Technical Report, No. 2013-TR-035, 2013
Consiglio Nazionale delle Ricerche - Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"

D4Science Facilities for Managing Biodiversity Databases

Leonardo Candela*, Donatella Castelli, Gianpaolo Coro, Federico De Faveri, Angela Italiano, Lucio Lelii, Francesco Mangiacrapa, Valentina Marioli, Pasquale Pagano

Abstract
During the last years, considerable progresses have been made in developing on-line species occurrence databases. These are crucial in scientific activities on biodiversity, including the generation of species distribution models, which play an important role in conservation efforts. 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 — Hybrid Data Infrastructure — Virtual Research Environment

[1] *Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy*
***Corresponding author**: leonardo.candela@isti.cnr.it

Contents

1. Introduction

Data sharing in the research domain is a practice whose benefits are nowadays well understood by both data *owners* and data *consumers* [1, 2, 3]. 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 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 recent 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 [4]. 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 to 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 is particularly relevant within the ecology scientific community [5, 6, 7]. Large scale initiatives have been launched in the past years, either at global – e.g., *GBIF* [8], *OBIS* [9], *VertNet* [10], *Catalogue of Life* [11] – or regional level – e.g., *speciesLink*[1] and *List of Species of the Brazilian Flora*[2] – to support the worldwide sharing of various collection of biodiversity data. The development of standards for data sharing has been promoted by establishing appropriate interest groups, e.g., the Biodiversity Information Standards (TDWG also known as Taxonomic Databases Working Group). Domain specific standards have been devel-

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

oped to focus on different interoperability aspects, e.g., *Darwin Core* [12] and *ABCD* [13] for data representation, *DiGIR* and *TAPIR* [14] for distributed data discovery, *LSIDs* [15] 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. [16] describe and analyse them by reviewing the current state of biodiversity data hosting and discussing the technological and social barriers affecting data sharing. 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 offering of "*seamless data access, integration, analysis, visualisation and use*" [17]. 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" [16]. 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 the 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 [18], DataONE [19] and Map of Life [20].

This paper introduces one of these new infrastructures, namely D4Science [21, 22], by discussing in particular the type of facilities it 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 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 supplanting) databases and information systems mandates and arrangements. They thus contribute to the implementation of the global biodiversity open endeavour envisaged by many [16, 24, 25].

[3]The term "as-a-Service" has been introduced in the context of the Cloud technologies [23], which help in assessing the "fitness for purpose" of the retrieved data. 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".

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 large initiatives 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 offers a rich array of data 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* [26, 27], 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". This means that such facilities cannot only be accessed through the D4Science portal, but can also be consumed automatically, via Internet, by other service providers. Among its facilities D4Science offers (*i*) a seamless access to third-party repositories and information systems and (*ii*) an open pool 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 praxis.

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 or to modify their strategies for data publication.

D4Science offers a service for species occurrence data discovery and access. This is conceived as a sort of mediator service [28] over a number of databases. The aim is to achieve the following key goals: (*i*) to *hide heterogeneity*, i.e., to abstract over differences in location, protocols, and models offered by each single database via dedicated plug-ins; (*ii*) to *embrace heterogeneity*, i.e., to allow for multiple loca-

