Ein Strategiepapier der AG-Biodiversitätsdaten

der Senatskommission für Biodiversitätsforschung der DFG

Zur Sicherung von umweltbezogenen biologischen Daten in Deutschland

September 2011

1. Ziel dieses Papiers

Organismen mit ihren Leistungen und Wechselbeziehungen zueinander sind die Grundlage von Ökosystemen und damit auch des menschlichen Lebens. Die insbesondere in den letzten Jahrzehnten erfolgten Veränderungen der Geosphäre, z.B. durch Klimawandel und veränderte Landnutzung, haben einen dramatischen Verlust an Organismen zur Folge. Es ist sicher, dass dies negative Auswirkungen auf die meisten Ökosysteme hat und weiterhin haben wird, auch wenn viele dieser Effekte wegen der Komplexität biologischer Wechselwirkungen zum gegenwärtigen Zeitpunkt noch nicht (exakt) vorhersagbar sind. Dennoch ist zu erwarten, dass sie einen eminenten wirtschaftlichen und gesellschaftlichen Wandel zur Folge haben werden. Das Erfassen von Veränderungen, ihrer Ursachen und Auswirkungen sowie das Prognostizieren zukünftiger Entwicklungen sind die wichtigsten Fragen der umwelt-bezogenen biologischen Forschung. Dies bedarf jedoch einer breiten Datenbasis und -zugänglichkeit. Die Verwertung der Ergebnisse moderner Forschungsansätze sowie auch der Vielzahl der in klassischen Ansätzen gewonnenen Daten ist bislang durch ihre begrenzte Verfügbarkeit limitiert, obwohl in Teilbereichen der taxonomischen und paläontologischen Forschung, z.B. naturhistorischen Sammlungen in Deutschland, sehr große Kompetenz im Bereich Datensicherung, Nutzung und Vernetzung auf nationaler und internationaler Ebene vorhanden ist (z.B. Global Biodiversity Information Facility, GBIF). Eine auf Ökosysteme zielende biologische Forschung ist Fächer- und Disziplin-übergreifend. Daher müssen die vielen verschiedenartigen Forschungsdaten unter Verwendung entsprechender informationstechnischer und fachlicher Standards gesichert, zusammengeführt und verfügbar gemacht werden. Dazu ist es notwendig, die bereits vorhandene wissenschaftliche Infrastruktur (Projekte, Geräte, Experimentierflächen, Sammlungen) auf der Ebene der Datenverarbeitung zu vernetzen. Für diejenigen Projekte der umweltbezogenen biologischen Forschung, die keine längerfristige Finanzierung haben, erfordert die

dauerhafte Datensicherung eine besondere Struktur, die geschaffen werden muss. Durch gezielte Fördermassnahmen zur optimierten Verfügbarkeit solcher Primärdaten[1] kann das Potential der umweltbezogenen biologischen Forschung in Deutschland effizient gesteigert werden, da viele hochkarätige, vielzitierte Publikationen auf Metaanalysen beruhen.

Ziel dieses Papiers ist es, die gegenwärtig existierenden Probleme bei der Datenspeicherung und –verarbeitung in der ökosystemaren biologischen Forschung in Deutschland aufzuzeigen, den resultierenden Bedarf seitens der Wissenschaftsgemeinschaft zu definieren, die Akzeptanz langfristiger zentraler Datenspeicherung durch die Forschenden selbst zu analysieren und Vorschläge zu unterbreiten, wie eine optimierte Datenlangzeitspeicherung und -zugänglichkeit von den Forschungsförderern unterstützt werden kann.

