Das DNA-Bank-Netzwerk – Eine Struktur für alle Fälle?

Gabriele Dröge, Holger Zetzsche, Birgit Gemeinholzer
Biodiversitätsinformatik und Labore
Botanischer Garten und Botanisches Museum Berlin-Dahlem
Königin-Luise-Str. 6-8, 14195 Berlin
{g.droege, h.zetzsche, b.gemeinholzer}@bgbm.org

The DNA Bank Network comprises four DNA banks of major biological research collections in Germany. A DNA bank is a technically optimized service facility for the storage of documented DNA as well as a new type of collection. Here we present the structure of the network's database system including all specially created, modified and applied software components. The shared web portal facilitates and visualises DNA data and specimen information of all available DNA samples. That includes a reference to the organism from which DNA was extracted and which had to be deposited in a scientific collection. Additionally, links to inferred molecular data are given if those are published in sequence databases. Unique identifiers are used to connect single DNA datasets with specimen data and links to digital multimedia units. Wrapper software as for the GBIF portal was applied to visualise data directly from multiple data sources. The DNA module was newly designed to record and manage DNA data. The module is as well made to set online references to specimen databases and molecular data sets. Numerous specimen databases of all relevant types can be linked to the DNA module. Furthermore, a DNA extension of ABCD schema (BioCASE Provider Software) was designed to enable a transfer of DNA data online via wrapper. The presented database architecture is appropriate to deal with any kinds of specimen and DNA databases compatible with GBIF. Therefore the network's webportal holds the potential to become a central internet platform for biological DNA banks.

1 Einleitung

DNA-Banken sind eine neue Form biologischer Sammlungen. Sie stellen spezialisierte Serviceeinrichtungen für die naturwissenschaftliche Forschung dar, die eine dauerhafte Lagerung von gut dokumentiertem DNA-haltigen Gewebe sowie von DNA-Proben ermöglichen. Derzeit werden weltweit zahlreiche biologische DNA-Banken aufgebaut, die ihren Sammlungsschwerpunkt meist auf bestimmten Organismengruppen (z.B. Arabidopsis Stock Center[1], Kew Gardens DNA Bank[2] für pflanzliche DNA, San Francisco Zoo DNA Bank[3] für tierische DNA und Centraalbureau voor Schimmelcultures[4] für pilzliche DNA) oder einen geographisch begrenzten Fokus haben (z.B. DNA Bank Kirstenborsch[5] für pflanzliche DNA aus Südafrika oder am Center for Plant Conservation Genetics[6] in Australien). Diese DNA-Banken unterscheiden sich außerdem in der Anzahl der vorgehaltenen DNA-Proben, in Umfang, Qualität und Art der angebotenen Informationen sowie der Möglichkeit die Proben zu bestellen voneinander. Aufgrund der Vielzahl an DNA-Banken mit jeweils eigenem Webauftritt und beschränktem Probenumfang ist es zeitlich sehr aufwändig, gut dokumentierte DNA für ein bestimmtes Taxon zu finden und zu beziehen.

Ein wesentliches Problem im Zusammenhang mit biologischen Sammlungsobjekten ist die Überprüfbarkeit molekularer Daten. Molekulare Daten (z.B. DNA-Sequenzen) werden in der Regel in einer der großen Sequenz-Datenbanken (GenBank[7], EMBL[8], DDBJ[9]) mit Angabe der taxonomisch korrekten Identifizierung des zugrunde liegenden Beleges hinterlegt. Ein Verweis zum Ort der Hinterlegung des Beleges sowie der aus ihm gewonnenen genomischen DNA fehlt aber oft oder ist unvollständig. So ist es zeitlich sehr aufwändig, die Informationen von der DNA-Sequenz zur DNA sowie vom angegebenen Taxonnamen zum Beleg zurückzuverfolgen. Eine kritische Überprüfung der Beleginformationen und der entsprechenden molekularen Ergebnisse erfolgt deshalb in der Regel nicht. Die wissenschaftliche Notwendigkeit einer Verifikation ergibt sich aber aus der Tatsache, dass bis zu 20% der Sequenzen in den großen Sequenz-Datenbanken fehlerhaft sind, eine falsche taxonomische Bezeichung haben oder falsch annotiert sind [Br03].

---

[1] http://arabidopsis.info/
[2] http://data.kew.org/dnabank
[3] http://www.sfzoo.com/openrosters/ViewOrgPageLink.asp?LinkKey=14507
[4] http://www.cbs.knaw.nl
[5] http://www.sanbi.org/research/dnabank.htm
[6] http://www.scu.edu.au/research/cpcg/
[7] http://www.ncbi.nlm.nih.gov/Genbank/index.html
[8] http://www.ebi.ac.uk/embl/ (European Molecular Biology Laboratory)
[9] http://www.ddbj.nig.ac.jp/ (DNA Data Bank of Japan)

Fehlende Datentransparenz sowie fehlende Verknüpfungen zwischen Beleg- und Sequenzdaten der waren Gründe, für die Initiierung eines neuen Netzwerks. Seit 2007 bilden die DNA-Banken von vier großen deutschen Forschungssammlungen das DNA-Bank-Netzwerk. Die Projektpartner ergänzen sich dabei in ihren Sammlungsschwerpunkten: Deutsche Sammlung von Mikroorganismen und Zellkulturen in Braunschweig (DSMZ), Zoologisches Forschungsmuseum Alexander Koenig in Bonn (ZFMK), Zoologische Staatssammlung in München (ZSM) sowie Botanisches Museum und Botanischer Garten Berlin-Dahlem (BGBM).