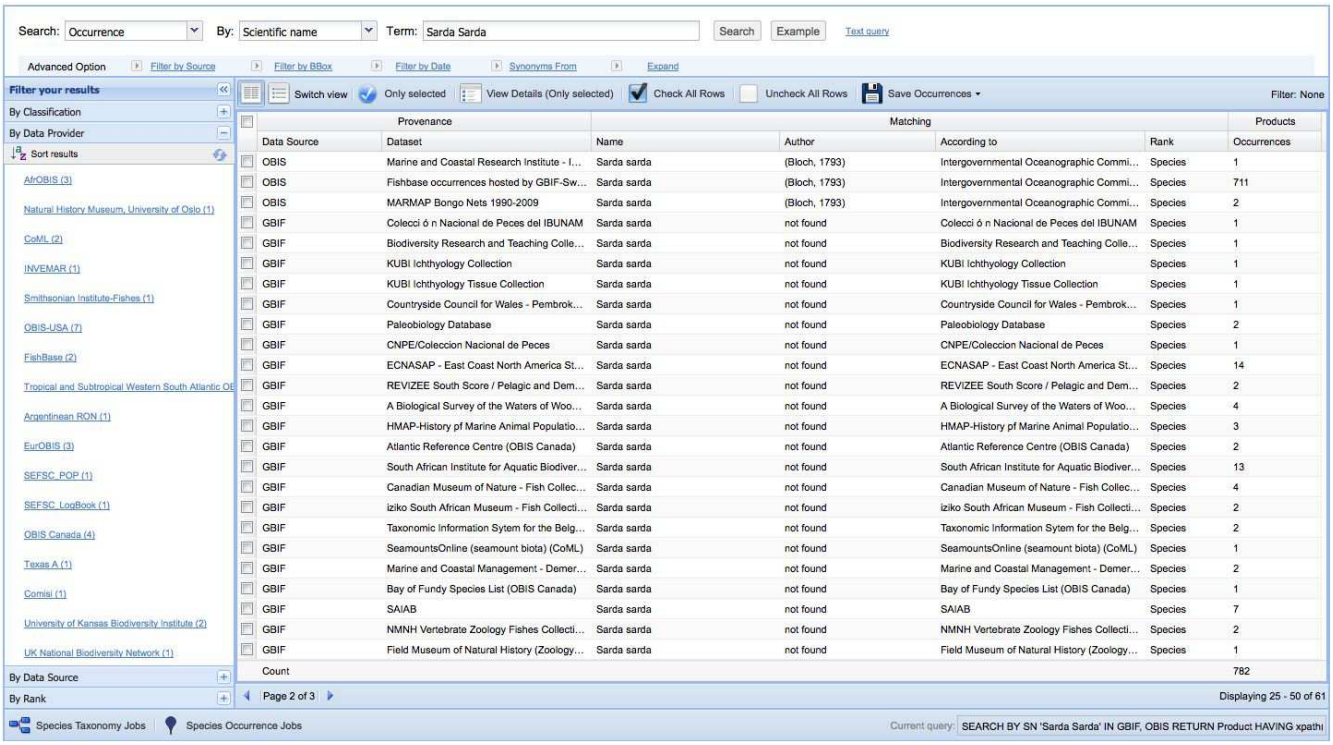

Figure 1. The SPD web interface with search facility running on *Sarda sarda*.

tions, protocols, and models by exposing the aggregated data in multiple ways; (*iii*) to *scale*, i.e., to retain good throughput under heavy load and high availability in the face of partial failures.

The D4Science service for species occurrences is named *Species Products Discovery* (SPD) and is endowed with a web interface. Fig. 1 depicts such interface, where it is possible to notice a search panel (on the top), a results view panel (on the right) and a classification panel (on the left). In addition to species occurrence data, the service supports discover and access to nomenclature data (Taxonomic items). However, the features associated with this type of information are out of the scope of this paper.

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. In order 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.

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 datasets discovered by means of a query. Furthermore, to overcome the potential issues related with taxonomy heterogeneities, the service relies on an automatic query expansion mechanism, i.e., the user query is automatically augmented with "similar" species names. In addition, 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 occurrence data discovered are presented to the user in an homogenised 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. Moreover, SPD provides the user with diverse aggregated views over the discovered datasets. It clusters the datasets with respect to the classification, the data provider, the database, and the rank.

