This is a pre-print of the article: L. Candela, D. Castelli, G. Coro, L. Lelii, F. Mangiacrapa, V. Marioli, P. Pagano An infrastructure-oriented approach for supporting biodiversity research, Ecological Informatics, Volume 26, Part 2, March 2015, Pages 162-172, ISSN 1574-9541, http://dx.doi.org/10.1016/j.ecoinf.2014.07.006

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

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, homogenized and pre-processed before being ready-to-use in scientific analysis and modeling. 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.

Keywords: Data integration, Data sharing, Data processing

1. Introduction

Data sharing in the research domain is a practice whose benefits are nowadays well understood by both data *owners* and data *consumers* (Gray et al., 2002; Hey et al., 2009; Boulton et al., 2012). Its adoption makes available to scientists a considerable amount of data that they can exploit in conducting their research. Sharing empowers them not only to access datasets produced

Email address: {name.surname}@isti.cnr.it (Leonardo Candela, Donatella Castelli, Gianpaolo Coro, Lucio Lelii,
Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano)

and collected by colleagues working in the same domain, 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, modeling, that were not possible few years ago. It also strongly facilitates the multi-disciplinary collaborations that are needed to address today 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, in selecting the most appropriate ones, homogenizing 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.

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 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 (Meng, 2004; Bach et al., 2012). 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, *LSIDs* (Clark et al., 2004) for data citation.

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*

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

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).

Paper contribution. This paper introduces one of these new infrastructures, namely D4Science (D4Science.org, 2012; Candela et al., 2009). 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 datasets collection and aggregation. Thus D4Science contributes to the implementation of the global biodiversity open endeavour envisaged by many (Goddard et al., 2011; Roberts and Moritz, 2011; Peterson et al., 2010).

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

[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".

the boundaries of disciplines, themes, regions, and taxonomies. A number of initiatives (e.g., GBIF) aggregate 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 to the databases 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, 2013), i.e., virtual spaces where 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 of the similarities in objectives and data between the data provider and SPD. In our experience in building tens of mediators, the average time to develop a real plug-in is two working days.[4]

Occurrence data discovery mechanism is based on a very simple procedure that allows a user to specify either the scientific name or a common name of the target species. The goal is to favour the *recall*, i.e., to maximise the amount of datasets discovered by means of a query. In practice, the system runs a query on all the databases and collects the results in a unified result set. To overcome the potential issues related with taxonomy heterogeneities, the service offers a mechanism enabling query expansion, i.e., the user query can be augmented with synonyms and "similar" species names from other data sources. For instance, it is possible to define queries by specifying that the scientific name to search is "*Sarda Sarda*" and the system might also use the synonyms names as reported in OBIS. This implies that every plug-in is called to report the results for all the datasets corresponding to the list of species specified in the target database. In addition to species names users can specifically select the databases to search among. They can also specify constraints on the spatial and temporal coverage of the data to which they are interested.

The service is offered through a web-based user interface (Fig. 1) consisting of (*i*) a search panel (on the top) for specifying the information need, (*ii*) a results view panel (on the right) for browsing the list of datasets resulting from a query, and (*iii*) a classification panel (on the left) offering clustered views of the results, e.g., the classification, the data provider, the database, and the rank. The occurrence data discovered are presented to the user in an homogenised

[4]Examples of developed plug-ins are available through the gCube webpage hosted by the Ohloh service `http://www.ohloh.net/p/gCube`. In particular, by using the Browse Code a user will have access to the entire gCube software code. SPD and its plug-ins are in the data-access area.

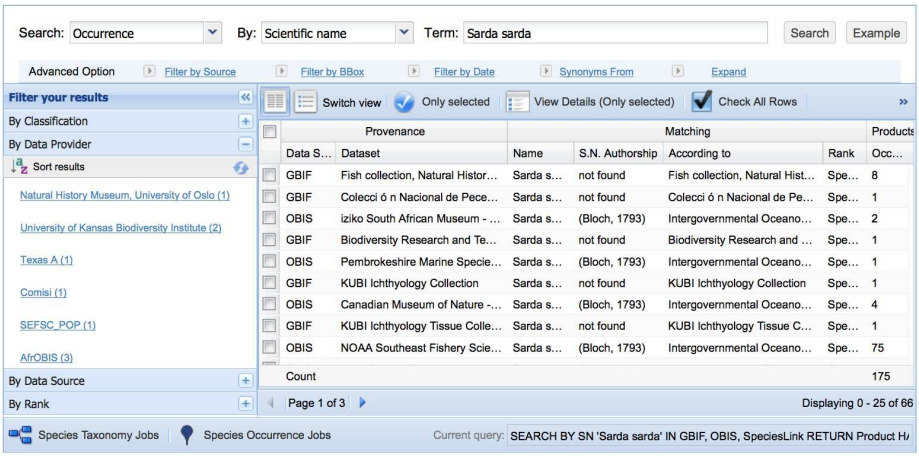

Figure 1: The SPD web interface with search facility running on *Sarda sarda*.

form, i.e., every dataset is described by carefully reporting typical Darwin Core information like (*i*) the original data provider, (*ii*) the author of the record, (*iii*) credits to the final provider, (*iv*) the species scientific name, (*v*) the coordinates of the occurrence, (*vi*) the basis of record, and (*vii*) the recording date.

The user is also provided with a number of facilities for inspecting the retrieved data. These allow to identify the "right" data, collect them and start forming a "research database". Among these facilities there are two diverse visualisations of the records belonging to the discovered occurrences datasets: a detailed one and a geospatial one (Fig. 2). Both these views allow to have access to a comprehensive description of every single occurrence point that has been identified via the SPD.

After selecting some occurrence points, SPD enables users to save such points in several formats, including CSV and Darwin Core. Such objects can be stored and shared with collaborators by relying on a *user workspace* (Assante et al., 2014). This is a core service of any virtual research environment. It is conceived to resemble a classical folder-based file system a user may be familiar with. The real added value of this file-system-like environment is represented by the large array of items it can manage in a seamless way and store in the Infrastructure.