2. Analyse des gegenwärtigen Zustands der Sicherung und Nutzung von Primärdaten im Bereich der biologischen Forschung in Deutschland

Nur für wenige Bereiche der biologischen Forschung gibt es nationale und/oder internationale Datenrepositorien[2], die eine Langzeitdatensicherung gewährleisten. Die bekanntesten internationalen Institute, die bestimmte Datenbanken, Repositorien und Analysefunktionen zur Verfügung stellen sind diejenigen für molekularbiologische Daten (z.B. als Bestandteil von NCBI, EBI, GenomeNet). Datenbankprojekte zur Speicherung und Bereitstellung von Primärdaten der umweltbezogenen biologischen Forschung werden in Deutschland, aber auch international, meist nur zeitlich befristet gefördert. Allerdings gibt es in verschiedenen Ländern der Welt auf nationaler Ebene Bestrebungen, die sich mit diesem Problem beschäftigen, z.B. DataOne (NSF/USA), Australian National Data Service (ANDS) und Joint Information Systems Committee. (JISC, UK). Die in Deutschland im Bereich der umweltbezogenen biologischen Forschung arbeitenden Repositorien sind meist individuelle Institutionen. Viele von ihnen benötigen kontinuierliche Zusatzfinanzierungen durch Projektmittel, Stiftungen oder andere Kofinanzierungs-Möglichkeiten, um die langfristige Datenverfügbarkeit zu gewährleisten. Die rechtlichen Aspekte der Dateneinspeisung bzw. -nutzung sind in den bestehenden deutschen Strukturen meist nur durch individuelle Einzel- und Kooperationsverträge gelöst, da in der Regel keine Rechtsberatung zur Seite steht. Die größten

[1] In diesem Kontext umfasst der Begriff "Primärdaten" alle Daten, die im Rahmen der umweltbezogenen biologischen Forschung erhoben werden. Dabei kann es sich z.B. um taxonomische, molekulare, phytochemische, ökologische und ökophysiologische, sowie verhaltensbiologische Daten, aber auch um abiotische Daten, z.B. von Ökosystemen, sowie Daten zum ökonomischen Wert von Organismen und Ökosystemen handeln.

[2] Im Gegensatz zur reinen Datenarchiven, wie sie früher betrieben wurde, wird hier der Begriff „Datenrepositorium“ für Datenbanken verwendet, die Daten im Hinblick auf eine breite und im Einzelnen vielleicht noch nicht absehbare Verwendbarkeit formatieren, kuratieren und speichern.

deutschen Datenrepositorien im Bereich der umweltbezogenen biologischen Forschung basieren zu 98% auf langjährig erprobten, stabilen rationalen Datenbanksystemen, wobei jedoch mehr als 70% der Datenrepositorien auf darunterliegende Ontologien (Definition von Begrifflichkeiten und ihre Beziehung zueinander) als flexible Lösung zum Austausch von Daten verzichten. Neuere Datenrepositorien implementieren Hilfskonstruktionen, sog. "work-arounds" zur Nachbildung von Ontologien, um diesem Problem zu begegnen. Maximal 70% der Datensätze in den bestehenden Strukturen und 85% zum Datensatz gehörende Zusatzinformationen (Metadaten) sind in Form international üblicher Standards hinterlegt. Eine nicht-standardisierte Hinterlegung von Daten resultiert aber in einer reinen Archivfunktion des Repositoriums, ohne einfache Möglichkeiten, die Daten über Suchalgorithmen wiederzufinden und zur Wiederverwendung unkompliziert umformatieren und mit anderen Datensätzen verschneiden zu können. Dagegen sind Datensätze, die standardisiert hinterlegt werden, international such- und vernetzbar, können einfach kuratiert, erweitert und für Analysen extrahiert werden und sind häufig an darauf aufbauende Softwaretools adaptiert, wodurch ihr Wert erheblich gesteigert wird. Die Qualität der in Deutschland hinterlegten Primärdaten der umweltbezogenen biologischen Forschung wird in der Regel vom Datenlieferanten definiert, nur zum Teil werden Datensätze auf ihre Vollständigkeit, Plausibilität oder anhand eines Datentyptests überprüft. Bisher gibt es an keinem der deutschen Datenrepositorien automatische „upload“-Funktionen. Manuelle Nacharbeiten zur Qualitätskontrolle und optimierten Datenhinterlegung sind überall notwendig. Unterstützung der Datenbanknutzer erfolgt in der Regel individuell durch E-mail- bzw. Telefonkontakt, nur selten über „online Helpdesks“. FAQs, Workshops, oder "step by step screenshots" sind weitere Möglichkeiten zur Unterstützung der Datenbanknutzung. Softwaretools steigern den Wert und die Attraktivität von Datenrepositorien deutlich.