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*, that is another ser-

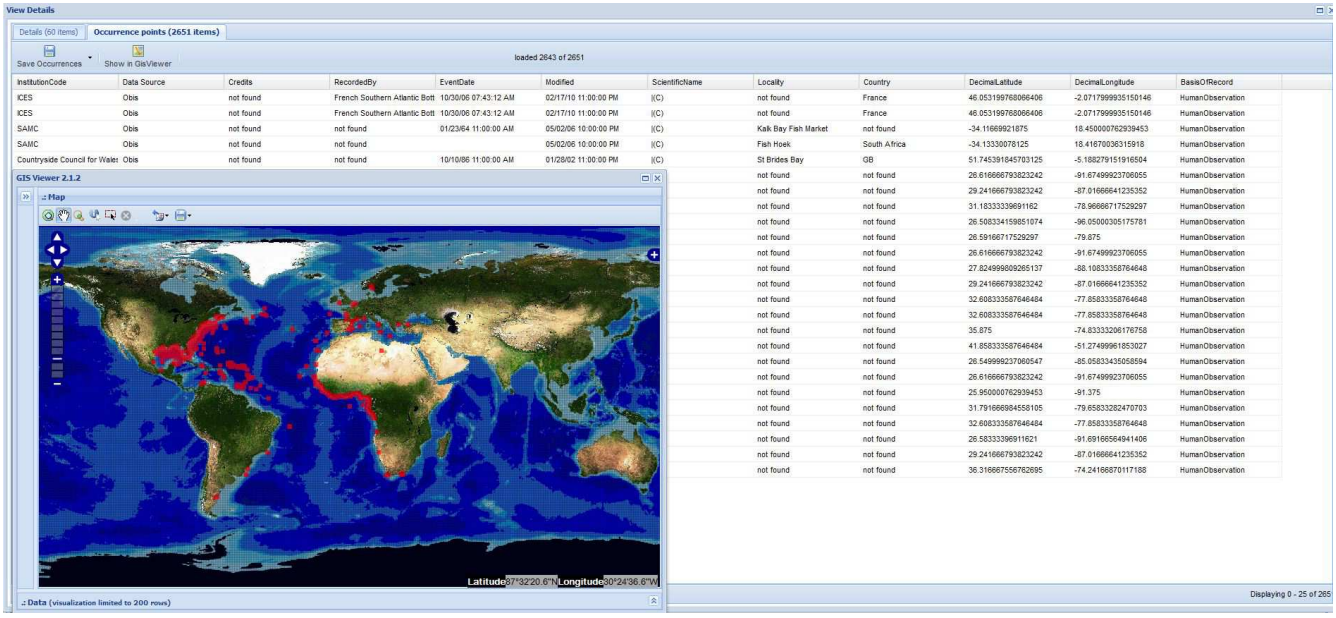

Figure 2. The SPD web interface displaying the selected datasets of species occurrences. The visible columns correspond to Darwin Core fields.

vice offered by the D4Science Infrastructure. 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

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 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, among other data manipulation facilities, 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 previously collected datasets.

The D4Science service for occurrence points manipulation is named *Occurrence Data Management* (ODM). It is endowed with a web interface 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* (SpT) is used to assess if two occurrences refer to the same point in the world, assuming a WGS-84 projection [29] for the coordinates. It represents the resolution at which a scientist considers two points to be the same: e.g., if $SpT = 0.5$ 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* (SyT) evaluates the lexical similarity between the scientific names and the "recordedBy"[4] fields in two records. A normalized lexicographic distance [30] $L_s(s_1, s_2)$ 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 $L \leq SyT$, 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, for our scopes. We assume, in fact, that

[4]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.

two occurrence records are equal if only if they refer to the same species, the same position and were reported by the same person or Institution in the same date.

At the end of the described comparison, two occurrence records are declared to be the same if (*i*) they are closer than SpT, (*ii*) they are similar over SyT, and (*iii*) the recording dates check is successful or does not apply. Between two similar records, the most recently modified is taken. It is obvious that when $SpT = 0.0$ and $SyT = 1.0$ the comparison reduces to a pure equality check. When $SyT = 0.0$ 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.

3. Results

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 this with two concrete examples. In particular we acquire data on the same species from two diverse databases and then compare these two dataset to highlight the differences among them.

