Biodiversity Informatics, 6, 2009, pp. 1-4

DARWIN CORE BASED DATA STREAMLINING WITH DIGIMUS 2.0

KAKODKAR A. P., KERKAR S. S., VARGHESE N. S., KAVLEKAR D. P. & C.T. ACHUTHANKUTTY

Bioinformatics Centre, National Institute of Oceanography,
Council of Scientific and Industrial research (CSIR), Dona Paula, Goa 403 004, India

Abstract.— Cataloguing biological specimen is an important activity of biological museums world over. Software developed especially for this purpose has evolved over time to achieve more accuracy in retrieving data from large and diverse datasets. Combining smaller datasets into a larger information system requires uniformity of data based on a single data standard. In the developing world smaller datasets are maintained by individual researchers or small college and university groups. To standardize data from such datasets software needs to be developed, requiring expertise and sufficient funds which are often unavailable. We present a simple open source web based tool developed using PHP to enable an individual, with little or no knowledge of information systems or databases, to effectively streamline specimen data with data standard Darwin Core 1.2 (DwC 1.2). Such data can then be shared and easily provided to data aggregators like Ocean Biogeographic Information Systems (OBIS - http://www.iobis.org) and Global Biodiversity Information Facility (GBIF - http://www.gbif.org). This tool can be accessed at http://www.niobioinformatics.in/digimus.php and its source code is freely available at http://www.niobioinformatics.in/digimus_source.php.

Key words.—biological specimen, Darwin Core, data standards, data streamlining, information exchange.

Darwin Core data standards facilitate robust information exchange between distributed datasets (Costello & Berghe 2006; Sautter et. al., 2007). Data retrieval tools such as DIGIR server have been meticulously designed for compatibility with Darwin Core data standards (Hobern, 2002; Greene, 2007; DiGIR, 2008). Integration of DIGIR server with Darwin Core enables querying diverse databases with the same Darwin Core schema, but independent of their location, data content or methods of development (Wieczorek, 2007). However, there aren't any accurate estimates for the number of collections adopting Darwin Core Data Standards. A current estimate, by GBIF enumerates 1683 collections from 226 providers (GBIF, 2008), which may not include Darwin Core datasets which are not data providers to GBIF.

Current scope of data standards (e.g. GBIF); do not consider small datasets which are with individuals and institutions. Most of these efforts take place in developed countries. Individual researchers from developing countries possess valuable data which is mostly not shared with global datasets (Schnase et. al., 2003). This data is unavailable to the global research community (Edwards, 2000). Until data standards such as Darwin Core are implemented by researchers from developing world the information in their possession may prove difficult to access.

Funding for biological data management in developing countries is much less or entirely nonexistent (Thomson, 2005). To solve this problem, a software tool needs to be developed that requires the user to possess little or no knowledge of web programming and database design. In this paper we present a web based open source software tool named "DigiMus 2.0" which enables individuals to manage biological specimen data from their collection and effectively structure it according to Darwin Core data standards v 1.2 (DwC v 1.2).

DIGIMUS 2.0 DEVELOPMENT

DigiMus 2.0 presents a web interface allowing a user to enter biological specimen data into a database. The database is structured to comply with Darwin Core 1.2 (DwC 1.2) data standards (Wouter, 2008). Data entered by individual users can be downloaded, transferred to another database, or shared with larger data aggregators (e.g. OBIS and GBIF) that employ a DIGIR server to query Darwin Core compliant datasets. As DigiMus 2.0

Correspondence email: *anaikkakodkar@nio.org*

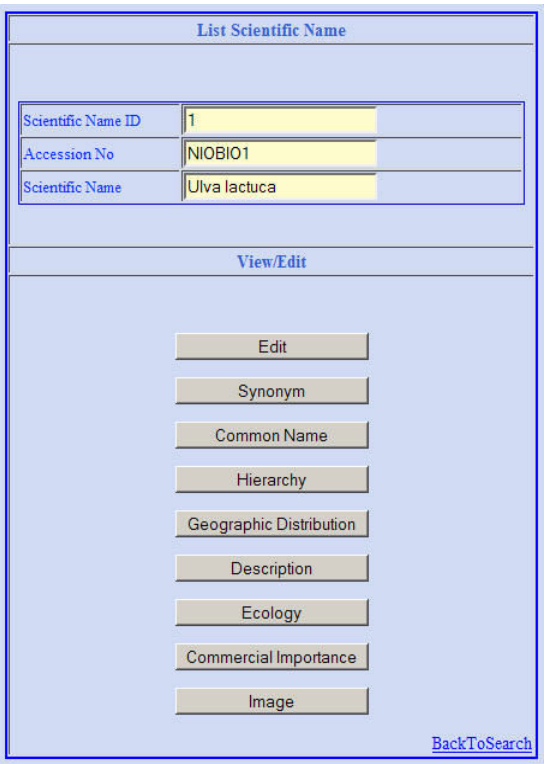

Figure 1. Interface for DigiMus 2.0. Page displayslinks to various data management modules.

was developed mainly for marine biological specimens, we have used OBIS version of Darwin Core 1.2, to enable better information exchange between an individual dataset and OBIS.