Das Ziel des DNA-Bank-Netzwerks ist es, die Hinterlegung und Verfügbarkeit von Belegen molekular untersuchter Individuen sowie ihrer DNA langfristig sicherzustellen, und damit die auf DNA basierende organismische Forschung effektiver und transparenter zu machen. Die Idee des Netzwerkes besteht darin, alle DNA-Proben der beteiligten Partnerbanken mit vollständiger Dokumentation über eine zentrale Internetplattform zugänglich und abrufbar zu machen. Dazu gehört auch die Verknüpfung publizierter Sequenzdaten mit den Informationen der zugrundeliegenden DNA-Proben. Es soll ebenso möglich sein, auf die DNA-Proben, DNA-Daten sowie auf ihre Belege und Belegdaten als Referenz verweisen zu können. Zudem soll DNA und Gewebe von taxonomisch sicher bestimmten Organismen für die Grundlagenforschung (Phylogenie, Populationsgenetik, Biogeographie) und für angewandte Analysen (z.B. Naturschutzgenetik) über das Netzwerk zu bestellen sein.

Das DNA-Bank-Netzwerk beabsichtigt damit, das fehlende Bindeglied zwischen biologischen Beleg-Datenbanken (z.B. BIODAT[10], Specify Software Project[11], DiversityCollection[12]) und DNA-Sequenz-Datenbanken zu sein und die Überprüfbarkeit molekularer Ergebnisse von der DNA-Sequenz bis zum untersuchten Beleg für alle hinterlegten Proben zu ermöglichen.

Da das DNA-Bank-Netzwerk für weitere Partner und deren Datenbank-Lösungen offen sein soll, musste eine komplexe Datenarchitektur realisiert werden, die im Folgenden vorgestellt wird. Dabei wird die Wahl der verwendeten Datenarchitektur diskutiert und die neu entwickelten Software-Komponenten erläutert.

[10] http://www.biodat.de/
[11] http://specifysoftware.org/
[12] http://www.diversityworkbench.net/Portal/DiversityCollection

2 Informationstechnische Voraussetzungen

2.1 Verwaltung und Bereitstellung von Belegdaten

Biologische Sammlungen werden weltweit mit spezialisierten Sammlungsdatenbanken verwaltet, die unterschiedliche Software verwenden und die entsprechend der Nutzeransprüche konfiguriert sind (z.B. Specify Software Project, DiversityCollection, BIODAT). Externe Belegdaten werden entweder gar nicht oder in Kopie in diesen Systemen erfasst.

Mit GBIF[13] existiert bereits ein zentrales Webportal mit dem annähernd 175 Mio. Beleg- und Beobachtungsdaten von 294 Institutionen weltweit frei und dauerhaft zugänglich gemacht werden [GBIFb]. Die Grundidee dabei ist, dass alle Beleg-Datensätze im Orginal durch die Datenprovider selbst verwaltet, und bei einer Suchabfrage live zu GBIF übertragen werden [Ch05]. Die Einträge der Sammlungsdatenbanken werden mittels sogenannter Wrapper in Standardformate übertragen, wodurch die Suchabfrage und eine einheitliche, strukturierte Darstellung über GBIF erst ermöglicht wird. Für die Datensynchronisation kommen drei Systeme zum Einsatz, die BioCASE Provider Software[14] [DG03], DiGIR[15] und TAPIR[16] [DG05], [Dö06]. Diese drei Wrapper bieten die Möglichkeit, die Tabellenstruktur einer Datenbank mittels Schema Mapping in ein einheitliches Format zu übersetzen (ABCD-Schema[17] bzw. Darwin Core[18]). Die Ausgabe der Daten erfolgt als XML-Datei. Es ist dadurch möglich, Informationen aus ganz verschiedenen Datenquellen und Datenformaten an GBIF anzubinden und dort zur Verfügung zu stellen.

Außer GBIF existieren mehrere Portale im Internet, die sich die Bereitstellung von Beleg- und Beobachtungsdaten zur Aufgabe gemacht haben. Allen ist jedoch gemein, jeweils nur bestimmte taxonomische Gruppen und/oder geographische Gebiete abzudecken (vergleiche Tabelle 1). Die einzige Ausnahme bildet das EDIT-Portal (Specimen and Observation Explorer for Taxonomists) [Zi09]. Über dieses Portal sind ebenso wie bei GBIF Daten von allen Kontinenten und aus allen taxonomischen Großgruppen angebunden. Das EDIT-Portal [Ke08] ist aus dem BioCASE-Portal hervorgegangen. BioCASE steht gleichzeitig für das Portal europäischer Beleg- und Beobachtungsdaten [HD08] sowie für die BioCASE Provider Software. Allerdings beziehen alle BioCASE-Portale ihre Daten ausschließlich aus dem SYNTHESYS-Cache [Ho08]. Dabei handelt es sich um ein Datenmodell, dessen Inhalte (jedoch nicht Struktur) identisch mit dem GBIF-Index sind, der in Kopie als Mirror am BGBM gehostet wird [GBIFb], [HD08].

---

[13] http://www.gbif.org (Global Biodiversity Information Facility)
[14] http://www.biocase.org/products/provider_software/index.shtml (Biological Collection Access Service)
[15] http://digir.sourceforge.net (Distributed Generic Information Retrieval)
[16] http://www.tdwg.org/activities/tapir/ (TDWG Access Protocol for Information Retrieval, TDWG = Taxonomic Databases Working Group)
[17] http://wiki.tdwg.org/twiki/bin/view/ABCD (Access to Biological Collection Data)
[18] http://digir.sourceforge.net/schema/conceptual/darwin/core/2.0/darwincoreWithDiGIRv1.3.xsd

Die BioCASE-Portal und somit auch das EDIT-Portal bieten einige zusätzliche Funktionen (z.B. Ausgabe in 11 Sprachen, Einbindung von taxonomischen Checklisten), die auf dem gegenwärtigen Stand für das DNA-Bank-Netzwerk nur bedingt geeignet sind (Portalsprache ist Englisch, z.Zt. sind überwiegend europäische, d.h. regional eingeschränkte Checklisten verfügbar).