3. Datenlangzeitspeicherung und –zugänglichkeit in der umweltbezogenen biologischen Forschung: Bedarf und Akzeptanz

Eine Umfrage unter 245 Wissenschaftlerinnen und Wissenschaftlern der umweltbezogenen biologischen Forschung aller Qualifikationsstufen (ENKE ET AL. SUBMITTED) im Jahr 2010/2011 ergab, dass mehr als 75% der Befragten grundsätzlich bereit sind, Ihre Daten in öffentlichen Repositorien zu hinterlegen. Allerdings konnte nur etwa die Hälfte der Wissenschaftlerinnen und Wissenschaftler Datenrepositorien benennen, die für die Aufnahme ihrer Datensätze in Frage kommen. Besonders großer Bedarf an Datenarchivierung wurde in den Bereichen ökologischer, physiologischer und mariner Biodiversitätsforschung genannt, für die es z.T. weder ausreichend nationale noch

internationale Möglichkeiten zur individuellen Datenhinterlegung gibt; auch ist der Bekanntheitsgrad solcher Datenrepositorien viel zu gering. Der Wunsch der Wissenschaftlerinnen und Wissenschaftler, ihre Daten in einem Langzeitarchiv zu hinterlegen, beruht auf ihrem Interesse an interdisziplinären Netzwerken und Kooperationen, sowie am Datenvergleich und der Transparenz von Forschungsergebnissen. Über 85% der befragten Wissenschaftlerinnen und Wissenschaftler bekundeten Interesse daran, Daten aus Langzeitrepositorien für ihre Forschung zu nutzen. Allerdings bestehen auch nicht unerhebliche Bedenken gegen die Langzeitarchivierung der eigenen Daten: So wird befürchtet, die Kontrolle über diese Daten zu verlieren und zu viel Zeit für die Eingabe der Daten in zentrale Repositorien investieren zu müssen. Häufig führt das Fehlen von "Datenmanagementplänen" zur unvollständigen Datenerhebung und laienhaften Datenverwaltung im Verlauf der Projekte. Der Datenumgang (Dokumentation, Aufbereitung, Einspeisung in Repositorien) muss erlernt werden, damit die Datenerhebung von Projektbeginn an im richtigen Format erfolgt. Der Mehrwert eines effizienten Datenmanagements sowie einer virtuellen Forschungsumgebung wird von den befragten Wissenschaftlerinnen und Wissenschaftlern durchweg erkannt und die Möglichkeit der Zitierung von hinterlegten Datensätzen wird gewünscht[3]. Allerdings stellen zum Teil fehlende Repositorienmöglichkeiten für biologische Daten, die unterschiedlichen Ausrichtungen der in Deutschland bestehenden Repositorien und die noch unzureichende Vernetzung dieser Datenbanken ein erhebliches Hindernis dar. Hier könnte eine nationale Dachstruktur in Form eines nationalen Datenservice-Zentrums mit zentraler Kontaktmöglichkeit Abhilfe schaffen. Die befragten Wissenschaftlerinnen und Wissenschaftler würde die Einrichtung einer solchen Dachstruktur ausdrücklich befürworten.

4. Vorschläge für eine optimierte Datenlangzeitspeicherung und -zugänglichkeit von Primärdaten der umweltbezogenen biologischen Forschung in Deutschland

4.1. Struktur und Kernkompetenzen eines Nationalen Datenservicezentrums für umweltbezogene biologische Daten

Die Einrichtung eines Nationalen Datenservicezentrums soll die bestehende Infrastruktur zur Speicherung und Verarbeitung von umweltbezogenen biologischen Daten grundlegend verbessern. Das Nationale Datenservicezentrum soll als Dachstruktur die bisher existierenden Datenbanken (sowohl Arbeits- als auch Repositorien-Datenbanken) zusammenführen, ihre Vernetzung/Interoperabilität entwickeln und für die weltweite Kompatibilität der Repositorien

[3] Piwowar HA, Day RS, Fridsma DB (2007): Sharing detailed research data is associated with increased citation rate. PLoS One. 21;2(3):e308.