A PHP[1] script written for DigiMus 2.0 connects to a MySQL[2] database and inserts user entered data. The script arranges the data into the tables designed to comply with DwC 1.2 schema. When the database is queried through the web interface, the PHP script runs a MySQL query and fetches results to the browser window in HTML.

In addition to DwC 1.2 compliant dataset, DigiMus 2.0 also allows users to enter additional specimen metadata (e.g., specimen description, ecological details and commercial importance), these tables are non DwC 1.2 compliant (see figure 1). HTML data is presented in the form of tables. A data download link below each table allows the users to download their data to a Comma Separated Values (CSV) file. Data from this CSV file can then be imported into a MS Excel worksheet or a local MySQL database. DigiMus 2.0 also proves a good choice for online user data backup. A user cannot directly delete any data from the main database server but can request the system administrator to do it instead, if data privacy needs to be protected. Custom unique record identifiers can be added by users in the "accession number" field. This is particularly useful in cases where users have predefined unique record identifiers.

DigiMus 2.0 has two web interfaces (i) Data Management Interface (DMI), It consists of a HTML data submission form, that enables a user to enter text and upload multimedia files. Various modules are available to the user to manage various types of information related to a single species (e.g., biogeography, taxonomy, ecology, description, commercial importance etc.). (ii) Record Display Interface (RDI), this is a output from the server in the form of HTML, it presents data in the web browser which enables users to view or download data entered by them through the data management interface, through the browser window (see figure 2). A minimum set of data fields are required to successfully complete a data record. These are programmatically implemented and are prompted within each module if minimum data fields are not met. These include scientific name (bionomial name, author and year) and basic taxonomy (from kingdom to family).

Evaluation of DigiMus 2.0 was carried out using data from seaweed herbarium collection of National Institute of Oceanography, Goa. A total of 729 seaweed herbarium records and data from other sources were manually entered using the web interface of data management interface. A demo web interface using HTML and CSS was developed to display records. The demo can be accessed here[3]

DISCUSSION

The objective behind developing DigiMus 2.0 was to create an easy to use interface for marine biologists to manage their biological specimen data and streamline it with current data standards such as Darwin Core 1.2 and we found that DigiMus 2.0 can create a downloadable data file concurrent with DwC 1.2 data standards, which can be uploaded to larger aggregators such as OBIS (Grassle, 2000) and GBIF (Lane, 2003). DigiMus 2.0 can prove an important tool for the marine biologists in the

[1] http://www.php.net.
[2] http://dev.mysql.com.
[3] http://www.niobioinformatics.in/digimus_demo.php

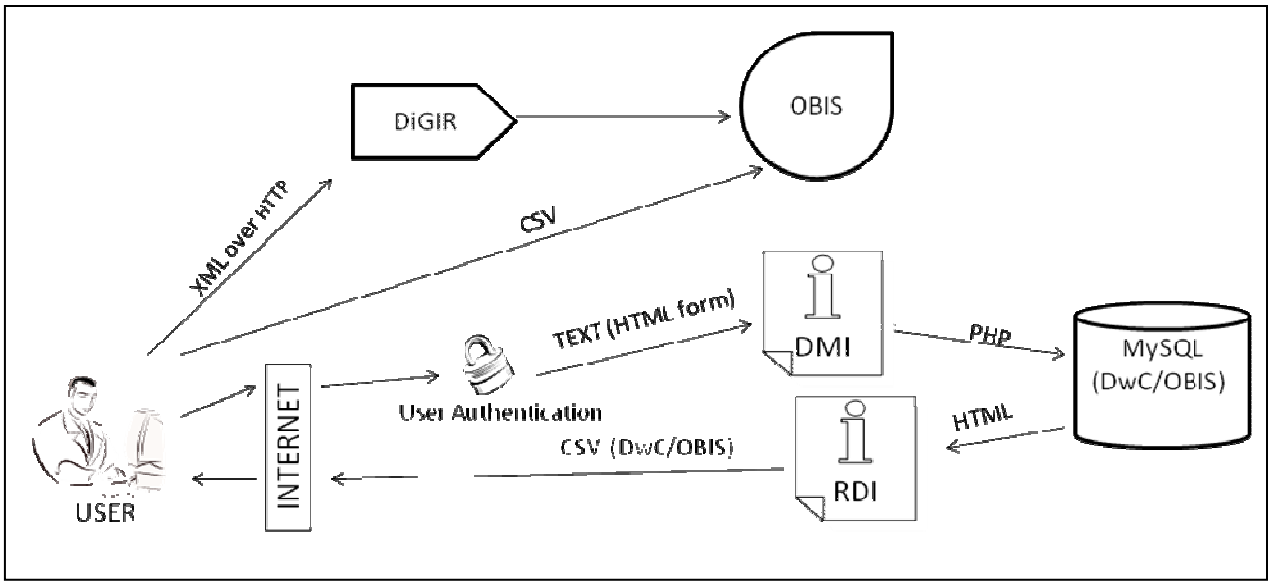

Figure 2. Representation of data flow through DigiMus 2.0. Data Management Interface (DMI) is a HTML form through which an user enters data. Data is processed by a SQL query and stored in the MySQL database. Record Display Interface (RDI), displays query response in a HTML table, which is downloadable. A user can share download data with OBIS as a text file (CSV) or through DiGIR server.