Tabelle 1: Übersicht vorhandener Portale für Beleg- und Beobachtungsdaten

Portal	**Taxonomie**	**Geographie**
Australian's Virtual Herbarium[19]	Pflanzen	Australische Provider
Mammal Networked Information System (MaNIS)[20]	Säugetiere	Nordamerikanische Provider
speciesLink[21]	Alle Gruppen	Brasilianische Belegdaten
European Natural History Specimen Information Network (ENHSIN)[22]	Alle Gruppen	Europäische Provider
Biological Collection Access Service for Europe (BioCASE)[23]	Alle Gruppen	Europäische Belegdaten
Global Biodiversity Information Facility (GBIF)	Alle Gruppen	Alle Kontinente
EDIT Specimen and Observation Explorer for Taxonomists[24]	Alle Gruppen	Alle Kontinente

Wegen seiner Systemunabhängigkeit gegenüber der Software der Datenprovider ist die GBIF/BioCASE-Technologie für die Verwirklichung der angestrebten Verknüpfung von bestehenden Belegdatenbanken mit DNA- und Sequenzdatenbanken derzeit ohne Alternative. Alle anderen in Tabelle 1 aufgeführten Portale nutzen ebenfalls die GBIF-Technologie oder sind selbst Datenprovider von GBIF, besitzen jedoch einen eingeschränkten Fokus.

[19] http://www.chah.gov.au/avh/avh.html
[20] http://manisnet.org/manis/
[21] http://splink.cria.org.br
[22] http://www.bgbm.org/BioDivInf/projects/ENHSIN/
[23] http://search.biocase.org/europe/
[24] http://search.biocase.org/edit/
(EDIT = European Distributed Institute of Taxonomy, http://www.e-taxonomy.eu/)

2.2 Verwaltung und Bereitstellung von DNA-Daten

Im Gegensatz zu den angebotenen Belegdaten ist es nicht möglich, dazugehörige DNA-Daten direkt über GBIF abzurufen. Die GBIF-Index-Datenbank sowie das GBIF-Portal sind derzeit nicht in der Lage, DNA-Daten zu integrieren oder auszugeben. Des Weiteren können DNA-Daten mit den vorhandenen Wrappern nur in stark begrenztem Maße (ABCD) oder überhaupt nicht (DarwinCore) übertragen werden.

DNA-Proben, die aus biologischen Sammlungsobjekten stammen, werden in der Praxis durch folgende Datensysteme verwaltet: Excel-Dateien, individuelle Datenbanklösungen, LIMS[25] oder die Integration der Daten in eine Sammlungsdatenbank (z.B. Specify Software Project, DiversityCollection).

Für die Realisierung der Datenarchitektur des DNA-Bank-Netzwerkes resultieren daraus zwei Probleme. Zum einen ist keines der aufgeführten Verwaltungssysteme für DNA-Daten in der Lage externe Beleg-Datenbanken so anzubinden, dass die Originaldaten live abgerufen werden können. Mit LIMS ist es möglich Beleginformationen zu speichern. DNA-Proben mit Belegdaten aus verschiedenen Sammlungen direkt („live“) zu referenzieren, wäre nur durch eine individuell angepasste Anbindung jeder einzelnen Sammlungsdatenbank aus der die Belege stammen machbar. Die benötigte Flexibilität der angestrebten Netzwerklösung kann damit nicht erreicht werden.

Software-Lösungen wie das Specify Software Project oder DiversityCollection haben zudem den Nachteil, dass nur wenige DNA-spezifische Attribute eingegeben werden können. Diese genügen den Ansprüchen des DNA-Bank-Netzwerkes nach hoher Transparenz vor allem gegenüber den Kunden nicht. Innerhalb des Projektes wurde vereinbart, 30 verschiedene DNA-spezifische Informationen zu erheben, wovon einige als Pflichtangaben deklariert wurden (z.B. Relation zum Beleg, Extraktionsmethode). Diese Software-Lösungen sind zudem ebenfalls nicht in der Lage externe Sammlungsdatenbanken live anzubinden.

3 Anforderungen an ein Datenbank-System für das DNA-Bank-Netzwerk

Das Hauptziel des DNA-Bank-Netzwerks ist ein web-basiertes Referenzsystem mit dem eine umfangreiche Dokumentation der DNA-Proben umgesetzt und die Informationskette zwischen DNA-Sequenzdaten und Beleg in beide Richtungen geschlossen werden kann. Unter Berücksichtigung der vorhandenen Probleme wurde dafür eine modulare Datenarchitektur vorgeschlagen und umgesetzt, die auf folgenden Annahmen beruht.

---

[25] http://www.lims.de/ (Labor-Informations-Management-Systeme)

Alle Daten zu einem Sammlungsobjekt sollen nur einmal erfasst werden. Auf sie kann für verschiedene Anwendungen oder abgeleitete Sammlungen bei Bedarf direkt zugegriffen werden. Der Zeitaufwand und die Anzahl von Eingabefehlern kann bei einmaliger Erfassung der Daten minimiert werden. Zudem müssen die Daten nur einmal korrigiert oder ergänzt werden, z.B. bei der Revision eines Taxonnamens oder wenn Annotationen zu einem Beleg hinzugefügt werden müssen. Kopien von Datensätzen könnten verwendet werden, wenn die Datensätze statisch bleiben. Das ist bei Belegdaten und insbesondere bei Taxondaten aber nicht der Fall.