The first example is based on the *Solea solea* marine species. The SPD service is used to acquire in a single step *Solea solea* all the occurrence data from GBIF and OBIS. All the discovered records are saved in two separate CSV files, one for GBIF and one for OBIS. Such files contain respectively 57,085 occurrence records (OBIS) and 2324 occurrence records (GBIF). We then applied the *duplicates deletion* operation by using $SpT = 0.0$ and $SyT = 1.0$ (pure equality check) and ended in 10,542 distinct records for OBIS and 1871 distinct records for GBIF.

To understand the differences between these two datasets and thus to demonstrate that there is a potential added value resulting from their identification (additional points), we perform a number of *subtraction* operation. A first subtraction operation was performed by using a pure equality check configuration between the OBIS unique points and the GBIF points. This operation revealed no overlap between the two sets. A second subtraction was performed by increasing the tolerance ($SpT = 0.0001$ and $SyT = 0.8$). Also this comparison revealed no overlap. Since OBIS is among the GBIF data publishers, this could mean that the representation of occurrences in native OBIS was different from the one in GBIF. A third subtraction was performed by increasing the spatial tolerance to 0.01 degree and again no superposition was found. This could mean that the "recordedBy" field or the scientific names were different in the two datasets. A forth subtraction was performed by using a lexical threshold equal to 0.0 (and $SpT = 0.01$) as to rely on the spatial distance only. This leads to the identification of 183 distinct records that are in both the datasets when compared with a 0.01 degree tolerance. By performing manual checks we confirmed that the "recordedBy" fields contained differences in the names formats. We found differences also in the scientific name fields. These results demonstrate that via the D4Science facilities it has been possible to collect a large number of *Solea solea* unique occurrence records which is neither available by interacting with GBIF nor with OBIS when using these systems in isolation. Thus, even if GBIF collects data from OBIS, the coverage is not complete. The user can also measure the degree of superposition between the two datasets. The choice to stop the analysis is on the user's side, who can decide to take into account the lexical similarities or to rely only on the spatial distance.

The second example is based on the Bermuda Grass (*Cynodon dactylon*) plant species, a very common plant that should have a large number of records. In this case we compare data coming from GBIF and speciesLink. The speciesLink database is smaller than the GBIF one. Via the SPD service we retrieved 8791 records from GBIF and 288 records from speciesLink. By applying a *duplicate deletion* with pure equality check to both the datasets, we obtained 6737 records for GBIF and 165 for speciesLink.

In order to assess that the two sets were disjoint we performed 3 *intersections* operations, varying the threshold configurations. In all the cases the intersection set was void, thus there are no overlaps. In the first comparison we set $SpT = 0.0$ and $SyT = 100$. In the second we removed the lexical comparisons ($SyT = 0.0$) and applied a pure equality check on the coordinates ($SpT = 0.0$). In the third comparison we used more spatial tolerance ($SpT = 0.01$ and $SyT = 0.0$) and again the intersection set was void. Increasing the spatial tolerance would have been not significant to our use, then we stopped the experiment. This experiment demonstrated that GBIF did not contain the speciesLink records at all, at 0.01 degree resolution.

Overall, this example highlights the possible usefulness of our approach even from the perspective of the data providers. A provider like speciesLink could use D4Science facilities to understand the amount of its owned records that are complementary with respects to homologous data published by other providers.

4. Discussion

Townsend Peterson et Al. [25] 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 simpli-

