, in order to include data also from palaeontological, mineralogical and geological digitalized collection data [16]. The ABCDEFG schema is exploited by GeoCASE (Geosciences Collection Access Service), an extended version of BioCASE.

In NNB, the user interface is provided by a GeoCASE web portal, fully supporting also BioCASE. Moreover, ABCDDNA (a theme specific extension for ABCD) is being adopted to support storage and exchange of data related to DNA collection units.

NNB coordinators can rely upon a dedicated tool, the Monitor Service Biocase [17]. Such tool records the global level of service provided by NNB in terms of active users, volumes of exchanged data, list of involved data sources etc. Moreover, the tool supports the validation and the quality check of the logical mapping of a data source vs. ABCD or ABCDEFG schema.
IV. NNB AND THE PUBLIC ITC INFRASTRUCTURE

NNB has been deployed as an additional system to the already existent ICT infrastructure of the public administration in Italy. Therefore, its operation has to take into account opportunities and constraints deriving by the relationships with the other systems and relevant legal rules.

A. INSPIRE Directive

The three Annexes the INSPIRE Directive [1] list many themes of potential interest for NNB, mainly: “protected sites”, “land cover”, “biogeographical regions”, “environmental monitoring facilities”, “habitats and biotopes”, “land use”, “species distribution”. The relevant Implementing Rules and Technical Guidelines, currently available on inspire.ec.europa.eu/index.cfm/pageid/2, have been mostly released after the start of NNB. This can imply a modification of the federated schema in order to express its content taking into account the INSPIRE data specifications. In general, biodiversity occurrence data do not directly fall into a specific INSPIRE theme; however, the related information relies also to some INSPIRE data themes. At least, the wrapper should take care of reconciling the data sources schemas and the federated one. Moreover, if the biodiversity data sets are be described in terms of INSPIRE metadata, their discovery and reuse are facilitated.

It appears to be useful to mention that, in principle, INSPIRE rules have a well delimited horizon (briefly: spatial data sets pertaining to the public sector; data of interest for environment management; data sets of new creation or thoroughly reshaped). However, in practice such horizon includes most entities and data that are managed in a biodiversity-related network as NNB. Therefore, the assessment of compatibility with INSPIRE should not be ignored.

It should also be noted that the interoperability-oriented approach (on which INSPIRE is based) is going to be exported to many other sectors, overcoming the apparent limits of the spatial data sets of interest for the environment. The European Interoperability Framework [18] has been aimed to facilitate the interoperability of services and systems between public administrations, and also between administrations and the citizens and industry. Moreover, the Digital Agenda for Europe [19] states that interoperability is essential to maximizing the social and economic potential of ICT in EU. Currently, the program ISA² [20] has been launched in order to exploit interoperability for modernizing the whole public sector. Such wealth of initiatives will apparently impact on networks as NNB.

B. Data openness and reuse

NNB makes its content reusable under the IODL – Italian Open Data License v.2.0.; the license text is available at www.sinanet.ispambiente.it/it/italian-open-data-license-v2-0.

In Italy, data have to be released by the public administration as open data, according to indications of the Code of the Digital Administration [21]. Various licenses can be adopted [22]; among them, IODL allows any content reuse, also for commercial initiatives, as far as the source is acknowledged. When data are of environmental interest, the access is granted both to allow the involvement of the citizens in the environment protection and to facilitate new economic initiatives. The legal framework originates from the Aarhus Convention [23] and the PSI Directive [24] [25].

Such commitment to the openness and reuse of public sector data is supported by a common ITC architecture: SPC [26]. SPC connects the heterogeneous systems of the public administration making them interoperable through specialized logical gates. The SPC architecture supports also SINAnet, the national system similar to EIONET. NNB could benefits of both infrastructures, as far as its BioCASE can be hosted on the relevant logical gates.

V. CONCLUSIONS

NNB has started and is still supporting an interesting cooperative effort among the biodiversity community in Italy. The benefits are expected also outside this community, in terms of best practices that are available for improving data interoperability in the public sector.

An amount of about 1.5 million records are already shared. It is foreseen that this amount of available knowledge will progressively increase.

While already exploiting open reliable SW solutions, NNB will probably have to improve their integration in the context of the current ITC infrastructure of the public administration, specifically in terms of data specifications.

REFERENCES


