Ecological Informatics 26 (2015) 162–172

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

An infrastructure-oriented approach for supporting biodiversity research

Leonardo Candela *, Donatella Castelli, Gianpaolo Coro, Lucio Lelii, Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi, 1, 56124 Pisa, Italy

ARTICLE INFO

Article history:
Received 2 January 2014
Received in revised form 18 July 2014
Accepted 21 July 2014
Available online 1 August 2014

Keywords:
Data integration
Data sharing
Data processing
Species occurrence data

ABSTRACT

During the last years, considerable progresses have been made in developing on-line species occurrence databases. These are crucial in environmental and agricultural challenges, e.g., they are a basic element in the generation of species distribution models. Unfortunately, their exploitation is still difficult and time consuming for many scientists. No database currently exists that can claim to host, and make available in a seamless way, all the species occurrence data needed by the ecology scientific community. Occurrence data are scattered among several databases and information systems. It is not easy to retrieve records from them, because of differences in the adopted protocols, formats and granularity. Once collected, datasets have to be selected, homogenised and pre-processed before being ready-to-use in scientific analysis and modelling. This paper introduces a set of facilities offered by the D4Science Data Infrastructure to support these phases of the scientific process. It also exemplifies how they contribute to reduce the time spent in data quality assessment and curation thus improving the overall performance of the scientific investigation.

1. Introduction

Data sharing in the research domain is a practice whose benefits are nowadays well understood by both data *owners* and data *consumers* (Boulton et al., 2012; Gray et al., 2002; Hey et al., 2009). Its adoption makes available to scientists a considerable amount of data that they can exploit in conducting their research. Sharing not only empowers them to access datasets produced and collected by colleagues working in the same domain, but it also enables the exploitation of very different data made available in other domains. This new data availability, especially the cross-domain one, is opening the way to new types of scientific practices, e.g., experiments, analysis, modelling, that were not possible few years ago. It also strongly facilitates the multi-disciplinary collaborations that are needed to address today's large research challenges. The attempts to exploit data in contexts different from where data has been produced have recently highlighted that an effective data reuse is often too challenging for the individual scientists (Borgman, 2011). Individual datasets are accessible with different protocols and through different user interfaces. This situation requires that a considerable amount of scientists' time is spent in understanding how to access the datasets, selecting the most appropriate ones, homogenising them and, more in general, preparing the datasets that fit the purpose of the planned scientific investigation. This lack is pushing researchers and technologists in computer science to think about new approaches for data sharing and management practices. These approaches must be flexible and powerful enough to adapt to the multitude of different and evolving situations, making the underlying complexity transparent to the scientists.

1.1. Data sharing and reuse in biodiversity research: state of the art

Data sharing and reuse is particularly relevant in modern biodiversity research to address large scale questions (Bendix et al., 2012; Costello, 2009; Enke et al., 2012; Michener and Jones, 2012). Large scale initiatives have been launched in the past years, either at the global – e.g., *GBIF* (Edwards et al., 2000), *OBIS* (Grassle, 2000), *VertNet* (Constable et al., 2010), *Catalogue of Life* (Jones et al., 2011) – or regional level – e.g., *speciesLink*[1] and *List of Species of the Brazilian Flora*[2] – to support the worldwide sharing of various collections of biodiversity data. The development of standards for data sharing has been promoted by establishing appropriate interest groups (Bach et al., 2012; Meng, 2004). Domain specific standards have been developed to focus on different interoperability aspects, e.g., *Darwin Core* (Wieczorek et al., 2012) and *ABCD* (TDWG, 2005) for data representation, *DiGIR* and *TAPIR* (TDWG, 2010) for distributed data discovery, and *LSIDs* (Clark et al., 2004) for data citation.

* Corresponding author.
E-mail addresses: Leonardo.Candela@isti.cnr.it (L. Candela), Donatella.Castelli@isti.cnr.it (D. Castelli), Gianpaolo.Coro@isti.cnr.it (G. Coro), Lucio.Lelii@isti.cnr.it (L. Lelii), Francesco.Mangiacrapa@isti.cnr.it (F. Mangiacrapa), Valentina.Marioli@isti.cnr.it (V. Marioli), Pasquale.Pagano@isti.cnr.it (P. Pagano).

[1] http://splink.cria.org.br/.
[2] http://floradobrasil.jbrj.gov.br/2012/.

http://dx.doi.org/10.1016/j.ecoinf.2014.07.006
1574-9541/

In spite of this large offer and initiatives, the biodiversity domain also suffers from the sharing and reuse problems highlighted above. Goddard et al. (2011) described and analysed them by reviewing the state of biodiversity data hosting and discussing the technological and social barriers affecting data sharing. Bach et al. (2012) analysed the technical solutions and standards implemented by existing information systems and repositories to support multidisciplinary biodiversity research. Well known initiatives aiming at simplifying biodiversity data access, like GBIF, are reacting to the need of simplifying biodiversity data access by carrying out strategic plans to further enhance the offering of "*seamless data access, integration, analysis, visualisation and use*" (Global Biodiversity Information Facility, 2011). There is a general awareness of the need to "seek a solution whereby these data are rescued, archived and made available to the biodiversity community" (Goddard et al., 2011). At the same time, it is clear that it is neither feasible nor reasonable to envisage a solution based on a single system in charge of maintaining and making available the entire production of biodiversity data. Rather it is expected that such a solution will be made available through an open endeavour in which (*a*) initiatives building databases for such data will continue to exist, (*b*) existing key players will continue to evolve towards larger federations, aiming at bringing the data out of these databases and promoting their sharing and reuse (e.g., GBIF and Catalogue of Life), and (*c*) increasingly more automatic support to the access and exploitation of shared data will be offered through new infrastructures working side-by-side with the rest — e.g., Pangea (Diepenbroek et al., 2002), DataONE (Michener et al., 2012) and Map of Life (Jetz et al., 2012).