Eine modulare Datenbankarchitektur, bei der in einer Datenbank jeweils alle Sammlungsobjekte eines Sammlungstyps (z.B. Belege, DNA etc.) einer Institution verwaltet werden, wird als vorteilhaft angesehen [ZDG08]. Demgegenüber ist die Verwendung einer gemeinsamen Datenbank für alle Sammlungen oder zahlreiche separate, personalisierte Datenbanken (eine Datenbank pro Wissenschaftler) innerhalb einer Institution, wie sie in der Praxis Verwendung finden, langfristig kaum zu handhaben. Eine gemeinsame Sammlungsdatenbank kann nicht immer für neue Anwendungen erweitert werden. Mehrere separate Datenbanken, die zudem ggf. verschiedene Software verwenden und/oder unterschiedlich konfiguriert sind, sind für Webanwendungen nur über die erwähnten Wrapper mit vertretbarem Aufwand zu erschließen.

Entsprechend wurde für das vorliegende Problem ein System entwickelt werden, bei dem die Belegdaten in den jeweiligen Sammlungsdatenbanken und die DNA-Daten in davon getrennten DNA-Datenbanken erfasst werden können. In einem zu entwickelnden Webportal sollten dann alle verfügbaren Daten vom Beleg, über die DNA bis hin zu Verweisen auf publizierte Sequenzdaten, Publikationen etc. zusammenlaufen und aus den entsprechenden Datenbanken direkt angezeigt werden. Die Anbindung der DNA-Daten an das Webportal sollte sich dabei an das oben erläuterte Konzept von GBIF anlehnen.

4 Datenarchitektur des DNA-Bank-Netzwerks

Für das DNA-Bank-Netzwerk wurde ein modular aufgebautes Datenbank-System entwickelt. Es besteht aus drei Ebenen: Beleg-Datenbanken, DNA-Datenbanken/DNA-Modul und Webindex/Webportal (vgl. Abb. 1).

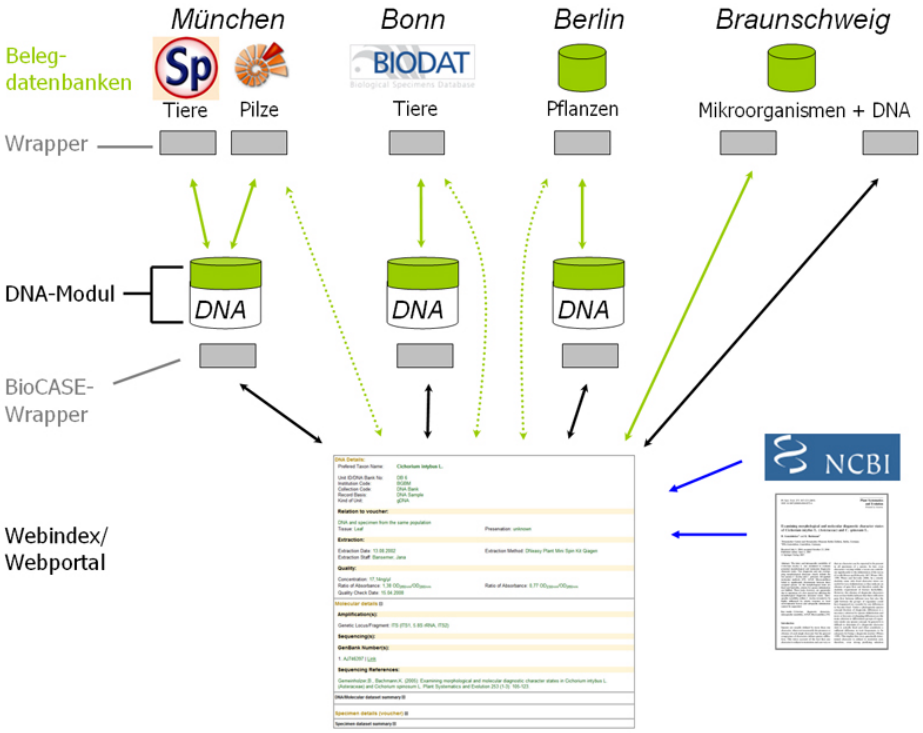

Abbildung 1: Datenflüsse im DNA-Bank-Netzwerk

Belegdaten (grüne Linien) werden über BioCASE- oder DiGIR-Wrapper (graue Boxen) zum DNA-Modul sowie zum Webportal/Webindex übertragen. Die DNA-Daten (schwarze Linien) werden zum Webindex/Webportal über einen zweiten BioCASE-Wrapper übertragen. Im Webportal werden alle Beleg- und DNA-Daten sowie Referenzen zu weiteren, molekularen Daten oder Publikationen zu einer DNA-Probe (blaue Linien) gebündelt und sind bei einer Suchabfrage auf einer Seite gemeinsam abrufbar [Dr08].

Alle verfügbaren Daten für Sammlungsbelege und DNA-Proben werden pro Partnerinstitution in zwei Datenbanken verwaltet, die Belege mit Informationen wie Fundort, geographische Koordinaten, Funddatum, Sammler, wissenschaftlicher Name etc. einerseits (→ Beleg-Datenbank) und die Proben mit genomischer DNA mit den Angaben zur Extraktionsmethode, verschiedenen Qualitätsparametern, PCR-Amplifikationen etc. andererseits (→ DNA-Datenbank). Die Verknüpfung der DNA-Daten mit den dazugehörenden Belegdaten erfolgt über die Wrapper, die auch GBIF nutzt. Dabei werden zur Identifizierung eines Beleges drei Attribute verwendet, die über das Mapping definiert werden: UnitID, SourceID und SourceInstitutionID (=Triple ID). Zusammen mit der Wrapper-URL sind die einzelnen Belegdatensätze damit eindeutig als Referenz nutzbar.