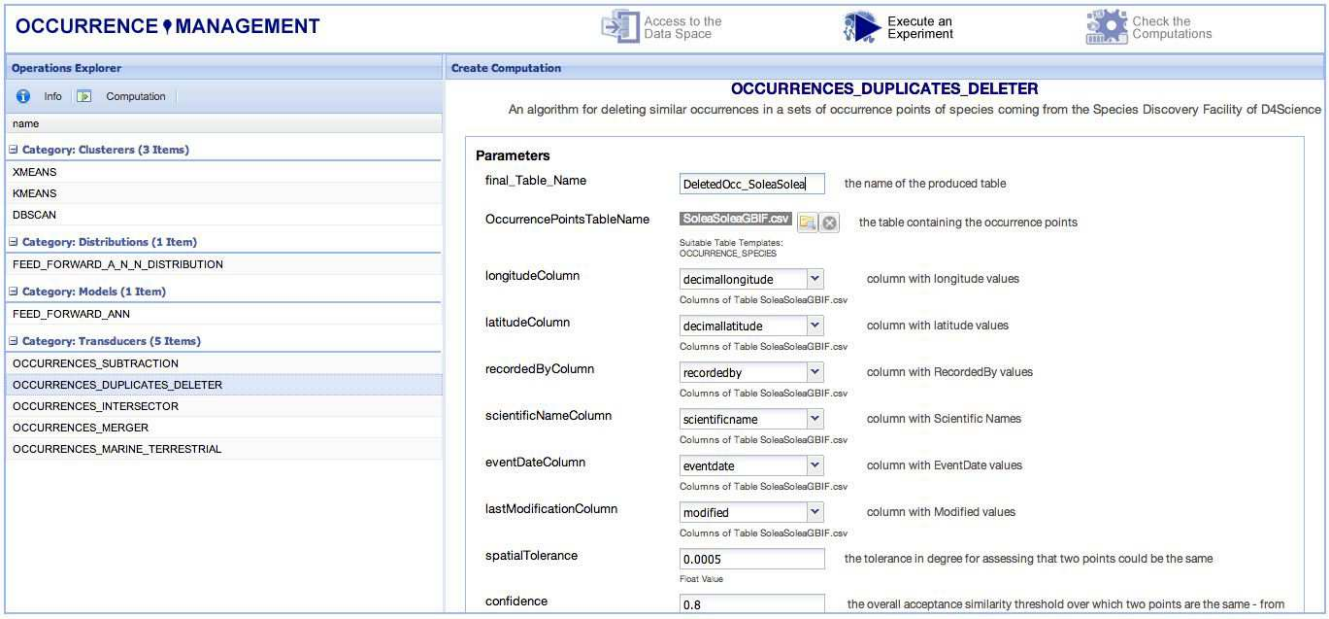

Figure 3. The setup phase of the "Occurrences Duplicates Deleter" procedure. On the left side a set of procedures is highlighted which can be applied to occurrence records.

fies 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 data preparation facility 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 ongoing projects dealing with species data: the i-Marine project [31] focusing on marine species and the EUBrazilOpenBio project [32] 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.

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

The infrastructure has been implemented in a such a way that the available set of facilities can be easily extended. In particular, for what concerns the class of those that have been described in this paper plans have already been made to improved them. 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. The ODM facility will be strengthened by exploiting the distributed computing capabilities offered by the D4Science infrastructure. This is justified by the fact that datasets comparison activities are computation intensive tasks when dealing with huge datasets and when serving hundred of users concurrently. On the analysis side, there are facilities for using more sophisticated techniques like occurrence clustering and anomaly points detection. Algorithms like DBScan [33] and KMeans [34] can be used to coordinates dimensions in order to assess the points density and to identify possible spatial outliers. 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).

References

[1] Jim Gray, Alexander S. Szalay, Ani R. Thakar, Christopher Stoughton, and Jan Vandenberg. Online scientific data curation, publication, and archiving. Technical Report MSR-TR-2002-74, Microsoft Research, July 2002.

[2] Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[3] Geoffrey Boulton, Philip Campbell, Brian Collins, Peter Elias, Dame Wendy Hall, Graeme Laurie, Onora O'Neill, Michael Rawlins, Dame Janet Thornton, Patrick Vallance, and Mark Walport. Science as an open enterprise. Technical report, The Royal Society, June 2012.

[4] Christine L. Borgman. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, pages 1–40, 2011.

[5] O. J. Reichman, Matthew B. Jones, and Mark P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331:703–705, 2011.

[6] William K. Michener and Matthew B. Jones. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, 27(2):85–93, 2012.

[7] M. J. Costello. Motivating online publication of data. *BioScience*, 59(5):418–427, 2009.

[8] James L. Edwards, Meredith A. Lane, and Ebbe S. Nielsen. Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, 289(5488):2312–2314, 2000.

