Deutsche Gesellschaft für Limnologie (DGL)
Erweiterte Zusammenfassungen der Jahrestagung 2013 (Potsdam-Berlin), Hardegsen 2014

Netzverfügbarkeit primärer Biodiversitätsdaten: Schritt für Schritt zur BioCASe-Anbindung limnologischer Datenquellen an das GBIF-Netzwerk

Wolf-Henning Kusber[1]*, Sabine von Mering*[1] *& Jörg Holetschek*[1]

[1] Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, w.h.kusber@bgbm.org, s.vonmering@bgbm.org, j.holetschek@bgbm.org

Keywords: BioCASe, Biodiversität, Datenverfügbarkeit, Global Biodiversity Information Facility, Observationsdaten

Einleitung

Eine der Kernaufgaben des *Consortium of European Taxonomic Facilities,* einem Verbund bedeutender europäischer Wissenschaftsinstitutionen (CETAF o. J.) ist die Veröffentlichung biologischer Sammlungsdaten. Die biodiversitätsinformatischen Erfordernisse von CETAF werden durch BioCASe (*The Biological Collection Access Service for Europe*) abgedeckt. BioCASe wurde als transnationales Netzwerk aufgebaut, um einen einheitlichen Internet-gestützten Informationszugang zu biologischen Sammlungen in Europa zu ermöglichen. BioCASe umfasst unterschiedliche Produkte, die diesen Zielen dienen und für Beleg-, Beobachtungs- und Ex-Situ-Daten genutzt werden (BioCASe 2013).

Für die Internetverfügbarkeit digitaler Datensammlungen wurde die BioCASe Provider Software entwickelt, die zwischen Biodiversitäts-Datenbanken und Netzwerken vermittelt und eine Implementierung des BioCASe-Protokolls zum Datenaustausch darstellt. Letzteres regelt, wie XML-Abfragen und Antworten zu Biodiversitätsdaten im Internet übermittelt werden. Grundlage dazu sind ABCD-Konzepte (*Access to Biological Collection Data*) des ABCD-Schemas (Holetschek et al. 2012). Für die Endnutzer entscheidend sind BioCASe-Portale, die für die Suche und Ausgabe von BioCASe-Datensätzen optimiert sind.

Das BioCASe-Netzwerk ist gegenwärtig ein bedeutender Partner für verschiedene Biodiversitätsdaten-Netzwerke. So wird beispielsweise ein Großteil der BioCASe-Datensammlungen an die *Global Biodiversity Information Facility* (GBIF), eine internationale Initiative zur Förderung von freiem Zugang zu primären Biodiversitätsdaten über das Internet, geliefert (GBIF 2013). Anfang 2014 wurden mehr als 430 Millionen Datensätze aller Organismengruppen für Forschung, Naturschutz und Bildung über das globale GBIF-Netzwerk zur Verfügung gestellt. Ausrichtung und Ziele von GBIF sowie Teilnahme von Staaten und nichtstaatlichen Organisationen werden durch ein *Memorandum of Understanding* geregelt (GBIF 2010), Regeln der Datenbereitstellung und -nutzung durch ein GBIF *Data Sharing Agreement* bzw. ein GBIF *Data Use Agreement* festgelegt (GBIF 2013).

GBIF-D (http://www.gbif.de) als nationaler GBIF-Knoten fördert die Anbindung komplexer Daten mithilfe von BioCASe und den Aufbau von *Special Interest*-Netzwerken (GBIF-D 2014). Das

ABCD-Datenschema ist vor allem für die Abbildung von vorliegenden Mehrfachbestimmungen zu einem Datensatz oder komplexen Beziehungen zwischen einzelnen Datensätzen zwingend erforderlich. Beispiele für letzteres sind z. B. Verknüpfungen zwischen Wirt und Parasit, aber auch zwischen Belegdaten und aus den Organismen gewonnenen DNA-Proben (DNA-Bank-Netzwerk, siehe Dröge et al. 2013).

Warum Biodiversitätsdaten im Internet zur Verfügung stellen?

Für alle Biodiversitätsdaten gilt, dass sie über gemeinsame Datenportale im Internet leichter gefunden, wissenschaftlich genutzt und häufiger zitiert werden können. Durch eine verbesserte Datenzugänglichkeit ist eine effektivere Datennachnutzung und Analyse einmal erhobener Daten möglich. Die Sichtbarkeit von Primär- und Metadaten, Projekten und Institutionen wird durch die Bereitstellung in Datennetzwerken deutlich erhöht.