sorgen (Ontologieentwicklung auf Basis internationaler Standards). Über das „Zentrale Portal“ sollen von außen eingehende Daten/Informationen im Nationalen Datenservicezentrum zum einen einer fachlichen Qualitätskontrolle zugeführt und zum anderen internationalen Datenbankstandards angepasst werden, bevor ihre endgültige Archivierung in den fachlich zuständigen Repositorien erfolgen kann. Da es für die sog. Arbeitsdatenbanken in zeitlich befristeten Forschungsprojekten noch keine Langzeitspeichermöglichkeit gibt, sollen solche Daten ggf. in einer eigenen Datenbank des nationalen Datenservicezentrums hinterlegt werden. Das Zentrum muss ein breites Spektrum von Serviceleistungen anbieten, wozu eine hohe IT Kapazität (große Rechenleistung) erforderlich ist. Eine mögliche Struktur des Nationalen Datenservicezentrums ist in Abb. 1 dargestellt.

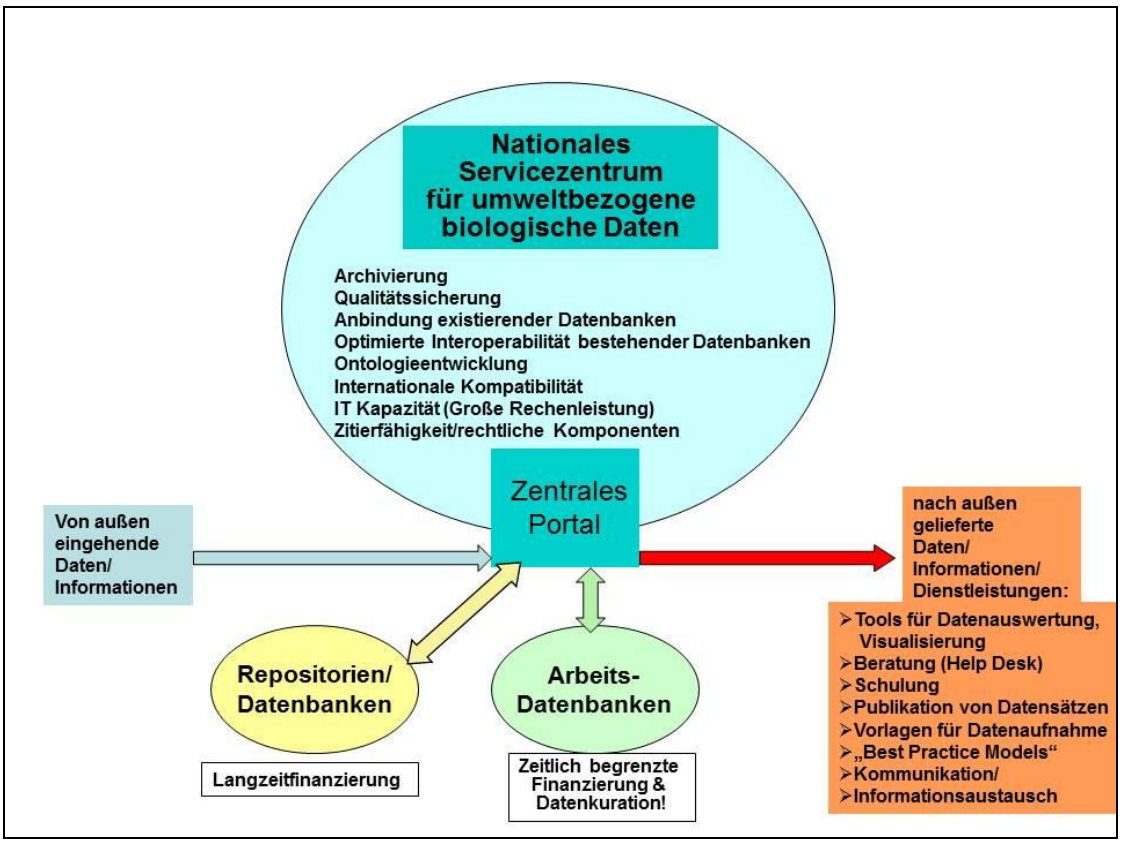

Abb.1: Struktur und Funktionen eines Nationalen Datenservicezentrums für die Biodiversitätsforschung.