developing countries, as it will enable them to manage data according to current data standards and without spending more time in developing software tools for data management or learning web programming. As a web based tool DigiMus 2.0, does not require specific hardware or software other than a personal computer with internet connectivity and a web browser. Such minimal requirements make it an ideal tool for individual researchers and small research groups in developing countries with limited data management budgets.

Web based biological specimen digitization software has always been developed to target a small number of data managers who have been specifically trained to have good knowledge of data management and information system concept. DigiMus 2.0 does not require the user to undergo any specific training. In fact it has been specifically designed to be user friendly. An undergraduate student or a researcher with basic knowledge of using a web browser can use DigiMus 2.0 with ease. Being in the development phase DigiMus 2.0 has some weaknesses. On the user end the functioning of DigiMus 2.0 depends upon the internet connection speed of the client computer. Low internet connection speed results in decreased number of managed data records. On the lower internet speeds uploading large image or multimedia files takes more time. At the server side, data storage can prove challenging with an increase in number of users. Managing separate databases for individual user at the server side is another concern. These are some challenges which need to be addressed in future versions.

DigiMus 2.0 exists as an open source tool that will ease marine biological data sharing. It prepares unorganized data into a standard format. Being an open source software and licensed under GPL, its source code can be modified for specific needs of a particular research group. Hopefully, DigiMus 2.0 will encourage marine biologists from developing countries to be a part of the larger worldwide data management culture and it will also motivate them to share biological specimen data. Online availability of data will increase its usability at wider levels which will be in the interest of the global biological community.

CONCLUSION

Streamlining biological specimen data, with current data standards increases the potential of data in terms of information exchange from distributed datasets to global initiatives such as OBIS and GBIF. DigiMus 2.0 is an effort to present a web based open source software tool for compilation of data to allow compatibility with Darwin Core data standards. We hope that this

tool shall be a handy utility to individual researchers and small research groups from developing countries for managing their biological specimen data.

ACKNOWLEDGEMENTS

The authors thank Department of Biotechnology (Govt. Of India), New Delhi for financial support through the BTISnet programme. This is contribution No- 4458 of NIO.

REFERENCES

Chapman, A. D. 2005. Principles of data Quality. Global Biodiversity Information Facility. 1-58.

Costello, M. J. and E. V. Berghe. 2006. 'Ocean biodiversity informatics': a new era in marinebiology research and management. Marine Ecological Progress Series. 316: 203-214.

DiGIR, 2008. Distributed Generic Information Retrieval. Accessed. June 2, 2008.[4]

Edwards, J. L., Lane, M. A., and E. S. Nielsen. 2000. Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. Science. 289: 2312-2314.

GBIF. 2008. Global Biodiversity Information Facility. Accessed. June 19, 2008.[5]

Grassle, J. F. 2000 The Ocean Biogeographic Information System (OBIS):an on-line, worldwide atlas foraccessing, modeling andmapping marine biological datain a multidimensionalgeographic context. Oceanography. 13: 5-7.

Greene, S. L., Minoura, T. Steiner, J. J. and G. Pentacost. 2007. WebGRMS: Prototype software for web-based mapping of biological collections. Biodiversity Conservation 16: 2611–2625

Hobern, D. 2002. Integrating Biodiversity Data Standards and Interoperability . Accessed. March 4, 2008.[6]

Lane, M. A. 2003 The Global Biodiversity Information Facility. Bulletin of the American Society for Information Science and Technology. Oct/Nov: 22-24.

Wieczorek, J. 2007. Darwin Core Wiki Site. Accessed. March 3, 2008.[7]

Wouter, A. 2008. Darwin Core Versions. Accessed. March 3, 2008.[8]

Schnase, J. L., Cushing, J., Frame, M., Frondorf, A., Landis, E., Maier, D., and A. Silberschatz. 2003. Information technology challenges of biodiversity and ecosystem informatics. Information Systems 28: 339-345.

Sautter, G., Bohm, K. and D. Agosti. 2007. A quantitative comparison of xml schemas for

Taxonomic publications. Biodiversity Informatics. 4: 1-13.

Thomson, K. S. 2005. Natural History Museum Collections in the 21st Century. Accessed March 4, 2008.[9]

[4] http://digir.net/.

[5] http://www.gbif.org/.

[6] http://www.cria.org.br/eventos/tdbi/bis/presentations/bis_dhobern.ppt.

[7] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome.

[8] http://wiki.tdwg.org/twiki/bin/view/DarwinCore/DarwinCoreVersions.

[9] http://www.actionbioscience.org/evolution/thomson.html.