Während sich in Forschungssammlungen weitgehend durchgesetzt hat, Biodiversitätsdaten so umfangreich wie möglich verfügbar zu machen, ist die Biodiversitätsforschung in einer Übergangsphase, hin zur obligatorischen allgemeinen Bereitstellung von Forschungsprimärdaten in öffentlich zugänglichen Datenrepositorien (Berendsohn et al. 2010, Ludwig & Enke 2013). Bei Landesämtern gibt es bezüglich von Biodiversitätsdaten aus dem Monitoring eine zunehmende Akzeptanz für die Bereitstellung in Biodiversitätsnetzwerken. Fragen u. a. der Datenrechte, der Datenqualität und des Umgangs mit den Umweltinformationsgesetzen der Länder bedingen allerdings längerfristige Abstimmungsprozesse.

Im Bereich der Bürgerwissenschaften („Citizen Science“) handelt es sich meist um ehrenamtlich erhobene Biodiversitätsdaten, deren Nutzbarkeit über einschlägige Citizen Science-Portale (z. B. GEO-Tag, Naturgucker u. a.) hinaus, durch Veröffentlichung in internationalen Biodiversitätsnetzwerken deutlich erhöht werden kann.

Wie Biodiversitätsdaten im Internet zur Verfügung stellen?

Wollen Datenbesitzer ihre Biodiversitätsdaten für Biodiversitätsnetzwerke zur Verfügung stellen, gibt es verschiedene Verfahren der technischen Anbindung (GBIF 2011). In den nächsten Abschnitten werden die Schritte vom vorhandenen Datenbestand bis zur Anbindung an das GBIF-Netzwerk mit Hilfe von BioCASe dargestellt. Der Ablauf beginnt mit der Analyse der Datenstruktur und der Datenbereinigung. Die folgenden Schritte sind die Vorbereitung der Metadaten, Installation der BioCASe Provider Software und Datenbereitstellung sowie das Mapping der individuellen Datenbank-Inhalte und die Indexierung bis zur Datenpublikation in Biodiversitätsdatenportalen.

Analyse der Datenstruktur und Datenbereinigung

Die Nutzbarkeit der Daten hängt unter anderem von der Datenstruktur ab. Ein möglichst hoher Grad an Atomisierung der Daten ist vorteilhaft, d. h. eine Speicherung unterschiedlicher Dateninhalte in verschiedenen Feldern. Datenbereinigung umfasst sehr unterschiedliche Felder der Datenbehandlung. Für einfachere Routinekontrollen und -korrekturen eignen sich automatisierte Verfahren, z. B. mit Hilfe von Webservices; komplexere Datenkontrollen sollten dagegen weiterhin von Experten vorgenommen werden (Chapman 2005). Wichtige Aspekte sind die Standardisierung von Daten und die Richtigkeit der Einträge. Fragen der Datenqualität vor der Datenpublikation in Biodiversitätsnetzwerken wurden in Kusber et al. (2012, 2013), von Mering & Kusber (2013) dargestellt und

diskutiert. Komplementär zur Datenkorrektur an der Basis, d. h. beim Datenhalter, sind fachwissenschaftliche Anmerkungen zu publizierten BioCASe-Datensätzen im Internet, die inzwischen in einfacher Weise möglich sind. Der elektronische Annotationsvorgang wird in Tschöpe et al. (2013) dargestellt und dokumentiert.

Vorbereitung der Metadaten

Metadaten geben Auskunft über die administrative und technische Zuständigkeit von Datensammlungen, den Datenbesitzer (bei öffentlich geförderten Datenerhebungen meist Institutionen), Namen und Codes der Institutionen und Datensammlungen, Angaben zu Rechten an den Daten und Lizenzen für Nutzer. Sie enthalten auch eine kurze Beschreibung der Datensammlung, die z. B. über Umfang, Geschichte und Forschungsinteresse der Objektsammlung Auskunft geben kann. Metadaten werden dazu verwendet, eine Datensammlung zu beschreiben, ihre regionale und globale Relevanz zu charakterisieren, auf die Datennutzung Einfluss zu nehmen und auf das Webangebot des Datenproviders bzw. die Originaldatenquelle zu verlinken.

Datenbereitstellung

In der Regel werden Originaldaten auf einem Internet-Server bereitgestellt. Daten, die nicht zur Veröffentlichung geeignet sind, müssen vor der Datenbereitstellung herausgefiltert werden. D. h. die inhaltliche Kontrolle (Datenkuration) liegt in den Händen des Datenhalters. Daher wird selten die Datenbank selbst abgefragt, sondern eine speziell für die Datenweitergabe an Datennetzwerke erstellte Sicht (View). Grundsätzlich sollten die Daten möglichst direkt durch den Datenhalter bereitgestellt werden, um deren Aktualität zu gewährleisten. In Fällen, in denen es technische oder administrative Schwierigkeiten bei der Datenbereitstellung gibt, ist ein Datenhosting durch GBIF-D Knoteninstitutionen (GBIF-D 2014) möglich, allerdings aufgrund der Notwendigkeit der Updates immer nur eine Notlösung.