Zudem wird über Links von der DNA-Probe auf alle abgeleiteten Informationen verwiesen, z.B. zu Abstracts oder Vollversionen relevanter Publikationen oder über die GenBank-Akzessions-Nummer zu den Sequenzdaten. Bei der GenBank-Akzessions-Nummer handelt es sich eine eindeutige Eingangsnummer, die in den Sequenzdatenbanken GenBank, EMBL und DDBJ gleichermaßen verwendet wird.

Eine gemeinsame Verwaltung von Beleg- und DNA-Daten in einer Datenbank bleibt bei dieser Datenarchitektur möglich und wird bei der DSMZ in Braunschweig angewendet. Da die DSMZ auf die Hinterlegung von lebendem Typusmaterial von Mikroorganismen und Zellkulturen spezialisiert ist, stammt der überwiegende Teil der DNA-Proben von eigenen Klonkulturen. Die wenigen Datensätze zu externem Beleg-Material werden an der DSMZ als Kopie eingegeben. Zu Beginn des Netzwerkprojektes wurden die für alle Partner notwendigen DNA-spezifischen Parameter diskutiert, die seitdem für jede DNA-Probe erfasst werden. Die Datenbank der DSMZ wurde so erweitert, dass mit ihr diese 30 DNA-Attribute verwaltet werden können. Für die anderen Partner kam diese Lösung nicht in Frage, da dort die Arbeit mit externen Belegen eher die Regel als die Ausnahme darstellt.

4.1 Beleg-Datenbanken

Alle vier Projektpartner sind GBIF-Knoten und verfügen über eigene große Sammlungen und Datenbanken, in denen die Informationen zu Belegen jeweils lokal verwaltet werden. Aus diesen Datenbanken kann das DNA-Bank-Netzwerk via Wrapper die benötigten Beleginformationen abrufen und mit den DNA-Daten über die oben genannte Triple ID und die Wrapper-URL verknüpfen.

Über dieselben Wrapper werden die Belegdaten auch im Webportal des DNA-Bank-Netzwerks angezeigt, wenn alle verfügbaren Daten einer bestimmten DNA-Probe abgerufen und zusammengestellt werden.

4.2 DNA-Datenbanken

Drei der vier Projektpartner verfügten zu Beginn des Projektes über keine Möglichkeit, DNA-Daten mit einer eigenen Datenbank eingeben und verwalten zu können. Einzig am BGBM existierte ein Microsoft Access-Frontend zur Eingabe einiger DNA-Parameter[26], die jedoch den Ansprüchen des Netzwerkes nicht genügten. Zudem konnten mit dieser Software keine Belegdaten aus externen Datenbanken direkt angebunden werden.

[26] http://www.bgbm.org/bgbm/research/dna/

Um eine möglichst große Flexibilität im Hinblick auf die beteiligten und mögliche zukünftiger Partner zu erreichen, wurde das DNA-Modul entwickelt. Dabei handelt es sich um eine MySQL-Datenbank und PHP-Skripte. Mit dem Modul können die DNA-Informationen aller DNA-Proben einer DNA-Bank lokal erfasst, verwaltet und mit den entsprechenden Beleginformationen verknüpft werden. An das Modul lassen sich beliebig viele Sammlungsdatenbanken über Wrapper anbinden.

Mit dem DNA-Modul können außerdem Kundenbestellungen sowie die Lagerung der DNA-Proben verwaltet und Links zu Publikationen und GenBank-Akzessions-Nummernen gesetzt werden, in den die DNA-Proben verwendet wurden. Die GenBank-Akzessions-Nummern beziehen sich auf publizierte Sequenzdaten, die bei EMBL, GenBank oder DDBJ hinterlegt sind. Nach diesen Nummern kann im Webportal ebenso wie im DNA-Modul gesucht werden, wodurch die Verknüpfung zwischen Sequenz und Beleg in beiden Richtungen sichergestellt ist.

Um die DNA-Daten an den Webindex/das Webportal zu übertragen, wird auch auf die DNA-Datenbank ein BioCASE-Wrapper aufgesetzt. Die dazu benötigte DNA-Erweiterung des ABCD-Schemas wird im nächsten Kapitel vorgestellt.

Um auch Belegdaten erfassen zu können, die über keine Datenbank verfügbar sind, die über GBIF erreichbar ist, bzw. die keine Wrapper-Installation aufweist, wurde mit der Entwicklung eines „Offline Specimen Tools“ begonnen. Damit können die wichtigsten Parameter zum Beleg aufgenommen und verwaltet werden (geographische und Standortangaben, Taxonname, bis zu vier Mehrfachbestimmungen, Ort der Aufbewahrung, digitale Bilder vom Beleg und/oder vom Fundort). Sofern Belegdaten aus dem Offline Specimen Tool übertragen werden sollen, muss ein zweites Mapping mit ABCD 2.06 vorgenommen werden. Damit können diese Belegdaten auch an GBIF übermittelt werden.

Des Weiteren wird ein Konfigurationstool entwickelt, mit dem es z.B. auch möglich sein soll, andere Datenbanksysteme als MySQL für die DNA-Datenerfassung zu verwenden. Für die Nutzer soll die Möglichkeit bestehen, das für sie am besten geeignete System anwenden zu können.

Das DNA-Modul kann sowohl lokal installiert werden, als auch über externe Server (Webapplikation) von autorisierten Nutzern bedient werden. Es ist als Open-Source-Produkt entwickelt worden und wird ab 2011 auf der Webseite des Netzwerks zur Verfügung stehen.