1.2. Paper contribution

This paper introduces one of these new infrastructures, namely D4Science (Candela et al., 2009; D4Science.org, 2012). In particular, the paper describes the facilities D4Science offers to support access and reuse of species occurrence data. D4Science provides scientists with an integrated and flexible computer-assisted environment, built on top of existing databases and information systems. It offers facilities for supporting two key phases of the reuse practice, i.e., *data acquisition* and *data preparation*. By "data acquisition" it is meant the action of discovering, selecting and accessing relevant data in diverse and disperse databases in a seamless way. By "data preparation" it is meant the action that precedes the actual reuse of the data, i.e., distilling and amalgamating discovered data as needed for "fitting the purpose" of the research activity. D4Science offers these facilities "*as-a-Service*",[3] i.e., community of practices can start using these facilities like off the shelf instruments without incurring in technology development and deployment efforts. The given facilities are developed by following an approach that supplements (while not supplants) databases and information systems mandates and arrangements for dataset collection and aggregation. Thus D4Science contributes to the implementation of the global biodiversity open endeavour envisaged by many (Goddard et al., 2011; Peterson et al., 2010; Roberts and Moritz, 2011).

2. Methods

As already discussed in the Introduction, data about species occurrences are now scattered among several databases and information systems. There is no single service that gives access to the entire spectrum of this kind of data across the boundaries of disciplines, themes, regions, and taxonomies. A number of initiatives (e.g., GBIF) aggregate a large amount of data from different databases and publish integrated versions of them through a single uniform interface. In order to implement such services they ask database providers to adhere to established publication guidelines, formats and protocols. Moreover, during the aggregation phase they apply specific transformations in order to generate the required unified view. Usually, these transformations are not only limited to the syntactic format. They often implement harmonisation and quality enhancement practices that are decided by the service provider and are not explicitly made known to the data consumers.

D4Science is a data e-Infrastructure which supports a different approach. It is built and operated by a dedicated software system: gCube (Candela et al., 2008). It offers a rich array of resources including datasets and data management facilities by leveraging on existing information systems and other data infrastructures. Further, it supports the creation and operation of *virtual research environments* (Candela et al., 2010, 2013b), i.e., virtual spaces where a group of scientists, remotely distributed, have access to the resources (data, tools and computing capabilities) needed to perform their specific works. D4Science makes its facilities available "as-a-Service" by two provision models: (*a*) a human-oriented model, i.e., the facilities are offered via a number of portlets via the D4Science portal, and (*b*) a service-provider-oriented model, i.e., the facilities are offered via a number of web based protocols and APIs.

Among its facilities D4Science offers (*i*) a seamless access to third-party repositories and information systems and (*ii*) an open set of functionalities for data transformations and quality improvement. In the rest of this paper we will describe these functionalities and highlight how they can be exploited in the scientific practices.

2.1. Occurrence data acquisition facilities

Differently from the other solutions provided so far in the biodiversity domain, D4Science does not impose any specific guideline or protocol/format to the databases or information systems it aggregates. Rather, it is conceived to deal with the heterogeneity and challenges resulting from a scenario where the providers are neither expected to be collaborative nor to modify their strategies for data publication. Moreover, D4Science does not build an aggregated database. Rather, it realises data aggregation dynamically, at query time.

D4Science offers a service for species occurrence data discovery and access named *Species Products Discovery* (SPD). In addition to species occurrence data, the service supports discovery and access to nomenclature data (Taxonomic items). However, the features associated with this type of information are out of the scope of this paper, they are discussed in Amaral et al. (2014).

SPD is conceived as a sort of mediator service (Wiederhold, 1992) over a number of databases. In order to give access to species occurrence data, the SPD service has been equipped with plug-ins interfacing with three major information systems: GBIF, OBIS, and speciesLink. To enlarge the number of information systems and data sources integrated into SPD, it is sufficient to implement (or reuse) a plug-in. A plug-in is able to interact with an information system or a database by relying on a standard protocol, e.g., TAPIR, or by interfacing with its proprietary protocol. Every plug-in mediates queries and results from the language and model envisaged by SPD to the peculiarities of a single database. In particular, every mediator relies on mappings (Lenzerini, 2002) supporting (*i*) the rewriting of queries from the unifying SPD query language to the query language supported by the specific data provider, and (*ii*) the transformation of results from the specific data provider format to the unifying SPD format. Details on the SPD query language, the SPD unifying data format and the mapping of retrieved data into the unifying format are extensively discussed by Candela et al. (2014). It is important to highlight that records, once described in the unified data model, contain details on their provenance produced accordingly to the citation policies promoted by each database. The effort needed to implement a new mediator depends on the complexity of the mappings between the data source query language and results format to the SPD ones. However, the definition of such mappings is quite easy because

[3] The term "as-a-Service" has been introduced in the context of the Cloud technologies (Foster et al., 2008). It refers to both a business model and a delivery model. These are based on the notion of "service", where a customer pays the provider on a consumption basis for such a "service".