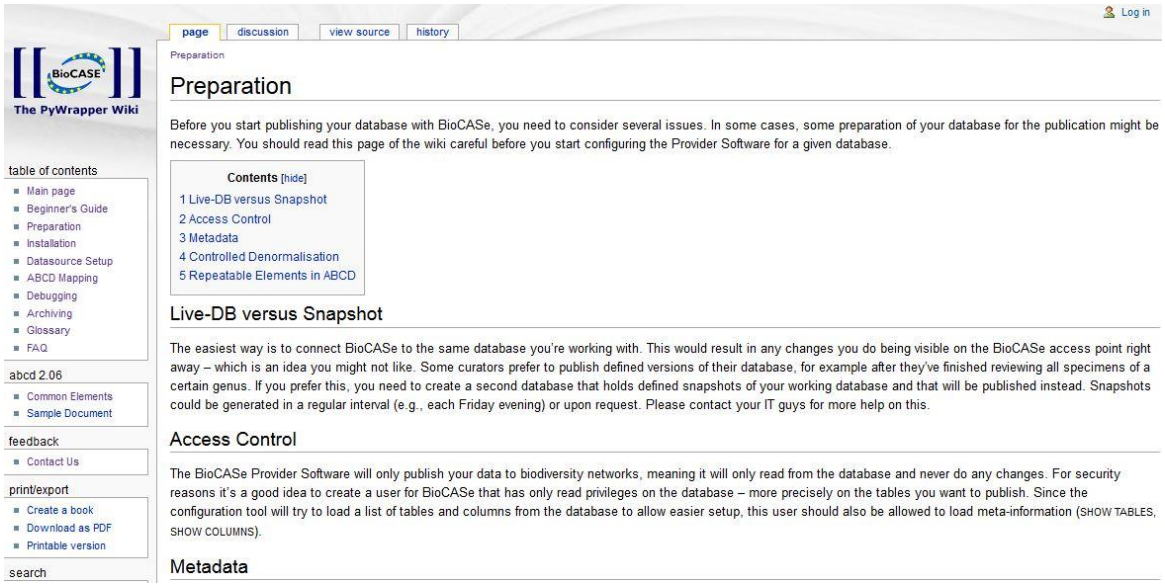

Abb. 1: Wiki zur Dokumentation der BioCASe Provider Software und der Einrichtung von ABCD-Datenquellen (N. N. 2011).

Installation der Provider Software

Um die vorbereiteten Daten im GBIF-Netzwerk zur Verfügung zu stellen, ist die Installation der BioCASe Provider Software notwendig. Umfassende Dokumentation bietet das BioCASe-Wiki (Abb. 1, N. N. 2011); persönliche Hilfestellung leistet der BioCASe-Support.

Mapping

Das Mapping ist die Zuordnung der Datenbankinhalte des Datengebers zu den entsprechenden Feldern des zum Datenaustausch genutzten ABCD Datenstandards. Diese Zuordnung ist mit Hilfe eines einfach zu bedienenden Konfigurationstools der BioCASe Provider Software möglich. Hilfestellung können auch hier BioCASe-Wiki (Abb. 1) und BioCASe-Support leisten.

Registrierung und Indexierung

Durch die Registrierung von Datenquellen bei BioCASe bzw. GBIF wird diese dem jeweiligen Netzwerk bekannt gemacht. Zur schnelleren Suche und besseren Darstellung der angebundenen Daten speichern die Netzwerke häufig verwendete Datenelemente in Cache-Datenbanken zwischen; dieser Vorgang wird Indexierung genannt. Das GBIF-Datenportal (GBIF 2013) beschränkt sich bei der Darstellung auf diese indexierten Informationen, BioCASe-Portale hingegen können alle vom Provider gelieferten Originaldaten sichtbar machen.

Abb. 2: Ansichten des 2013 neu entwickelten Portals der Global Biodiversity Information Facility (GBIF 2013). A. Zugriff auf 433 Millionen Observationen. B. Observationsdaten zu Deutschland (oben), Observationsdaten deutscher Datenanbieter (unten). C. Metadaten-Darstellung am Beispiel einer BioCASe-Datenquelle zu Belegproben am BGBM aus der limnologischen Forschung

Datendarstellung in Biodiversitätsportalen

In verschiedenen Portalen werden im GBIF-Netzwerk publizierte Biodiversitätsprimärdaten ausgegeben und suchbar gemacht. Das GBIF-Datenportal (Abb. 2, http://www.gbif.org) ermöglicht die gemeinsame Suche über alle geographischen Regionen und alle Organismengruppen hinweg. BioCASe-Portale nutzen häufig Teilmengen dieser Daten, um sogenannte *Special Interest Networks* zu bilden. Das 2013 veröffentlichte GBIF-D Algae & Protozoa Datenportal (http://protists.gbif.de)

stellt aktuell 7,6 Millionen Datensätze zu limnischen und marinen Protisten mithilfe von BioCASe-Technologie zur Verfügung (GBIF-D 2013, Kusber et al. 2013).