4.3 DNA-Erweiterung des ABCD-Schemas

Die BioCASE Provider Software benutzt das ABCD-Schema, um die individuelle Struktur einer Datenbank in eine allgemeine, GBIF-kompatible Struktur zu übersetzen. Diese übersetzten Strukturen ermöglichen effiziente Suchabfragen, da individuelle Abfragebefehle für die lokalen Datenbanken entfallen.

Das ABCD-Schema enthält über 1000 spezifische Felder für die Übertragung ganz unterschiedlicher biologischer Sammlungsdaten[27] [Be05]. So können z.B. die Informationen zur Bestimmung eines Beleges (Taxoninformationen) ebenso übermittelt werden wie Aufsammlungsdaten (Fundort, geographische Koordinaten, Standort, Sammler, Sammeldatum, etc.) oder die Informationen zu Ort und Art der Hinterlegung eines Beleges in einer wissenschaftlichen Sammlung. Ferner gibt es Erweiterungen des ABCD-Schemas für die Übertragung weiterer fachspezifischer Daten, z.B. ABCDEFG[28] für geowissenschaftliche Informationen [Ki05].

Der DNA-Teil im aktuellen ABCD 2.06-Schema[29] war nicht ausreichend, um alle für das DNA-Bank-Netzwerk relevanten DNA-Inhalte zu übertragen. Er umfasst lediglich die GenBank-Akzessions-Nummer, die Sequenzier-Methode, den Ausführenden der Sequenzierung, die Sequenzlänge, die Information, ob es sich um DNA, RNA oder ein Protein handelt sowie eine Referenzangabe.

Im Rahmen des DNA-Bank-Projektes wurde daher eine ABCD-Erweiterung für DNA-Daten (ABCDDNA) mit 76 spezifischen Elementen entwickelt. Das Schema ist auf der Webseite des Netzwerks als HTML- und XML-Datei kommentiert und verfügbar.[30]

Mit dieser Erweiterung ist die Übertragung von DNA-spezifischen Datenbankeinträgen wie der Extraktionsmethode, DNA-Konzentration und verschiedenen Parametern der DNA-Qualität, Primern und Erfolg von PCR-Amplifikationen, Informationen zu Klonierung, Sequenziermethode, Sequenzierprimer, Elektropherogramme aller ermittelten DNA-Sequenzen, Konsensus-Sequenzen und DNA-Barcodes sowie Links zu GenBank-Einträgen und Publikationen möglich.

Der DNA-Datensatz wird durch die Angabe der Triple ID und gegebenenfalls der Wrapper-URL ebenso wie der Beleg-Datensatz eindeutig referenzierbar. Als assoziierter Datensatz (UnitAssociation[31]) wird mit demselben Schema u.a. auch die Referenz zum Beleg übertragen (Triple ID des Beleges).

Die Verwendung von DiGIR/DarwinCore statt BioCASE/ABCD kam nicht in Frage, da DarwinCore keine DNA-Erweiterung anbietet und nicht hierarchisch aufgebaut ist. D.h. mehrere GenBank-Akzessions-Nummern anzugeben, die mit ein und derselben DNA-Probe in Verbindung stehen, wäre damit nicht möglich. Das ist aber unerlässlich, da aus einer DNA-Probe zahlreiche und im Extremfall hunderte oder tausende DNA-Sequenzen bzw. GenBank-Einträge gewonnen werden können.

---

[27] http://www.bgbm.org/TDWG/CODATA/Schema/ABCD_2.06/HTML/ABCD_2.06.html
[28] http://www.geocase.eu/ (Access to Biological Collection Data Extended For Geosciences)
[29] http://www.bgbm.org/TDWG/CODATA/Schema/ABCD_2.06/HTML/ABCD_2.06.html#complexType_Sequence_Link031A1EA0
[30] www.dnabank-network.org/schemas/ABCDDNA/DNA.html, www.dnabank-network.org/schemas/ABCDDNA/DNA.xml
[31] http://www.dnabank-network.org/schemas/ABCDDNA/ABCDDNA.html#element_UnitAssociation_Link07C06E60

Ebenso wie die Wrapper-Software wurde die DNA-Erweiterung als Open-source-Produkt entwickelt und kann von der Netzwerk-Webseite heruntergeladen werden[32]. Die Seite enthält eine detaillierte Anleitung zur Installation und Benutzung des ABCDDNA-Schemas[33].

4.4 Webportal und Webindex

Das Webportal des DNA-Bank-Netzwerks wurde als zentrale Plattform für die DNA-Banken aller derzeitigen und möglichen zukünftigen Partner entwickelt. Damit können alle DNA-Proben der Partnerbanken und alle verfügbaren Daten, die sich auf die einzelnen Proben beziehen, gebündelt dargestellt werden. Der Webindex ist eine MySQL-Datenbank und enthält die suchbaren Beleg- und DNA-Daten sowie die dazugehörigen GenBank-Akzessions-Nummern in Kopie. Der Import der Daten in den Webindex erfolgt mit Hilfe einer Routine, wobei über die DNA-Wrapper zuerst alle vorhandenen Datensätze gezählt und unter ihrer eindeutigen Nummer abgespeichert (UnitID der DNA-Probe) werden. Danach werden die benötigten Daten inklusive der Triple ID des zugrundeliegenden Beleges schrittweise in den Webindex übertragen. Nach den Importen folgen diverse Routinen, die z.B. das Sammeljahr aus der Datumsangabe und die Meere aus den Fundortangaben herausfiltern.