2.2. Occurrence Data Preparation Facilities

After occurrence data have been retrieved, a processing phase is often required to clean data or to discover complementarity between different datasets. Data providers, in fact, can be collectors of other providers and this means that taking occurrence records from many sources could introduce redundancy in the extracted dataset. Furthermore, it cannot be assumed that data providers supervise the quality of the datasets they publish. Datasets can contain errors or can miss information. In other cases the providers can mix specimens and catch reports or survey observations. Cleaning operations are useful to make

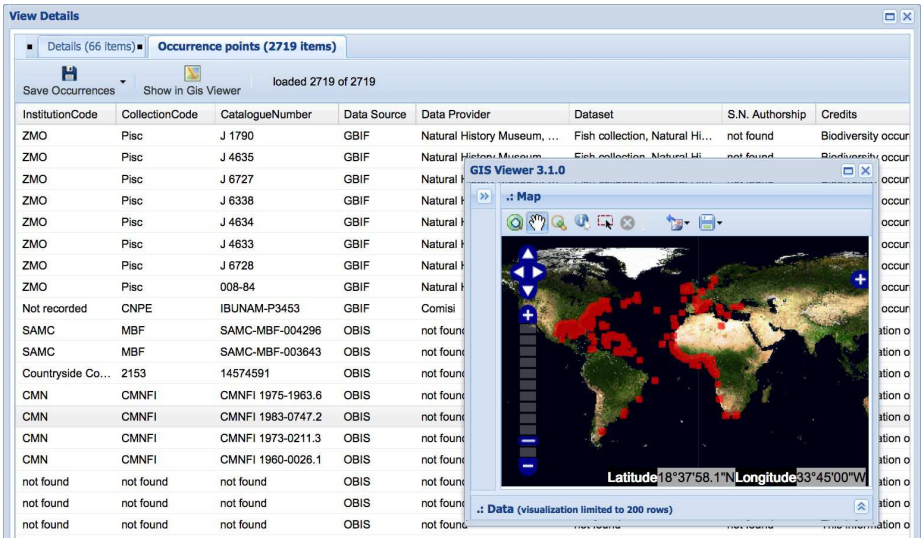

Figure 2: The SPD web interface displaying the selected datasets of species occurrences. The visible columns correspond to Darwin Core fields.

scientists able to use occurrence records in their experiments. Such experiments can focus on niche modelling, where occurrence points are assumed to approximate presence points and the environmental features attached to each point are correlated to the preferred habitat by a species. Otherwise, a scientist could use the occurrence records to evaluate the coverage or the quality of a field survey for a certain species. In this case, the aim is to understand if a representative number of occurrences has been collected, which contains a good heterogeneity of species habitat variables. Another possible usage of cleaning operations is the estimation of the degree to which the projection of a niche model on a certain area makes sense, i.e., how much the values produced by a model are reliable. If the range of variation of the environmental parameters taken at the presence (or absence) points of a species is representative for a certain projection area, then the projection domain is very similar to the training set. Thus, projecting the model onto that area is coherent with the information that was provided to the model during the preliminary training session.

D4Science is endowed with processing facilities for occurrence points which address all the above issues. Cleaning operations are performed by means of processing tools that allow to delete duplicates, to merge two datasets of occurrences or to select only points falling in marine areas. Errors or outliers detection is addressed by means of clustering techniques, which isolate zones having a low density of points or being far from the others. Furthermore, D4Science supplies facilities to evaluate the similarity among the features attached to presence points with respect to those attached to some other locations. These techniques are mainly based on the Principal Component Analysis (Jolliffe, 2005), a mathematical method that assesses the degree to which the dimensions of a vector

are independent on each other.

The application of these techniques to the occurrence data coming from the D4Science discovery facilities is not trivial. Each occurrence, in fact, is a rich information set containing additional data, other than a pair of geographical coordinates. Stating that two occurrence records are equal requires not only to evaluate if two points are close, but also to check if the scientific names and the authorships are equal or at least lexicographically similar. Furthermore, also the dates of creation and update should coincide. Such information set must be properly taken into account, in order to avoid situations in which two occurrences refer to the same scientific name, but with two different authorships, which could mean they refer to completely different species.

In the next subsections we will describe the operations D4Science supplies to its users to perform processing operations on occurrence records. These take into account the complete information context around each occurrence record extracted by the data discovery facilities. This set of operations is the result of a collaboration with many communities of practice involving biodiversity practitioners dealing with occurrence data.[5]. These communities principally aim at producing species distribution models and at enriching occurrence records with information about the environmental characteristics of species presence locations. By analysing their requirements, three main categories of processes have been identified: algebraic operations (cf. Sec. 2.2.1), clustering and outliers detections (cf. Sec. 2.2.2). Several examples can be found of applications of these processing categories to biodiversity data, e.g., Cumming et al. (2012); Graham et al. (2004, 2008). The main limitation of these examples, is that they are not included in the context of an integrated e-Infrastructure allowing users to explore data and to process them in an efficient way. Furthermore, algebraic operations applied to occurrence records are not usually flexible enough to rapidly account for coordinates approximations and discrepancies in recording dates. On the other hand, among the several approaches to clustering and outliers detection present in literature, we propose common algorithms that suit with the elicited requirements. Moreover, D4Science is an open development platform that uses a plugin approach for the deployment of new functionalities (Candela et al., 2009; Coro et al., 2014), thus accommodating requests for new algorithms requires low development effort.

2.2.1. Algebraic Operations on Occurrence Datasets

Differently from other solutions in the biodiversity domain, D4Science provides every scientist with a computer-assisted environment enabling to inspect the collected datasets and to understand which are the discrepancies and overlaps among such datasets. In fact, even if the datasets are somehow homogenised during the acquisition phase, this does not mean that their contents are comparable or ready to be used in a scientific experiment. For example, coordinates

[5]These communities include that operated by the iMarine project (`http://www.i-marine.eu`) and the EUBrazilOpenBio project (`www.eubrazilopenbio.eu/`)

could be given at different precision and authors names (or species names) could be written in different formats. There is no single "data format" that suits with any scientific experiment, then scientists need an environment facilitating their data preparation activities.

D4Science offers a number of algebraic operations specifically conceived to deal with species occurrence data. These include *union*, *intersection*, *subtraction* and *duplicates deletion* that use a probabilistic approach. Algebraic operations allow scientists to retrieve complementary or duplicate information among datasets.

