Proceedings of the 1st Central European Diatom Meeting 2007
Kusber, W.-H. & Jahn, R. (ed.)
Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin
ISBN 978-3-921800-63-8, © BGBM, Berlin 2007.
doi:10.3372/cediatom.109 (available via http://dx.doi.org/)

Common access to distributed biodiversity information

Anton Güntsch, Wolf-Henning Kusber, Markus Döring, Pepe Ciardelli & Walter G. Berendsohn

Botanic Garden and Botanical Museum Berlin-Dahlem, Dept. of Biodiversity Informatics and Laboratories, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany; biodiversityinformatics@bgbm.org

INTRODUCTION

Over the last decade, several international networks have been implemented to give common access to distributed primary biodiversity data holdings, such as collection and observational data. Significant examples include: the SpeciesAnalyst network (Vieglas 1999), focussing primarily on North American Natural History Collections; the European Natural History Specimen Information Network ENHSIN (Berendsohn 2003, Güntsch 2003); and the Biological Collection Access Service for Europe (Berendsohn 2002, BioCASE 2007a, b), networking diverse European biological collections. Unfortunately, a wide range of different network architectures, access protocols, and data standards is in use, posing a major impediment to the development of a common data portal for all biodiversity data repositories world-wide. With the advent of the Global Biodiversity Information Facility (GBIF 2007), this obstacle has been largely overcome by agreeing upon a common set of protocols DiGIR (DiGIR 2005) and BioCASE (BioCASE 2007a) and data standards for collection information (ABCD and Darwin Core). This article gives an introduction to the BioCASE protocol and to how standard software components can be used to link data sources to international networks with relative ease. An example using the AlgaTerra project demonstrates that BioCASE technology can be used to give access to collection data ranging from text information to multimedia files.

THE BIOCASE PROTOCOL

Unified access to data sources using different data models and database technologies requires a query mechanism which wraps the local database query system of the individual data holder with a common access layer. The BioCASE network addressed this requirement by developing two XML-based specifications to form a solid language for sending and receiving commonly understood messages across the collection network (Döring & Güntsch 2003).

The ABCD XML-schema (Access to Biological Collection Data, see Berendsohn 2005) provides a comprehensive definition of terms associated with diverse types of biological collections and their corresponding meanings. If an ABCD-compliant network component uses the term *ISO3166Code*, for example, all other components “know” that this refers to a 2-, 3-, or 4-letter code following the ISO3166 standard for representing names of country of origin. ABCD covers both the unit level (e.g. specimens and observations) and the collection level, including IPR and copyright issues, and in its current version 2.06 comprises around 1000 element definitions.

Apart from the definition of terms and their meanings, the network defines a set of valid queries and responses. The BioCASe protocol provides an XML-based specification of three basic query types which can be sent to every BioCASe compliant data provider:

Capabilities: a capability request returns a list of the data definitions for which this provider has been configured (e.g. ABCD versions), as well as the set of network data elements which have been configured. Typically, the capabilities request is used to inform client software components about the properties of a BioCASe provider installation.

Scan: a scan request returns the set of distinct values for a given term. For example, a scan for the term “ISO3166Code” will return all country codes used by the provider. Typically, the scan operation is used to build data indices in centralized portals.

Search: a search request is the actual query for data and can be composed of multiple nested comparison operations and Boolean expressions (see Fig. 1).

```
<request>
 <header>
  [...]
  <type>search</type>
 </header>
 <search>
  <requestFormat>http://www.tdwg.org/schemas/abcd/1.2</requestFormat>
  <responseFormat start="0" limit="5">http://www.tdwg.org/schemas/abcd/1.2</responseFormat>
  <filter>
   <and>
    <like path="/DataSets/DataSet/[...]/TaxonIdentified/NameAuthorYearString">Navicula*</like>
     <or>
      <like path="/DataSets/[...]/TaxonIdentified/HigherTaxa/HigherTaxon">Bacillariophy*</like>
      <and>
       <equals path="/DataSets/DataSet/[...]/GatheringSite/Country/CountryName">Germany</equals>
       <greaterThan path="/DataSets/DataSet/[...]/ISODateTimeBegin">2002-04</greaterThan>
      </and>
     </or>
    </and>
   </filter>
  <count>false</count>
 </search>
</request>
```

Fig. 1. BioCASE compliant search request.