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

[10] Heather Constable, Robert Guralnick, John Wieczorek, Carol Spencer, A. Townsend Peterson, and The VertNet Steering Committee. Vertnet: A new model for biodiversity data sharing. *PLoS Biol*, 8(2):e1000309, 02 2010.

[11] Andrew C Jones, Richard J White, and Ewen R Orme. Identifying and relating biological concepts in the catalogue of life. *Journal of Biomedical Semantics*, 2(7), 2011.

[12] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato De Giovanni Tim Robertson, and David Vieglais. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. *PLoS ONE*, 7(1), 2012.

[13] TDWG. Access to Biological Collections Data - ABCD, 2005. Version 2.06.

[14] TDWG. TAPIR - TDWG Access Protocol for Information Retrieval, 2010. Version 1.0.

[15] Tim Clark, Sean Martin, and Ted Liefeld. Globally distributed object identification for biological knowledgebases. *Briefings in Bioinformatics*, 5(1):59–70, 2004.

[16] Anthony Goddard, Nathan Wilson, Phil Cryer, and Grant Yamashita. Data hosting infrastructure for primary biodiversity data. *BMC Bioinformatics*, 12(Suppl 5):S5, 2011.

[17] Global Biodiversity Information Facility. GBIF Strategic Plan 2012-2016: Seizing the Future, 2011.

[18] Michael Diepenbroek, Hannes Grobe, Manfred Reinke, Uwe Schindler, Reiner Schlitzer, Rainer Sieger, and Gerold Wefer. Pangaea – an information system for environmental sciences. *Computers & Geosciences*, 28(10):1201 – 1210, 2002.

[19] William Michener, Dave Vieglais, Todd Vision, John Kunze, Patricia Cruse, and Greg Janée. DataONE: Data Observation Network for Earth – Preserving Data and Enabling Innovation in the Biological and Environmental Sciences. *D-Lib Magazine*, 17(1/2), 2011.

[20] Walter Jetz, Jana M. McPherson, and Robert P. Guralnick. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution*, 27(3):151 – 159, 2012.

[21] D4Science.org. D4Science Hybrid Data Infrastructure, 2012.

[22] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. D4Science: an e-Infrastructure for Supporting Virtual Research Environments. In Maristella Agosti, Floriana Esposito, and Costantino Thanos, editors, *Post-proceedings of the 5th Italian Research Conference on Digital Libraries - IRCDL 2009*, pages 166–169. DELOS: an Association for Digital Libraries, 2009.

[23] Ian Foster, Yong Zhao, Ian Raicu, and Shiyong Lu. Cloud Computing and Grid Computing 360-Degree Compared. In *Grid Computing Environments Workshop, 2008. GCE '08*, pages 1–10, 2008.

[24] Dave Roberts and Tom Moritz. A framework for publishing primary biodiversity data. *BMC Bioinformatics*, 12:I1, 2011.

[25] A. Townsend Peterson, Sandra Knapp, Robert Guralnick, Jorge Soberón, and Mark T. Holder. The big questions for biodiversity informatics. *Systematics and Biodiversity*, 8(2):159–168, 2010.

[26] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Making Virtual Research Environments in the Cloud a Reality: the gCube Approach. *ERCIM News*, (83):32–33, October 2010.

[27] Leonardo Candela, Donatella Castelli, and Pasquale Pagano. Virtual research environments: an overview and a research agenda. *CODATA Data Science Journal*, 12:GRDI75–GRDI81, 2013.

[28] Gio Wiederhold. Mediators in the Architecture of Future Information Systems. *Computer*, 25(3):38–49, 1992.

[29] S.A. True. Planning the future of the World Geodetic System 1984. In *Position Location and Navigation Symposium, 2004. PLANS 2004*, pages 639 – 648, April 2004.

[30] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[31] I-Marine. The i-Marine European Project, 2011.

[32] EUBrazilOpenBio. EUBrazilOpenBio European Project: the EUBrazilOpenBio web portal, 2012.

[33] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. 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*, pages 226–231. AAAI Press, 1996.

[34] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.