The D4Science service for occurrence points manipulation is named *Occurrence Data Management* (ODM). It is endowed with a web-based user interface (Fig. 3) and it supports the above algebraic operations by using tolerance thresholds for assessing when two occurrence records are to be considered equal. Thresholds can be defined by every single user for every single operation and involve a spatial tolerance and a syntactic tolerance.

The *spatial tolerance* (T_{Sp}) is used to assess if two occurrences refer to the same point in the world, assuming a WGS-84 projection (True, 2004) for the coordinates. It represents the resolution at which a scientist considers two points to be the same: e.g., if $T_{Sp} = 0.5$ degree and the distance between two points is lower than 0.5 degree then the two points will be identified as potentially the same point.

The *syntactic tolerance* (T_{Sy}) evaluates the lexical similarity between the scientific names and the "*recordedBy*"[6] fields in two records. A normalized lexicographic distance (Levenshtein, 1966) $L_s(s_1, s_2)$ is used between the scientific names (s_1 and s_2) reported in two records. The same measure $L_r(r_1, r_2)$ is used on the "*recordedBy*" fields (r_1 and r_2) of the same records. Eventually, the product $L = L_s(s_1, s_2) * L_r(r_1, r_2)$ gives an overall lexical similarity between the records. If $L \leq T_{Sy}$, then the two records are declared to be similar.

A further comparison applies to the recording dates: if recording dates are reported in both the two records, they are checked to be the same, otherwise the check does not apply. A mismatching in the recording dates means that the two records are different.

We based the similarity comparison on the coordinates, the scientific name, the "*recordedBy*" field, and the recording date, as they contain the minimal information to identify an occurrence point. We assume, in fact, that two occurrence records are equal if and only if they refer to the same species, the same position, and were reported by the same person or Institution in the same date.

The motivations leading to this similarity comparison are: (i) as demonstrated by Vanden Berghe et al. (2014), the Levenshtein distance for lexicographic comparisons is a reliable measure to estimate the lexical similarity between two species scientific names. In particular, it is more indicated than other

[6]The term "recordedBy" refers to the Darwin Core specification. It indicates a list of names of people, groups, or organizations responsible for recording the original occurrence point.

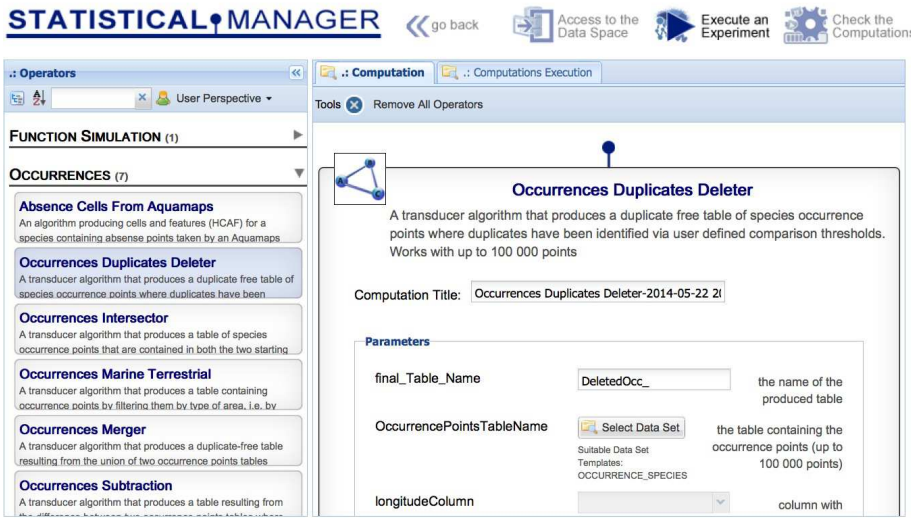

Figure 3: The setup phase of the "Occurrences Duplicates Deleter" procedure. On the left side a set of procedures is highlighted which can be applied to occurrence records.

approaches when no assumption can be made on the misspelling errors that are present in the transcriptions of species scientific names. Domain oriented measures gain better performance when names are written by taxonomists. Since our system addresses heterogeneous data providers, we assume that entries can contain a large variety of misspelling errors. Thus, we chose to adopt the most flexible, effective and domain independent measure; (*ii*) a strict comparison between dates has been adopted because dates carry very delicate information. Dates differing for even one day or one hour could indicate completely different records. Occurrence reports, in fact, are usually very precise about the recording dates and times of the occurrences.

At the end of the described comparison, two occurrence records are declared to be the same if (*i*) they are closer than T_{Sp}, (*ii*) they are similar over T_{Sy}, and (*iii*) the recording dates check is successful or does not apply. Between two similar records, the most recently modified is taken because it is assumed that the editing activity improved the record. It is obvious that when $T_{Sp} = 0.0$ and $T_{Sy} = 1.0$ the comparison reduces to a pure equality check. When $T_{Sy} = 0.0$ the system ignores the lexical comparisons.

The ODM algebraic operations rely on the above similarity evaluation. In the *union* procedure the service joins two occurrences sets A and B, excluding all the elements in B which are "similar to" elements in A. In the *intersection*, the system takes only the elements in A which are "similar to" elements in B. In the *subtraction*, it takes all the elements in A which are not "similar to" any element in B. Finally, in the *duplicates deletion* the service only takes the most recent records of A, excluding "similar" records in A itself.

2.2.2. Clustering and Outliers Detection

Clustering techniques allow to group real valued vectors by relying on their density or distance in an N–dimensional space. In the biodiversity context vectors usually refer to coordinates pairs or to environmental features. A side effect of clustering is that outliers can be identified as those points that are far from all the other groups. Thus, clustering also simulates a similarity measure for the vectors. Applications of such techniques range from automatic classification and characterization of information spaces to grouping and outliers detection. Examples of clustering procedures are density based algorithms like DBScan (Ester et al., 1996) and OPTICS (Ankerst et al., 1999), or distance based vector quantizations like K-means (MacQueen et al., 1967) and X-means (Pelleg and Moore, 2000). Clustering is also used to address the classification of outliers directly, for example in the case of the Local Outlier Factor (LOF) (Breunig et al., 2000).