Several networks have been built using this relatively simple query language, from smaller thematic systems (e.g. GBIF-D “German Flora”, see Kirchhoff et al. 2007) to large-scale international networks providing simultaneous access to hundreds of collection databases (e.g. BioCASE 2007a).

HOW TO BECOME A BIOCASE PROVIDER

As a result, programming know-how is no longer necessary to connect a data source to one or more data networks. A provider software package from the BioCASE initiative can be installed on a web server and configured to map attributes in the local collection database to data elements used in the network (see Fig. 2).

Since configuring such mappings in text files is still an obstacle for less technically adept users, the provider software comes with a configuration tool that walks the user through all necessary installation steps with a graphic user interface. Personal support

(email, phone) is available from the European helpdesk for BioCASE providers (support@biocase.org).

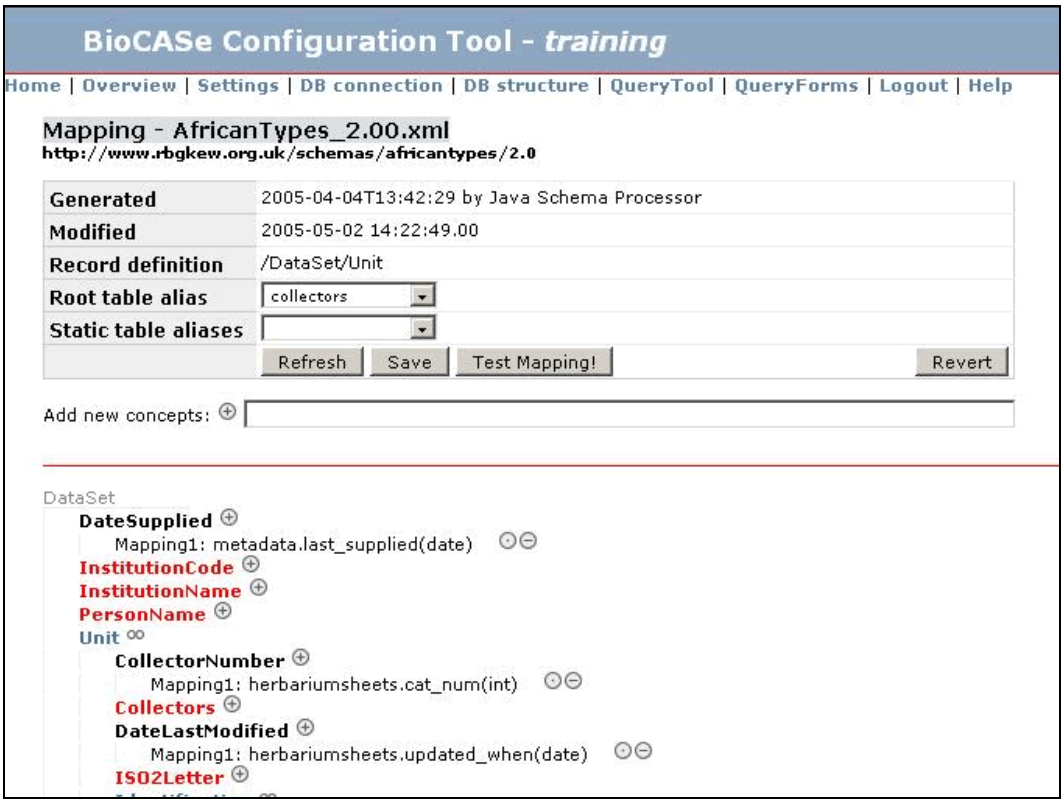

Fig. 2: BioCASe configuration tool.

THE ALGATERRA EXAMPLE

The AlgaTerra Information System on Algae (Jahn 2001, Jahn et al. 2004, Jahn & Kusber 2007) is an excellent example of the flexibility of both the BioCASe protocol and the ABCD schema. In 2005, the type section of the Algaterra database was linked to GBIF using the BioCASe provider toolkit. Type specimen label information is now available via the global GBIF data portal and through the European BioCASE search interface.

In the same year, AlgaTerra decided to publish movies of micro-algae via the international data networks (Fig. 3).