Das Datenservicezentrum muss Kernkompetenzen erfüllen, sollte aber im Hinblick auf künftige Fragestellungen und technische Weiterentwicklungen flexibel konzeptioniert sein. Eine enge Verschneidung zwischen der zu planenden Dachstruktur des Datenservicezentrums mit anderen nationalen und internationalen Infrastrukturen ist äußerst wichtig, um den Datenfluss zu sichern und die Implementierung von Standards und Normen, die für eine Langzeitsicherung notwendig sind, zu gewährleisten. Hierbei kann auf Vorarbeiten des Kompetenznetzwerkes Langzeitarchivierung NESTOR und den dort entwickelten Kriterienkatalogen bzw. des Open Archival Information Systems (OAIS) als ISO-Referenzmodell[2] zurückgegriffen werden.

Das **zentrale Portal** des Nationalen Datenservicezentrums ist eine Anlaufstelle für Wissenschaftlerinnen und Wissenschaftler der umweltbezogenen biologischen Forschung sowie für externe Datenbankadministratoren (*Abbildung 1*). Das Zentrum soll vielfältige Dienstleistungen „nach außen“ für Nutzer anbieten. So z.B. bedürfen Wissenschaftlerinnen und Wissenschaftler, die bisher selten mit internationalen Datenbankstandards konfrontiert wurden, der Beratung und Schulung, und es muss eine Kommunikation bei Problemen der Dateneinspeisung bzw. des Datenexports möglich sein. Hierzu wird qualifiziertes Personal benötigt, das die Probleme der Wissenschaftlerinnen und Wissenschaftler versteht und diese mit den technischen Voraussetzungen der zusammengeschlossenen Datenbanken in Einklang bringen kann. Eine weitere Dienstleistung ist die Bereitstellung standardisierter Arbeitsblätter für die Erhebung von umweltbezogenen biologischen Daten und die später benötigten Kontextinformationen bei Projektbeginn, um die Einspeisung von Daten in das Archiv zu optimieren und bei Projektende eine zeitaufwändige und häufig qualitätsmindernde Umformatierung zu meiden. Desweiteren sollen Softwaretools zur Auswertung und Visualisierung der Daten bereitgestellt werden. Diese können in einem gewissen Umfang die Akzeptanz von Datenbanken der umweltbezogenen biologischen Forschung steigern, wie z.B. Verbreitungskarten (z.B. FishBase) oder das BLAST®-Tool der internationalen molekularen Datenbanken (NCBI, EBI, GenomeNet) zeigen. Die Dachstruktur soll ferner im Namen aller angebundenen Datenbanken als Ansprechpartner für die interessierte Öffentlichkeit dienen, für rechtliche Fragen der Datenkustodie zur Verfügung stehen und politische Lobbyarbeit für eine sinnvolle Nutzung öffentlich relevanter Daten betreiben.

4.2. Ausstattung des nationalen Datenservicezentrums

Der Aufbau eines zentralen Datenrepositoriums ohne darüber hinausgehende Servicefunktionen würde weder unter den Wissenschaftlerinnen und Wissenschaftlern, noch unter den IT-Administratoren bestehender Repositorien die notwendige Akzeptanz finden, d.h. keinen wesentlichen Mehrgewinn im Vergleich zu bestehenden Strukturen darstellen. Ohne standardisierte, kuratierte Datenhinterlegung ist die Wiederfindungsmöglichkeit der Daten nicht gewährleistet und der Mehraufwand der Dateneinspeisung nicht gerechtfertigt. Außerdem sollten nur Daten, die einer angemessenen Qualitätsprüfung unterzogen wurden, langfristig gespeichert werden. Überprüfungen sollten in Bezug auf den Autor, den Datentyp, außerhalb der Norm befindliche Daten (Ausreißer), Thesauri und Publikationen erfolgen.