D4Science hosts the DBScan, K-means, X-means and LOF algorithms, with a general purpose approach which can be applied to any table containing vectors of real values. The output of these procedures is a table were each row reports the elements of one vector along with two additional columns. These report information about the belonging cluster of each row and a boolean value indicating if the vector is an outlier. In the case of LOF, the algorithm reports the local deviation of a vector with respect to its neighbours.

Clustering algorithms in D4Science can be used in combination with the cleaning procedures described in the previous subsection. For example, after having deleted the duplicates from a dataset, the DBScan algorithm can highlight the presence of outliers, which could correspond, for example, to specimens stored in a museum or to reporting errors. Different clustering algorithms can detect different kinds of outliers. In the case clustering is applied to the environmental features attached to some occurrence points, it can detect precise locations that share similar habitat. If clustering is applied to biodiversity indicators, like frequency of observation, marginality, specialization or widespreadness, the grouping can highlight similar characteristics of species, for example their commonness or their threatening status.

2.2.3. Occurrence Points Representativeness

In the case environmental characteristics were attached to occurrence points, several techniques allow to discover similarity between such points. This is of particular interest for those scientists that want to investigate the similarities between the locations in which a species has been observed, but also to evaluate if the coverage given by a survey for a species is sufficient to describe its habitat. From the point of view of niche modellers, preliminary processing of features vectors can highlight useless features or can evaluate the potential robustness of the models to produce. One of the best known techniques in this context is the Principal Component Analysis (PCA) (Jolliffe, 2005), a mathematical procedure that aims to reduce the dimensionality of the features space. PCA uses an orthogonal transformation in the features space for producing independent variables called principal components. This transformation can be useful for

investigating the correlations among the environmental features used in niche models. Adding more dependent variables, in fact, usually does not result in better models.

A specific technique for occurrence points is the Habitat Representativeness Score (HRS) (MacLeod, 2010), which extends PCA by directly addressing biodiversity data. It was originally created to measure the degree to which sampled habitats are representative for a certain area of study. For example, HRS has been used for assessing the minimum number of surveys on a study area that are needed to cover a good heterogeneity of species habitat variables. HRS can be applied to two datasets of environmental features, one representing a sampled area and the other a geographical region of interest. A score is produced for each feature, ranging from 0 to 2, with 2 representing completely non-overlapping distributions of values. The lower the HRS the more similar data obtained from a survey are to the study area. Thus, HRS can be also used to assess how much the features associated to species occurrences are representative for the environmental characteristics of a certain area of study.

D4Science provides an implementation of the HRS algorithm applicable to two sets of environmental features: on one side those associated to the occurrence points of a species, on the other side those associated to locations that cover a certain area of study. An example of usage of such technique to assess the potential quality of a niche model can be found in Coro et al. (2013).

3. Experiments

The goal of the facilities discussed so far is to simplify the data acquisition and preparation phases as to enhance the availability of potential occurrence data. In this section we demonstrate their use by means of concrete examples. These experiments occur in many typical use cases identified by the D4Science community. In particular, we acquired data from diverse databases and then processed these data to highlight duplicates, differences between the datasets and possible outliers.

In the following, we present three experiments. In the first, we investigate complementarity between two big datasets referring to a well known list of species names. The datasets are not fully overlapping, and the experiment demonstrates how our facilities help to evaluate such overlap. In the second, we scale down to one single species, having a large amount of observations. We use this species to give a qualitative insight of the complementarity between the datasets hosted by diverse providers. The aim is to further demonstrate that the datasets contain complementary information and that our method is applicable also to single species investigation. In the third experiment, we used clustering to qualitatively demonstrate how a user can simply detect outliers by means of D4Science facilities. The meaning of the term outlier is here dependent on the parameters of the algorithm. Basic outliers (0 coordinate points) are discovered in the experiment and others are suggested by the system.

Table 1 reports the number of observations and data providers involved in each experiment. Overall, the experiments highlight the possible usefulness

	Data Provider	N. of Obs.	Distinct Obs.	Benchmark
Exp.1	GBIF	2,805,784	1,102,158	FAO Fact Sheets spp.
	speciesLink	8942	6576	FAO Fact Sheets spp.
Exp.2	GBIF	8791	6737	Cynodon dactylon
	speciesLink	288	165	Cynodon dactylon
Exp.3	GBIF+OBIS	2278	2104	Cetorhinus maximus

Table 1: Observation records involved in the experiments, with indication about the belonging benchmark dataset.

of our approach even from the perspective of data providers. A provider like speciesLink, for example, could use the D4Science facilities to understand the amount of its owned records that are complementary with respects to homologous data published by other providers. In other cases, applying clustering can automatically detect outliers and thus possible errors. This can lighten the cleaning effort that the data providers may manually apply.

Species name	Average Data Acquisition Time			
	GBIF	OBIS	speciesLink	SPD
Cynodon dactylon	378 (s)	n.a.	455 (s)	357 (s)
Cetorhinus maximus	113 (s)	101 (s)	n.a.	67 (s)

Table 2: Comparison of average times for accessing and retrieving occurrence records datasets. We compare SPD to a "manual" interaction, where the user accesses the website of a single data provider, searches for occurrence records using the scientific name of a species and asks for producing one dataset containing all the records. We stopped the time calculation when the production of the dataset is complete. Time has been calculated as an average on 5 interactions.

Table 2 reports an indication of the average data acquisition time for Experiments 2 and 3. Although this is not a systematic study, it gives a flavour of the enhancements introduced by SPD. The table reports the time a user needs for "manually" interacting with the website of each data provider compared with the time when using SPD. In particular, it reports the duration of the search for the occurrence records of a species, plus the time required to produce a dataset containing all the occurrence records. In the case of SPD this is done in one single step dealing with both the data providers in parallel. The table reports the average time on 5 repetitions of the same interaction. From this it emerges that the time required by SPD is less than the one required by each single data provider website interaction. This is due to the fact that SPD sends simultaneous requests to each provider and collects data using parallel threads.

Furthermore, since the invoked web services are the same as those of GBIF and specieLink, this means that SPD provides a more lightweight web interface and a faster production process for occurrence records lists after the retrieval process.