Bei der Detailansicht im Portal werden zwei Anfragen verschickt. Zum einen werden die aktuellen DNA-Daten aus der jeweiligen DNA-Datenbank abgerufen und angezeigt, zum Zweiten die dazugehörigen Belegdaten aus der entsprechenden Beleg-Datenbank (vgl. Abbildung 1). Für beide Anfragen werden die eindeutigen Parameter zur Identifizierung verwendet (Triple ID des DNA- und des Beleg-Datensatzes). Die Anfragen resultieren in jeweils einer XML-Datei, die mittels XSLT in ein HTML-Dokument umgewandelt und im Webportal auf einer gemeinsamen Seite dargestellt werden.

Um auch von den Beleg-Daten auf die DNA-Daten verweisen zu können, kann in den jeweiligen Beleg-Datenbanken ein Verweis auf das Webportal generiert werden, der alle verfügbaren DNA-Proben zu einem bestimmten Beleg zeigt. Dabei muss die Triple ID des Beleg-Datensatzes nach folgendem Muster übergeben werden:

http://www.dnabank-network.org/Query.php?UnitIDS=DSM 14247&
CollCodeS=Prokarya&InstCodeS=DSMZ

Über das Webportal können die DNA-Proben aller Partner außerdem bestellt werden. Zur Abwehr von Spams ist ein Login mit vorheriger Registrierung nötig. Die Freischaltung eines Accounts erfolgt nach manueller Authentifizierung der Adressinformationen der Kunden. Alle verwendeten Passwörter und Aktivierungscodes werden verschlüsselt im Webindex abgelegt.

---

[32] http://www.dnabank-network.org/Downloads.php
[33] http://www.dnabank-network.org/Extension.php

Bei einer (Vor-)Bestellung wird eine entsprechende Email mit den gewünschten DNA-Proben an die betreffenden DNA-Banken verschickt. Die Bestätigung, Bearbeitung und Abrechnung der Bestellungen erfolgt dann in der Verantwortung der Partner. Eine allgemeine Lösung über das Webportal ist hierfür nicht vorgesehen, da die Bearbeitungs- und Abrechnungsmodi der jetzigen und potenziellen Partner zu verschieden sind.

Es existiert zusätzlich die Möglichkeit DNA-Proben und -Daten zeitlich oder generell zu sperren. Sofern sie generell gesperrt werden (z.B. bei geschützten Arten) wird der Datensatz nicht in den Webindex importiert. Bei einer zeitlichen Sperre (z.B. wenn Sequenzdaten noch nicht publiziert sind) ist die DNA-Probe im Webportal einsehbar, jedoch nicht bestellbar.

Das Portal bietet die Möglichkeit, gezielt nach Beleginformationen zu suchen, z.B. nach dem Taxonname einer Probe, Fundortangaben oder Sammeljahr. Des Weiteren ist es möglich, nach einer bestimmten DNA-Qualität zu suchen, die für eine molekulare Analyse mindestens erforderlich ist, die Suche auf Belege zu beschränken für die ein digitalisiertes Bild (Voucher) vorhanden ist, oder nach einer GenBank-Akzessions-Nummer zu suchen, für die DNA verfügbar ist.

In Planung sind die Übertragung weiterer Beleginformationen, z.B. Dubletten in anderen Herbarien oder Populationsaufsammlungen, die Hinterlegung eines taxonomischen Baumes sowie die Übertragung molekularer Rohdaten, d.h. von Elektropherogrammen, Einzelsequenzen etc. Letztere sind für Molekularbiologen äußerst wertvoll, weil sie eine Einschätzung der Qualität von DNA-Konsensus-Sequenzen erlauben. Über die DNA-Sequenzdatenbanken werden diese Rohdaten derzeit nicht angeboten.

4.5 Nutzung der Strukturen

In den DNA-Banken des Netzwerkes werden derzeit ca. 13.900 DNA-Proben gelagert. (BGBM 4500, DSMZ 4000, ZFMK 1400, ZSM 4000). In die DNA-Banken des Netzwerkes wurden auch ca. 4.200 DNA-Proben aktueller Forschungsprojekte von Wissenschaftlern der Partnerinstitutionen integriert.

Die Datenbank des Webportals wurde im April 2009 mit 9.239 DNA-Proben und 5.243 Taxa online gestellt. Nachdem die Mitteilung im Internet publik gemacht hatte, gab es einen sprunghaften Anstieg der Besuche von 2300 im März auf über 3100 im April. Danach pegelte sich die Anzahl der Besuche bei ca. 700 pro Woche ein. Es werden, wie erwartet und beabsichtigt, Taxa aus dem gesamten Organismenreich gesucht, von Pilzen, Diatomeen (Kieselalgen), Gefäßpflanzen und Wirbellosen bis hin zu Primaten (Menschenaffen, inklusive Mensch). Bis Juni 2009 hatten sich bereits 12 Kunden angemeldet. Am BGBM sind bis dato beispielsweise doppelt so viele Bestellungen über das Webportal eingegangen wie in den zwei Jahren zuvor.

4.6 Zusammenfassung

Die Datenarchitektur des DNA-Bank-Netzwerks ist so angelegt, dass keine spezifischen Lösungen für einzelne Partnerinstitutionen entwickelt werden mussten und vorhandene Strukturen genutzt werden konnten (z.B. Wrapper-Technologie von GBIF). Mit dem DNA-Modul wurde eine flexibel einsetzbare Software zur Erfassung und Verwaltung von DNA-Daten entwickelt. Das Modul richtet sich vor allem an Institutionen und Wissenschaftler, die eine eigene DNA-Bank inklusive Datenbank neu einrichten möchten. Ein Umstieg auf das DNA-Modul ist ebenso möglich wie die Anbindung einer individuellen DNA-Datenbank, wie es z.B. für die DSMZ realisiert wurde.

