



NeFo-Interview

Datenmanagement: DFG fördert Infrastruktur für öffentlich zugängliche Biodiversitätsforschungsdaten und bestimmt Richtlinien zum Datenmanagement

PD Dr. Birgit Gemeinholzer (Universität Gießen), Vorsitzende der AG Biodiversitätsdaten der DFG-Senatskommission für Biodiversitätsforschung



Im Rahmen von Biodiversitätsforschungsvorhaben produzierten Primärdaten sind bisher in der Regel nicht öffentlich zugänglich. Sie verschwinden meist in Schubladen, auf CDs und Sticks, werden unlesbar aufgrund von Softwareweiterentwicklungen und geraten in Vergessenheit, während die auf ihnen basierenden Analysen Eingang in Publikationen finden. Dies gilt für Einzelprojekte aber auch für größere Verbundprojekte, deren Datenhaltung nach Projektende meist



nicht mehr gewährleistet ist. Die Verfügbarkeit von Primärdaten kann jedoch sehr wertvoll sein, z.B. um Folgeuntersuchungen durchzuführen, Zeitreihen zu analysieren, Forschungsergebnisse zu reproduzieren oder indem die Primärdaten in einem anderen Kontext weiterverwendet werden. Dies spielt für immer mehr Forschungsbereiche eine wachsende Rolle.

Zur Unterstützung der Forschenden beim Forschungsdatenmanagement fördert die Deutsche Forschungsgemeinschaft deshalb den Auf- und Ausbau einer Infrastruktur GFBio, die künftig einen zentralen Zugang zu den relevanten, bereits jetzt existierenden Infrastrukturen, Analysewerkzeuge für Forschungsdaten anbieten soll. Um eine einheitliche und nutzbare Datenbasis zu erhalten, hat eine DFG-Expertengruppe Leitlinien zum Umgang mit Forschungsdaten in der Biodiversitätsforschung erarbeitet. PD Dr. Birgit Gemeinholzer ist Leiterin des Fachbereichs Spezielle Botanik an der Justus-Liebig Universität Gießen. Als Vorsitzende der AG Biodiversitätsdaten der Senatskommission für Biodiversitätsforschung der DFG koordinierte sie die Erstellung der Leitlinien. Im NeFo-Interview erzählt sie, was die Infrastruktur künftig leisten soll und was das für die Arbeit der Biodiversitätsforschenden bringt bzw. bedeutet.

Frau Gemeinholzer, welche Motivation und welchen Zweck haben die Richtlinien zum Umgang mit Forschungsdaten in der Biodiversitätsforschung, die von der DFG jetzt neu formuliert wurden?

Gemeinholzer: Dieser Leitfaden wurde von der AG Biodiversitätsdaten der Senatskommission für Biodiversität entwickelt um Antragstellern von Biodiversitätsprojekten Hilfestellung zur optimierten Datenplanung, Erfassung, Verarbeitung und Aufbewahrung an die Hand zu geben. Es handelt sich also um Empfehlungen, um die Forschenden in diesem Bereich zu unterstützen. Im Leitfaden für die Antragstellung der DFG gab es bereits einen Passus, in dem Antragsteller dazu Stellung nehmen sollten, wie sie mit den im Projekt erzielten Forschungsdaten umgehen möchten. Diese Richtlinien spezifizieren genauer, was bei der Sicherung, Aufbewahrung und Nachnutzbarkeit von Daten zu berücksichtigen ist.

Die Liste der Aspekte, die in den Leitlinien abgefragt werden, ist ziemlich lang. Befürchten Sie hier keinen Abschreckungseffekt?