Experiment 1

As first experiment, we report about the effectiveness of our procedures in highlighting the diversity between two datasets. To such aim, we selected the GBIF and speciesLink data providers, which are likely to be disjoint (to some extent) because of their different contributors and spatial coverages, i.e., global and regional respectively. We show how our procedure allows a user to evaluate such complementarity. In particular, we calculated the intersection on a large amount of species observations under different configurations of our procedures. In this experiment, intersection is reported as the percentage of matching observations with respect to the total observations in the speciesLink dataset (the smaller dataset). For such reason, we used the percentage of agreement as quantitative quality measure.

For the selected datasets, the coverage of speciesLink by GBIF was expected to be only partial, and we wanted to confirm this by means of our procedures. We used an evaluation benchmark made by a well known list of species, i.e., the FAO Fact Sheets list [7]. This list includes 548 aquatic species of commercial interest. Via the SPD service, we discovered that for such list (a) GBIF contains 2,805,784 observation records, of which 1,102,158 are different according to a pure equality check, and (b) speciesLink, contains 8942 observation records, of which 6576 records are strictly distinct.

Figure 4 reports on the effect of the variation of T_{Sp} (with T_{Sy} fixed to 1) on the intersection amount between these two datasets. In other words, the chart reports the portion of the speciesLink observations covered by GBIF when the spatial resolution for the comparison gets coarser, considering an exact match between the names in the records. The chart shows that the datasets have partial superposition, as we expected.

Figure 5 reports on how much the missing superposition is due to lexicographic distances. Thus, records corresponding to the same species were compared without considering their lexical distances ($T_{Sy} = 0$). The chart shows that the percentages slightly increase but the superposition is still partial. This effect confirms the complementarity between the datasets.

Figure 6 reports on the effects of variations of the T_{Sy} parameter. We fixed a coarse resolution for the occurrence records ($T_{Sp} = 1$) and changed the syntactic threshold. The figure highlights the fall of the similarity between records when the lexical threshold becomes more restrictive.

A similar effect is reported by Figure 7, which reports on the variation of the T_{Sy} parameter when the observations are required to be exactly at the same

[7] http://www.fao.org/fishery/species/search/en

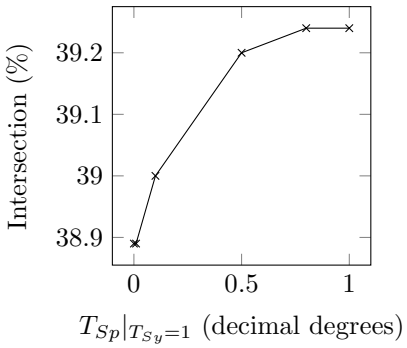

Figure 4: Effect of the variation of T_{Sp} on the intersection between the GBIF and the speciesLink records for the FAO Fact Sheets species. The chart keeps T_{Sy} fixed to 1 (pure equality check for scientific names and recorders).

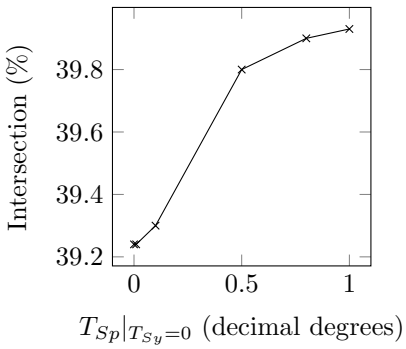

Figure 5: Effect of the variation of T_{Sp} on the intersection between the GBIF and the speciesLink records for the FAO Fact Sheets species. The chart keeps T_{Sy} fixed to 0 (no lexical checks for corresponding species).

place ($T_{Sp} = 0$). The trend is similar to the previous one, and this means that the intersection points are likely to be the exactly the same in both the datasets.

Overall, the charts highlight that the complementary points remain distinct even when a coarser spatial granularity and a more flexible lexical similarity are used. This highlights the importance of the disjoint observations in speciesLink, which are related to complementary information about the species.

Experiment 2

As second experiment, we report a deeper analysis on a particular case, taken from the data providers involved in experiment 1. We selected a specific and representative use case to verify the general behaviour detected by the charts in the previous experiment. The aim of this experiment is to confirm that GBIF and speciesLink contain complementary information for a species with a large number of observations associated. This experiment also demonstrates the effectiveness of our solution in scaling down to one single use case, other than in analysing large datasets.

In particular, we compared the occurrence points of *Cynodon dactylon* (bermuda grass) coming from the GBIF and the speciesLink data providers separately. As the bermuda grass is a very common plant it has a large number of records, thus it was likely to be reported several times in both the datasets. On the other side, the speciesLink dataset was smaller than the GBIF one, and we explored their complementarity in details. Via the SPD service, we retrieved 8791 records from GBIF and 288 records from speciesLink. By applying a *duplicate deletion* with pure equality check ($T_{Sy} = 1$) to each dataset, we obtained 6737 distinct records for GBIF and 165 for speciesLink.

In order to assess if the two sets were disjoint we performed 3 *intersection* operations of speciesLink with respect to GBIF, varying the thresholds config-

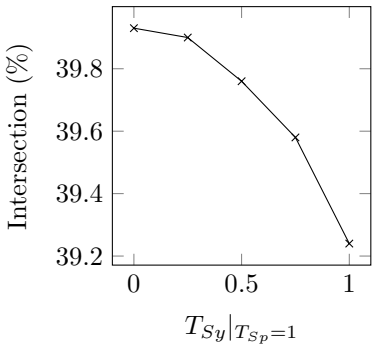

Figure 6: Effect of the variation of T_{Spy} on the intersection between the GBIF and the speciesLink records for the FAO Fact Sheets species. The chart keeps T_{Sp} fixed to 1 degree.