Die Optimallösung des Datenservicecenters entspricht einem "Data Warehouse" mit hoch-integrierter Datenbasis, aktiver Datenpflege durch Kuratoren, umfangreichen Zugriffs- und Analysefunktionen, einem semantischen Web für Suchfunktionen, sowie wissenschaftlichen Werkzeugen und Softwaretools zur Meta-Analyse und Synthese.

Das Nationale Datenservicezentrum hat drei Kernbereiche:

1. Anlaufstelle und Ansprechpartner,
2. Technische Unterstützung, und
3. Entwicklung im IT-Bereich.

Personelle Ausstattung

Die Leitung des Nationalen Datenservicezentrums muss unter Wahrung der strategischen Ausrichtung die effiziente Funktionsweise der Institution organisieren und langfristig sicherstellen, Dialogpartner für die wissenschaftlichen Communities sein und auf nationaler und internationaler Ebene an der Entwicklung von Datenstandards mitarbeiten.

Eine wirkungsvolle Öffentlichkeitsarbeit ist besonders zu Beginn des Zentrums von großer Bedeutung, um es national und international bekannt zu machen, denn viele bisherige Initiativen waren infolge mangelnder Bekanntheit nicht erfolgreich. Die Abteilung für Öffentlichkeitsarbeit des Zentrums soll über alle Medien Informationsmaterial verbreiten und Schulungen und Workshops durchführen.

Helpdesk-Aufgaben können häufig von bestehenden Datenbanken in der biologischen Forschung in Deutschland nicht abgedeckt werden, da hierfür die personelle Kapazität nicht ausreicht. Die Helpdesk-Abteilung soll nicht nur das Datenservicezentrum selbst bedienen, sondern auch die angeschlossenen Datenbanken/Repositorien. Außer der Bearbeitung von Nutzeranfragen und Nutzerwünschen soll das Helpdesk "Best practice Modelle" und die Vorlagen für eine effektive Datenaufnahme entwickeln. Weitere Aufgaben sind die angemessene Datenkuration und die Einleitung der Qualitätskontrolle, bevor die Daten in den zuständigen Repositorien deponiert werden können.

Die IT-Abteilung des Zentrums soll die Vernetzung der bestehenden Datenbanken/Repositorien im Datenservicezentrum betreiben und die Pflege der Daten sicherstellen. Neben der Instandhaltung der Infrastruktur ist die IT-Abteilung für die Aktualisierung der Software und für die Archivierung und Datenversionierung verantwortlich. Sie entwickelt und pflegt die Ontologien des Zentrums. Softwareoptimierungen können im Rahmen von Projekten durchgeführt werden, z.B. die Programmierung von projektbezogenen Aufgabenstellungen oder die Programmierung von

Routinealgorithmen zur Qualitätssicherung oder von Import-Werkzeugen. Hier ergeben sich Möglichkeiten zur Einwerbung von Drittmitteln

5. Fazit

Das hier vorgeschlagene „Nationale Servicezentrum für umweltbezogene biologische Daten“, das die langfristige Speicherung und nachhaltige Nutzung von umweltbezogenen biologischen Primärdaten zum Ziel hat, wird der Forschung enorme Impulse verleihen. Neue Dimensionen können der biologischen Forschung in Deutschland erschlossen werden, um sie in der internationalen Spitze zu positionieren. Die Bedeutung der umweltbezogenen biologischen Forschung ist gerade im Zusammenhang mit dem globalen Wandel kaum zu überschätzen. Viele für politisch kluges Handeln notwendige Informationen können über die effektive Vernetzung deutscher einschlägiger Datenbanken/Repositorien im Nationalen Datenservicezentrum schnell zur Verfügung gestellt werden. Um die weitere Entwicklung der Aufnahme und Nutzung umweltbezogener biologischer Daten auf internationaler Ebene mitbestimmen zu können, benötigt die deutsche Gemeinschaft der Biologen und Ökologen eine Institution in der ihre Kompetenzen gebündelt und die vorhandene Information bestmöglich genützt werden kann. Das hier vorgeschlagene Nationale Datenservice-Zentrum entspricht dieser Institution. Allerdings drängt die Zeit, da die Entwicklung auf internationaler Ebene rasch voran schreitet.