Fig. 3. AlgaTerra movie of *Cymatopleura librile* (Ehrenb.) Pant. (copyright Oliver Skibbe).

This was made possible by simply modifying the BioCASE provider configuration so that the requested object's ABCD record now includes the corresponding movie's URL on the Botanic Garden and Botanical Museum Berlin-Dahlem's web server.

This example demonstrates that the very same technology and installation can be used to transmit very different types of content, ranging from simple XML-encoded text information to videos, high resolution images, and sound files. Today, more than 14 million specimens and observational records have been linked using BioCASE technology, and this number continues to grow rapidly. By integrating the massive amounts of data stored in formerly disparate and scattered data repositories, we hope to create a new homogeneous information space and give birth to a new generation of scientific applications.

ACKNOWLEDGEMENTS

The development of access systems for specimen and observational data as well as support for data providers is sustained in the framework of the EU 6th-framework project SYNTHESYS (RII3-CT-2003-506117).

REFERENCES

Berendsohn, W. G. (ed.) 2005: ABCD Schema - Task Group on Access to Biological Collection Data [online]. – Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. [cited 2007-09-30]. Available from <http://www.bgbm.org/TDWG/CODATA/default.htm>.

Berendsohn, W. G. 2003: ENHSIN in the context of the evolving global biological collection information system. – Pp. 21-32 in: Scoble, M. (ed.): ENHSIN – The European Natural History Specimen Information Network. The Natural History Museum. – London.

Berendsohn, W. G. 2002: BioCASE - A Biological Collection Access Service for Europe. Alliance News 29(6): 6-7.

BioCASE 2007a: Biological Collection Access Service [online] BioCASE Secretary, Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. [cited 2007-09-30]. Available from http://www.biocase.org/.

BioCASE 2007b: What is BioCASE [online] BioCASE Secretary, Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. [cited 2007-09-30]. Available from <http://www.biocase.org/whats_biocase/unit_net.shtml>.

DiGIR 2005: Distributed Generic Information Retrieval (DiGIR) [online]. [cited 2007-09-30]. Available from <http://digir.sourceforge.net/>.

Döring, M. & Güntsch, A. 2003: Technical introduction to the BioCASE software modules. 19th annual meeting of the Taxonomic Databases Working Group (TDWG 2003), Oeiras, Lisbon 2003, Abstract.

GBIF 2007 GBIF Data Portal [online]. – Global Biodiversity Information Facilty [cited 2007-09-30]. Available from <http://www.gbif.org/>.

Güntsch, A. 2003: The ENHSIN Pilot Network. – Pp. 33-40 in: Scoble, M. (ed.): ENHSIN – The European Natural History Specimen Information Network. The Natural History Museum. – London.

Jahn, R. 2001: AlgaTerra - an information system for terrestrial algal biodiversity: a synthesis of taxonomic, molecular and ecological information. – Pp. 230-231 in: Anonymous (ed.): BIOLOG - German Programme on Biodiversity and Global Change. Status Report 2001. – Bonn.

Jahn, R. & Kusber, W.-H. (ed.) 2007: AlgaTerra Information System [online]. Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin. [cited 2007-09-30 September]. Available from <http://www.algaterra.org>.

Jahn, R., Kusber, W.-H., Medlin, L. K., Crawford, R. M., Lazarus, D., Friedl, T., Hepperle, D., Beszteri, B., Hamann, K., Hinz, F., Strieben, S., Huck, V., Kasten, J., Jobst, A., Glück, K. 2004: Taxonomic, molecular and ecological information on diatoms: The information system AlgaTerra. – Pp. 121-128 in: Poulin, M. (ed.): Seventeenth International Diatom Symposium 2002. – Bristol.

Kirchhoff, A., Hahn, A., Holetschek, J., Kelbert, P., Jahn, R. & Berendsohn, W. G. 2007: Open Access to Biodiversity Collection Data – GBIF Germany and the Botanical Node. – Pp. 79-82 in: Kusber, W.-H. & Jahn, R. (ed.): Proceedings of the 1st Central-European Diatom Meeting 2007. – Berlin. [CrossRef]

Vieglas, D. 1999: Integrating disparate biodiversity resources using the information retrieval standard. Z39.50. – TDWG 1999 Abstracts. – Cambridge, USA.