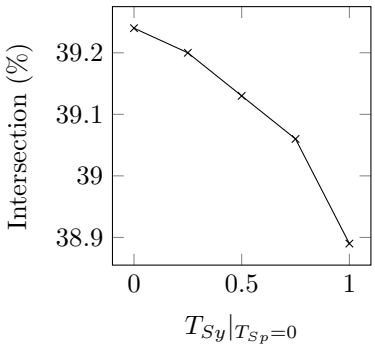

Figure 7: Effect of the variation of T_{Spy} on the intersection between the GBIF and the speciesLink records for the FAO Fact Sheets species. The chart keeps T_{Sp} fixed to 0 degrees.

urations. In all the cases the intersection set was empty, thus there were no overlaps. In the first comparison we set $T_{Sp} = 0.0$ and $T_{Sy} = 1$. In the second we removed the lexical comparisons ($T_{Sy} = 0.0$) and applied a pure equality check to the coordinates ($T_{Sp} = 0.0$). In the third comparison we used more spatial tolerance ($T_{Sp} = 0.01$ and $T_{Sy} = 0.0$) and again the intersection set was empty. Increasing the spatial tolerance would have been not significant to our use, then we did not go forward with the experiment. This experiment demonstrated that GBIF did not contain the speciesLink records at all, at 0.01 degree resolution. Thus, this is in line with the general behaviour highlighted by the charts in the previous experiment.

Experiment 3

In the third experiment, we demonstrate the effectiveness of outliers detection. The concept of outlier observation is dependent on the user's definition. Scientists could define as outliers some points well separated from other points. In other cases, they could decide that points on which they do not agree with the reporters are outliers. Thus, the definition of outlier can be subjective. Here we report a qualitative evaluation of our outliers detection procedure, which shows how to simply detect common 0 coordinate outliers. Furthermore, we also show how the output of the clustering changes when using different input parameters. This allows scientists to tune our system for better grouping points or for defining outliers.

We used the SPD service to acquire occurrence records for *Cetorhinus maximus* (basking shark) from GBIF and OBIS. The discovered data were automatically saved in one dataset, containing records from both the datasets. For this experiment, we took occurrences coming from the collections of the University of Oslo, as well as from FishBase and from the Northeast Fisheries Science Center (NEFSC). We kept the intersections between the OBIS and the GBIF

Figure 8: The distribution of the basking shark occurrence records coming from the collections of the University of Oslo, of FishBase and of the Northeast Fisheries Science Center, coming from a query to SPD on the OBIS and GBIF datasets.

datasets under control. In particular, we took the complete FishBase collection from OBIS, while from GBIF we took only the FishBase occurrences hosted by GBIF-Sweden. Furthermore, we took the collections from the University of Oslo from GBIF only, and the NEFSC datasets from OBIS only. The total number of occurrences from both the datasets was 2278, which contained duplicate occurrences only for a subset of FishBase. As first cleaning step, we applied the *duplicates deletion* operation using $T_{Sp} = 0.5$ and $T_{Sy} = 0.8$. Thus, we admitted half degree spatial tolerance and low lexical mismatching on the species name and authorships transcriptions.

Figure 8 depicts the distribution of the occurrence points we used for this experiment. The dataset resulting from the duplicate deletion process contained 2104 elements, which resulted just from the deletion of the FishBase duplicates. As it can be visually noticed from Fig. 8, some points having (0,0) coordinates are present among the occurrences, along with isolated points. The (0,0) coordinate records are common errors. They are uninformative records that are likely to be deleted in most of the analyses. In order to automatically highlight the presence of such points and of other possible outliers, we applied clustering techniques.

Figure 9 depicts the result of the application of the DBScan algorithm to our basking shark dataset without duplicates, while Figure 10 reports the output of X-means. In this experiment DBScan was run using the following parameters configuration (Ester et al., 1996): *epsilon* = 10 and *minimum number of points required to form a cluster* = 2. On the other side, X-means was run using the following parameters (Pelleg and Moore, 2000): *max iterations* = 100, *minimum number of clusters* = 1, *maximum number of clusters* = 50 and *minimum number of points required to form an outlier set* = 2. DBscan automatically

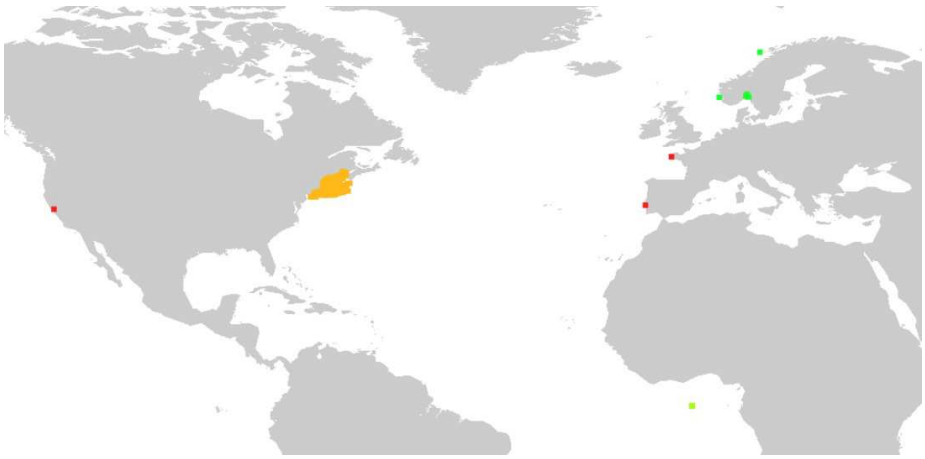

Figure 9: Representation of the application of the DBScan (density based) algorithm to the occurrence points of the basking shark. The algorithm was applied to the collections of the University of Oslo, of FishBase and of the Northeast Fisheries Science Center, coming from a query to SPD on the OBIS and GBIF datasets and after duplicates deletion. Each colour indicates a different cluster.

detected 5 clusters, while X-means detected 3 clusters. Both the algorithms indicated that the (0,0) coordinate points were outliers, which demonstrates the effectiveness of the procedure in detecting common errors. When we configured the algorithm by setting to 1 the minimum number of points required to form an outlier, then DBScan detected that the points in the West coast of USA, and the ones in France and Spain were other possible outliers. In this case X-means did not detect outliers. By looking at the maps, it is possible to notice that DBScan considered the points near France and Spain as not belonging to the same density cluster of the points in North Europe, while X-means was more coarse from this point of view. X-means calculated that all the points in Europe belong to the same cluster because it only considered their relative distance with respect to the points in USA. As result of this analysis, users are able to obtain a dataset without duplicates, from which they could cut the outliers off. Furthermore, they could concentrate on some of the clusters, which contain clean information coming from the two data sources, and possibly decide to identify also these as outliers.

4. Discussion and Conclusion

Peterson et al. (2010) well highlighted the benefits for biodiversity-related tasks resulting from information infrastructures and approaches improving data and analytical software availability.