Danksagung

GBIF-D Pflanzen, Algen & Protisten wird vom Bundesministerium für Bildung und Forschung (BMBF), Projekt LI1001A finanziert. Der deutsche Beitrag für GBIF wird durch BMBF und DFG getragen.

Literatur

Berendsohn, W. G., Chavan, V., Macklin, J. A. (2010): Recommendations of the GBIF Task Group on the Global Strategy and Action Plan for the Mobilisation of Natural History Collections Data. J. Biodiversity Informatics 7: 67-71.

BioCASe (2013): Biological Collection Access Services Homepage. Published on the Internet http://www.biocase.org [25.01.2014].

CETAF (o. J.): CETAF - Consortium of European Taxonomic Facilities. Published on the Internet http://www.biocase.org [03.02.2014]

Chapman, A. D. (2005): Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. 58 S.

Droege, G., Barker, K., Astrin, J., Partels, P., Butler, C., Cantrill, D., Coddington, J., Forest, F., Gemeinholzer, B., Hobern, D., Mackenzie-Dodds, J., Ó Tuama, É., Petersen, G., Sanjur, O., Schindel, D., Seberg, O. (2013): The Global Genome Biodiversity Network (GGBN) Data Portal. Nucleic Acids Research. 42 (D1): D607-D612. DOI:10.1093/nar/gkt928

GBIF (2013): Global Biodiversity Information Facility Published on the Internet http://www.gbif.org [25.01.2014]

GBIF (2011): Getting Started: An overview of data publishing in the GBIF network, (contributed by Remsen, D., Ko, B., Chavan, V., Raymond, M.). Copenhagen: Global Biodiversity Information Facility, 16 S. ISBN: 87-92020-28-3. Published on the Internet http://links.gbif.org/getting_started_publishing_en_v1 [10.02.2014]

GBIF (2010): GBIF Memorandum of Understanding, Copenhagen: Global Biodiversity Information Facility, S., Published on the Internet at http://www.gbif.org/orc/?doc_id=2955 [11.02.2014]

GBIF-D (2014): Global Biodiversity Information Facility – Deutschland. Published on the Internet http://www.gbif.de 11.02.2014]

GBIF-D (2013): GBIF.DE Algae & Protozoa Data Portal. Published on the Internet http://protists.gbif.de [10.02.2014]

Holetschek J., Dröge G., Güntsch A., Berendsohn W.G. (2012): The ABCD of primary biodiversity data access. Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology. 146: 771-779.

Kusber, W.-H., Dröge, G., von Mering, S., Jahn, R. (2012): GBIF-D Pflanzen, Algen & Protisten: Mobilisierung und Publikation primärer Biodiversitätsdaten für die Nutzung in internationalen Datennetzwerken. Deutschen Gesellschaft für Limnologie, Erweiterte Zusammenfassungen der Jahrestagung 2011 (Weihenstephan), Eigenverlag der DGL, Hardegsen: 406-410.

Kusber, W.-H., von Mering, S., Jahn, R. (2013): GBIF-Dateninfrastruktur: Limnologische Beobachtungs- und Belegdaten publizieren, abfragen und analysieren. Deutschen Gesellschaft für Limnologie, Erweiterte Zusammenfassungen der Jahrestagung 2012 (Koblenz), Eigenverlag der DGL, Hardegsen: 485-489.

Ludwig, J. & Enke, H. (Hrsg.) (2013): Leitfaden zum Forschungsdaten-Management. Glückstadt: Verlag Werner Hülsbusch: 118 S.

N.N. (2011): BioCASe Provider Software. Published on the Internet http://wiki.bgbm.org/bps/index.php/Main_Page [25.01.2014].

Tschöpe, O., Macklin, J. A., Morris, R. A., Suhrbier, L., Berendsohn, W. G. (2013): Annotating Biodiversity Data via the Internet. Taxon 62: 1248-1258.

von Mering, S., Kusber, W.-H. (2013): GBIF data network – infrastructure for biodiversity research. Open access to occurrence data of steppe species. Thüringer Ministerium für Landwirtschaft, Forsten, Umwelt und Naturschutz (TMLFUN), Steppenlebensräume Europas - Gefährdung, Erhaltungsmaßnahmen und Schutz, Erfurt: 441-446 S.