Das ABCDDNA-Schema wurde entwickelt, um eine ähnlich flexible Plattform wie das GBIF-Portal speziell für DNA-Daten etablieren zu können. Insbesondere dieses Schema ermöglicht es, später auf einfache Weise weitere DNA-Banken an das Portal anzubinden. Diese können entweder das DNA-Modul oder eigene DNA-Datenbank-Lösungen verwenden. Eine wesentliche Voraussetzung für zukünftige Partner ist eine „GBIF-fähige“ Sammlungsdatenbank (Wrapper-Installation). Da nicht alle Institutionen mit Interesse an Anbindung an das DNA-Bank-Webportal über eine solche Beleg-Datenbank verfügen, wird zudem ein Offline-Specimen-Tool entwickelt, das Teil des DNA-Moduls ist. Sofern es zukünftig auch für DiGIR/DarwinCore eine vergleichbare DNA-Erweiterung wie nun für ABCD geben sollte, kann auch diese in die Datenarchitektur des Netzwerkes integriert werden.

Das Webportal des DNA-Bank-Netzwerkes birgt somit das Potenzial, eine internationale Plattform für biologische DNA-Banken zu werden, ähnlich dem GBIF-Portal, das das führende internationale Webportal für biologische Beleg- und Beobachtungsdaten ist.

Die Datenarchitektur des DNA-Bank-Netzwerks erweist sich bis zum jetzigen Zeitpunkt als eine äußerst flexible Struktur die in der Lage ist, ganz unterschiedliche Datenbanklösungen integrieren zu können. Über das Webportal des Netzwerks können beliebig viele DNA-Datenbanken angebunden werden, über das DNA-Modul beliebig viele Beleg-Datenbanken. Es sind Verweise zu allen abgeleiteten Informationen möglich (z.B. Publikationen, Bilder). Insbesondere durch das ABCDDNA-Schema ist die Übertragung komplexer Datensätze möglich.

5 Danksagung

Die Autoren bedanken sich bei der DFG für die finanzielle Förderung des Projektes, bei den Web-Administratoren des ZFMK, der ZSM, der DSMZ und des BGBM für die Unterstützung bei der Anbindung der Datenbanken sowie bei Wolf-Henning Kusber für kritische Durchsicht des Manuskripts.

Literaturverzeichnis

[Be05] Berendsohn, W.G. et al.: ABCD - the proposed standard XML schema for access to biological collection data. P. 3 in Berendsohn, W.G.; Rissoné, A. (ed.): Taxonomic Databases Working Group, 2005 Annual Meeting, Abstracts, St. Petersburg, Berlin, London.

[Br03] Bridge, P.D. et al.: On the unreliability of published DNA sequences. New Phytologist 2003; S. 43 - 48.

[Ch05] Chapman, A.D.: Uses of primary species-occurrence data. The Global Biodiversity Information Facility. Kopenhagen, 2005.

[DG03] Döring, M.; Güntsch, A.: Technical introduction to the BioCASE software modules. 19th annual meeting of the Taxonomic Databases Working Group (TDWG), Abstract, Lissabon, Portugal, 2003.

[DG05] Döring, M. et al.: The integration of DiGIR and BioCASe, 20th TDWG meeting, Christchurch, New Zealand, October 2006.

[Dö06] Döring, M.: Using TAPIR in biodiversity networks. 22nd TDWG meeting, St. Louis, USA, October 2006.

[Dr08] Dröge, G. et. al.: The DNA Bank Network - How GBIF technology enables a new generation of communicating DNA repositories. 24th annual meeting of the Taxonomic Databases Working Group (TDWG), Abstract, Perth, Australien, 2008.

[GBIFa] The Global Biodiversity Information Facility: GBIF to Establish Mirror Sites on 3 Continents. http://www.gbif.org/Stories/STORY1113468321, Copenhagen, Denmark, 2005.

[GBIFb] The Global Biodiversity Information Facility: GBIF Data Portal. http://data.gbif.org, Copenhagen, Denmark, 2009.

[HD08] Holetschek, J.; Döring, M.: Publishing Specimen & Observation Records Using BioCASe Technology. In: Gradstein S. R.; Klatt S.; Normann, F.; Weigelt, P.; Willmann, R.; Wilson, R. (eds): Systematics 2008 Programme and Abstracts, Göttingen 7-11 April 2008. Universitätsverlag Göttingen, Göttingen.

[Ho08] Holetschek, J. et al.: The SYNTHSYS Specimen and Observation Portal. eBiosphere Conference, London 2009 (im Druck).

[Ke08] Kelbert, P. et al.: The new EDIT Specimen and Observation Explorer for Taxonomists. In: Weitzman, A.L., Belbin, L. (eds.). Proceedings of TDWG (2008), Fremantle, Australia.

[Ki05] Kiessling, W. et. al.: ABCDEFG – a draft Extension for Geosciences to the ABCD XML schema. 21st annual meeting of the Taxonomic Databases Working Group (TDWG), Abstract, St. Petersburg, Russland, 2005.

[ZDG08] Zetzsche, H.; Dröge, G.; Gemeinholzer, B.: Die Etablierung eines DNA-Bank-Netzwerkes in Deutschland. Aufbau des Netzwerkes und Management. In: Osnabrücker Naturwissenschaftliche Mitteilungen 33/34; Tagungsband: Botanische Gärten gestalten Zukunft – Umweltkommunikation, Artenschutz und Genetische Ressourcen, 2008.

[Zi09] Zippel, E. et al.: EDIT Specimen and Observation Explorer for Taxonomists. Eine Nützliche Komponente der taxonomischen EDIT-Arbeitsplattform im Internet. Pp. 18-21 in GfBS-Newsletter 21/2009, Dresden.