This paper has introduced an innovative infrastructure-based approach aiming at offering data acquisition and data preparation facilities on species occurrences data. In particular, it has presented a data acquisition facility that

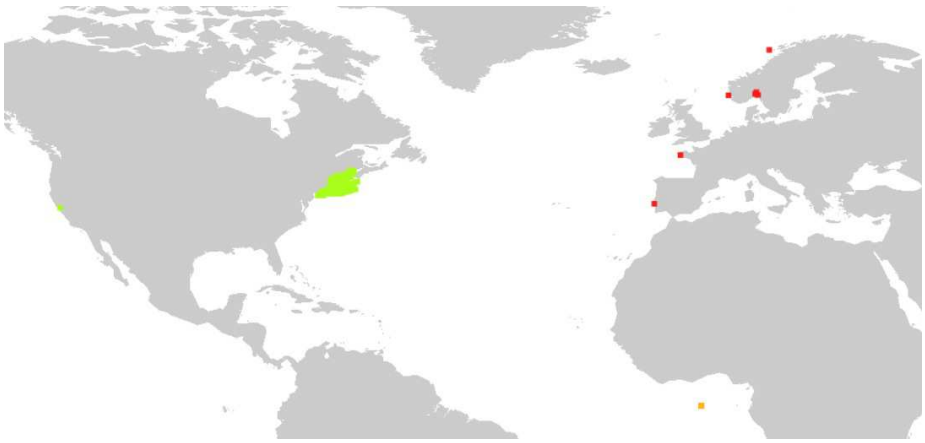

Figure 10: Representation of the application of the X–Means (distance based) algorithm to the occurrence points of the basking shark. The algorithm was applied to the collections of the University of Oslo, of FishBase and of the Northeast Fisheries Science Center, coming from a query to SPD on the OBIS and GBIF datasets and after duplicates deletion. Each colour indicates a different cluster. The difference with respect to DBScan is that X–Means identifies all the points in Europe as belonging to the same cluster.

simplifies the discovering of and access to relevant data by abstracting over the peculiarities of the data owners/publishers while guaranteeing provenance and attribution. Moreover, it has illustrated a set of data preparation facilities that empowers scientists to deeply analyse the collected data in order to identify potential duplications and discrepancies that depends on scientist's specific needs.

The implementation of these facilities is nicely integrated with existing efforts on databases and information systems development by following an approach that supplements these initiatives contributing to enlarge the visibility and use of the published data.

The described facilities have been developed and used in two projects dealing with species data: the iMarine project[8] focusing on marine species and the EUBrazilOpenBio project[9] focusing on plants. These facilities are currently made publicly available via the portals operated by these projects and can be used by any scientist willing to exploit them. The feedback received in these contexts[10] is positive, the facilities offered are simplifying scientists practices in accessing data and managing them.

Besides the facilities illustrated in this paper, the D4Science infrastructure offers a large variety of other facilities to support also the management of other

[8] iMarine project website `http://www.i-marine.eu`

[9] EUBrazilOpenBio project website `www.eubrazilopenbio.eu/`

[10] Dedicated meeting and events have been organised to collect this feedback. Moreover, community members are using the system to produce periodic validation reports, e.g., Ellenbroek and Pagano (2012).

biodiversity related data like taxonomic items. For instance, it is possible to easily build checklists of species names from diverse databases via the SPD and then compare these checklists with the aim to identify discrepancies across diverse taxonomies (Amaral et al., 2014).

The following enhancements for the so described facilities are planned. The lexical similarity supporting data preparation will be enhanced in order to take into account more information associated with occurrence records. Moreover, an appropriate weighing scheme will be defined. Moreover, facilities aiming at integrating and enriching occurrence records with environmental information are under development.

Acknowledgements

The work reported has been partially supported by the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644) and the *EUBrazilOpenBio* project (FP7 of the European Commission, FP7-ICT-2011.EU-Brazil, Contract No. 288754). We thank the two anonymous reviewers for their comments and suggestions which significantly contributed to improve the quality of the paper.

References

Amaral, R., Badia, R. M., Blanquer, I., Braga-Neto, R., Candela, L., Castelli, D., Flann, C., De Giovanni, R., Gray, W. A., Jones, A., Lezzi, D., Pagano, P., Perez-Canhos, V., Quevedo, F., Rafanell, R., Rebello, V., Sousa-Baena, M., Torres, E., 2014. Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure. Concurrency and Computation: Practice and Experience n/a, n/a.

Ankerst, M., Breunig, M. M., Kriegel, H.-p., Sander, J., 1999. OPTICS: ordering points to identify the clustering structure. In: Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 49–60.

Assante, M., Candela, L., Castelli, D., Pagano, P., January 2014. The D4Science Research-Oriented Social Networking Facilities. ERCIM News 96, 44–45.

Bach, K., Schäfer, D., Enke, N., Seeger, B., Gemeinholzer, B., Bendix, J., 2012. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. Ecological Informatics 11, 16–24.

Bendix, J., Nieschulze, J., Michener, W. K., 2012. Data platforms in integrative biodiversity research. Ecological Informatics 11, 1–4.

Borgman, C., 2011. The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology, 1–40.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, D. W., Laurie, G., O'Neill, O., Rawlins, M., Thornton, D. J., Vallance, P., Walport, M., June 2012. Science as an open enterprise. Tech. rep., The Royal Society.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J., May 2000. Lof: Identifying density-based local outliers. SIGMOD Rec. 29 (2), 93–104.
URL http://doi.acm.org/10.1145/335191.335388

Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F., 2013. Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience.
URL DOI:10.1002/cpe.3030

Candela, L., Castelli, D., Pagano, P., 2009. D4Science: an e-Infrastructure for Supporting Virtual Research Environments. In: Agosti, M., Esposito, F., Thanos, C. (Eds.), Post-proceedings of the 5th Italian Research Conference on Digital Libraries - IRCDL 2009. DELOS: an Association for Digital Libraries, pp. 166–169.

Candela, L., Castelli, D., Pagano, P., October 2008. gCube: A Service-oriented Application Framework on the Grid. ERCIM News (72), 48–49.

Candela, L., Castelli, D., Pagano, P., October 2010. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. ERCIM News (83), 32–33.