Gemeinholzer: Eine Untersuchung von Enke et al. 2012 unter deutschen Biodiversitätsforschenden hat gezeigt, dass Forschende grundsätzlich gegenüber einer Datensicherung und Nachnutzbarkeit nicht negativ eingestellt sind. Allerdings ist häufig eine Umformatierung und Einspeisung in Biodiversitätsdatenbanken zeitaufwändig und kompliziert. Dies kann optimiert werden, wenn man sich bereits bei Projektbeginn über Fragen zum Datenmanagement Gedanken gemacht hat, damit die Daten gleich im richtigen Format erfasst und verarbeitet werden. Alle Punkte, die in den Richtlinien angeführt werden, sind in dieser Hinsicht von Bedeutung und stellen eine Orientierungshilfe im Sinne der Antragsteller dar.



Sekundäranalysen und Reproduktion von Forschungsergebnissen sind ja keine neuen Aspekte. Wieso kommt der Anspruch auf Zugänglichkeit der Forschungsdaten gerade jetzt?

Gemeinholzer: Lange Zeit war es für Forschende nicht einfach, die Primärdaten aus Forschungsergebnissen in Datenbanken zu hinterlegen. Häufig wussten die Forschenden meist nicht, wer die Daten wie und wo archiviert und eine Umformatierung in die entsprechenden Datenformate zur Langzeitarchivierung war häufig zeitaufwändig und umständlich. Ebenso umständlich war eine Wiederverwertung.

Untersuchungen des Arbeitskreises Biodiversitätsdaten der Senatskommission für Biodiversitätsforschung der DFG ergaben, dass eine dezentrale Vernetzung bestehender Datenrepositorien und Datenbanken mit zentralen Help-Desk-Strukturen und Funktionalitäten für den Nutzer einen sehr großen Mehrwert liefert. Diese für den Nutzer vorteilhafte zentrale Zugangsstruktur sollte jedoch nicht auf Kosten von langjährigen, gut etablierten, qualitativ hochwertigen dezentralen Datenbanken unterschiedlichster Expertise und Fokussierung erfolgen.

Um beiden Aspekten Rechnung zu tragen, mündete dies in die Etablierung von GFBio (German Federation for Biological Data, gfbio.org), ein dezentrales serviceorientiertes Dateninfrastruktursystem für biologische Daten und Umweltdaten mit einem zentralen Zugangsportal für den Forschenden. GFBio wird seit 2013 durch die DFG gefördert. GFBio befindet sich noch im Aufbau, bietet aber bereits jetzt Möglichkeiten, Forschungsdaten zu hinterlegen und wiederzuverwenden.

Was bedeutet „Im Aufbau befindliche“ Infrastruktur konkret, wie weit ist man tatsächlich?

Gemeinholzer: GFBio liegen zum einen funktionale Datenbanken zugrunde, die voll funktionsfähig sind und nur ein gemeinsames Zugangsportal bekommen, zum anderen werden bislang fehlende Strukturen zur Harmonisierung und Vernetzung von Sammlungs-, Molekularen- und Umweltdatenbanken neu etabliert, bzw. werden Nutzerschnittstellen als Serviceeinrichtung neu geschaffen. Zu diesen gehören, z.B. die Erstellung von Daten Management Plänen, die man vor Projektbeginn bei GFBio erstellen kann, um die Daten während dem Projektverlauf gleich im richtigen Format zu erheben, und geeignete Archive zu identifizieren, damit Daten nach Projektende unkompliziert in eine Langzeitarchivierung überführt werden können.

Andere Funktionalitäten, die gerade entwickelt werden, sind Visualisierungstools und Analysekomponenten verfügbarer Daten. Man arbeitet an der Etablierung von Qualitätsstandards, Hilfefunktionalitäten in Bezug zu Rechten an den Daten, Lernangeboten zur korrekten Datenerhebung und -archivierung, Suchfunktionalitäten, Dateneingabemasken, und vielem mehr, so dass insbesondere die Forschenden als Datenerheber aber auch Datennutzer einen großen



Mehrwert erhalten, ohne selbst Computerspezialisten werden zu müssen. Für Nutzer kann z.B. ein Blick auf das „[Data Life Cycle - fact sheets](#)“ im Support & Helpdesk zum Umgang mit den eigenen Forschungsdaten informativ sein. Die ersten Datensätze sind in GFBio eingespeist und jeder kann die ersten Funktionalitäten testen und sich online über das [Webportal von GFBio](#) informieren.