Candela, L., Castelli, D., Pagano, P., 2013. Virtual research environments: an overview and a research agenda. CODATA Data Science Journal 12, GRDI75–GRDI81.

Candela, L., De Faveri, F., Lelli, L., Mangiacrapa, F., Marioli, V., Pagano, P., 2014. Accessing biodiversity databases: a domain specific query language and a unifying data model. Technical Report 2014-TR-26, Istituto di Scienza e Tecnologie dell'Informazione A. Faedo, CNR.

Clark, T., Martin, S., Liefeld, T., 2004. Globally distributed object identification for biological knowledgebases. Briefings in Bioinformatics 5 (1), 59–70.

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A. T., The VertNet Steering Committee, 02 2010. Vertnet: A new model for biodiversity data sharing. PLoS Biol 8 (2), e1000309.
URL http://dx.doi.org/10.1371%2Fjournal.pbio.1000309

Coro, G., Pagano, P., Ellenbroek, A., 2013. Combining simulated expert knowledge with neural networks to produce ecological niche models for latimeria chalumnae. Ecological Modelling 268 (0), 55 – 63.
URL http://www.sciencedirect.com/science/article/pii/S0304380013003980

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo L., 2014. Parallelising the Execution of Native Data Mining Algorithms for Computational Biology. Submitted to Concurrency and Computation: Practice and Experience

Costello, M. J., 2009. Motivating online publication of data. BioScience 59 (5), 418–427.

Cumming, G. S., Gaidet, N., Ndlovu, M., 2012. Towards a unification of movement ecology and biogeography: conceptual framework and a case study on afrotropical ducks. Journal of biogeography, 39(8), 1401-1411.

D4Science.org, 2012. D4Science Hybrid Data Infrastructure.
URL www.d4science.org

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G., 2002. Pangaea – an information system for environmental sciences. Computers & Geosciences 28 (10), 1201 – 1210.
URL http://www.sciencedirect.com/science/article/pii/S0098300402000390

Edwards, J. L., Lane, M. A., Nielsen, E. S., 2000. Interoperability of biodiversity databases: Biodiversity information on every desktop. Science 289 (5488), 2312–2314.

Ellenbroek, A., Pagano, P., 2012. EA-CoP Validation. D3.4 iMarine Project Report.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., Gemeinholzer, B., 2012. The user's view on biodiversity data sharing - investigating facts of acceptance and requirements to realize a sustainable use of research data. Ecological Informatics 11, 25–33.

Ester, M., Kriegel, H., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining. AAAI Press, pp. 226–231.

Foster, I., Zhao, Y., Raicu, I., Lu, S., 2008. Cloud Computing and Grid Computing 360-Degree Compared. In: Grid Computing Environments Workshop, 2008. GCE '08. pp. 1–10.

Global Biodiversity Information Facility, 2011. GBIF Strategic Plan 2012-2016: Seizing the Future.
URL http://links.gbif.org/sp2012_2016.pdf

Goddard, A., Wilson, N., Cryer, P., Yamashita, G., 2011. Data hosting infrastructure for primary biodiversity data. BMC Bioinformatics 12 (Suppl 5), S5.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., 2008. The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45(1), 239-247.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A. T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in ecology & evolution, 19(9), 497-503.

Grassle, J., 2000. The Ocean Biogeographic Information System (OBIS): an online, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography 13 (3), 5–7.

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C., Vandenberg, J., July 2002. Online scientific data curation, publication, and archiving. Technical Report MSR-TR-2002-74, Microsoft Research.

Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Jetz, W., McPherson, J. M., Guralnick, R. P., 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology & Evolution 27 (3), 151 – 159.

Jolliffe, I., 2005. Principal component analysis. Wiley Online Library.

Jones, A. C., White, R. J., Orme, E. R., 2011. Identifying and relating biological concepts in the catalogue of life. Journal of Biomedical Semantics 2 (7).

Lenzerini, M., 2002. Data integration: a theoretical perspective. In: PODS '02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM Press, New York, NY, USA, pp. 233–246.

Levenshtein, V., 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707–710.

MacLeod, C., 2010. Habitat representativeness score (hrs): a novel concept for objectively assessing the suitability of survey coverage for modelling the distribution of marine species. Journal of the Marine Biological Association of the United Kingdom 90 (07), 1269–1277.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. California, USA, p. 14.

Meng, C., December/January 2004 2004. Biological information standards. Bulletin of Association for Information Science and Technology 30 (2), n/a.

Michener, W. K., Allard, S., Budden, A., Cook, R. B., Douglass, K., Frame, M., Kelling, S., Koskela, R., Tenopir, C., Vieglais, D. A., 2012. Participatory design of DataONE - enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics 11, 5–15.

Michener, W. K., Jones, M. B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27 (2), 85–93.

Pelleg, D., Moore, A., 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, pp. 727–734.

Peterson, A. T., Knapp, S., Guralnick, R., Sobern, J., Holder, M. T., 2010. The big questions for biodiversity informatics. Systematics and Biodiversity 8 (2), 159–168.
URL http://www.tandfonline.com/doi/abs/10.1080/14772001003739369

Roberts, D., Moritz, T., 2011. A framework for publishing primary biodiversity data. BMC Bioinformatics 12, I1.

TDWG, 2005. Access to Biological Collections Data - ABCD. Version 2.06.
URL http://www.tdwg.org/activities/abcd/

TDWG, 2010. TAPIR - TDWG Access Protocol for Information Retrieval. Version 1.0.
URL http://www.tdwg.org/activities/abcd/

True, S., April 2004. Planning the future of the World Geodetic System 1984. In: Position Location and Navigation Symposium, 2004. PLANS 2004. pp. 639 – 648.

Vanden Berghe E., Bailly N., Coro G., Fiorellato F., Aldemita C., Ellenbroek A., Pagano P., 2014. BiOnym: a flexible workflow approach to taxon name matching. Technical Report 2014-TR-22, Istituto di Scienza e Tecnologie dell'Informazione A. Faedo, CNR.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Robertson, R. D. T., Vieglais, D., 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1).

Wiederhold, G., 1992. Mediators in the Architecture of Future Information Systems. Computer 25 (3), 38–49.