Welche bestehenden Dateninfrastrukturen sind für das Vorhaben relevant und wie sollen sie eingesetzt werden?

Gemeinholzer: Momentan haben sich 19 verschiedene Partnerinstitutionen mit ihren jeweiligen biodiversitätsrelevanten Datenbanksystemen in Deutschland mit finanzieller Unterstützung durch die DFG zu GFBio zusammengeschlossen. Anvisiert ist, möglichst alle auf internationalen Standards basierenden Dateninfrastrukturen der deutschen Biodiversitätsforschung mit ins Boot zu holen und mittels eines dezentralen Zugangsportals zu vernetzen. Selbstverständlich muss es Keimzellen einer Entwicklung geben, die jetzt mit der Entwicklung von GFBio in Gang gesetzt wurde. Ein Desiderat war jedoch von Beginn an, die Netzwerkstruktur so flexibel zu halten, um auch auf zukünftige Bedürfnisse der Biodiversitätsforschung flexibel reagieren zu können.

Wie geht man mit der allgemeinen Problematik der Eigentumsrechte und IPRs um?

Gemeinholzer: Eigentumsrechte und Intellectual Property Rights an Daten sind sowohl auf nationaler als auch auf internationaler Ebene relativ ungeklärt, weshalb es zu diesem Thema diverse internationale Arbeitsgruppen gibt, um Empfehlungen zu erarbeiten (z.B. CODATA Working Group on Legal Interoperability of Research Data). Diese werden in GFBio Eingang finden. Momentan werden die bislang üblichen Verfahren der einzelnen Datenzentren weiter beibehalten, die meist auf Creative Commons Lizenzen basieren, ehe neue Erkenntnisse vorliegen.

Werden nur Daten aus DFG-geförderten Projekten einbezogen, oder soll die Infrastruktur auch anderen Einrichtungen, Verbänden oder Ehrenamtlichen zur Verfügung stehen?

Gemeinholzer: Alle Primärdaten aus biologischen oder umweltbezogenen Forschungsprojekten können (abhängig von den verfügbaren Ressourcen) Eingang in GFBio finden, unabhängig davon, ob eine DFG-Finanzierung zugrunde liegt oder nicht. Bestimmte Datenstandards und die Vollständigkeit der Information sind für die Datenhinterlegung wichtig. Dies sichert zum einen die Qualität der Daten, zum anderen ermöglicht es eine computerlesbare Wiederfindung und Zitierung. Dafür müssen Begleitinformationen zu den Daten mit erfasst werden – wie dies funktioniert, [kann man online erfahren](#).

Geht es um rein naturwissenschaftliche Daten oder darüber hinaus?

Gemeinholzer: Zum momentanen Zeitpunkt stehen biologische Daten und Umweltdaten aus wissenschaftlichen Forschungsprojekten im Fokus von GFBio, bzw. Begleitforschung hierzu. Diese können aus verschiedenen Wissenschaftsbereichen stammen, z.B. von Ökologen, Physiologen,



Genetikern, Taxonomen, Systematikern o.ä. Sehr viele qualitativ hochwertige Datenbanken in Deutschland werden jedoch auch von Vereinen, Verbänden und ehrenamtlich Tätigen gepflegt, deren Einbindungsmöglichkeit langfristig auch Berücksichtigung finden soll.

Ist eine Anbindung an bestehende Infrastrukturprojekte wie GBIF oder GEOBON geplant?

Gemeinholzer: Selbstverständlich stand die Möglichkeit der Anknüpfung an bestehende internationale Dateninfrastrukturen von Anfang an im Fokus von GFBio. Viele Konsortiumsmitglieder in GFBio sind an GBIF-Deutschland beteiligt. Ein Memorandum of Understanding wurde zwischen EU BON, dem europäischen Ableger von GEO BON, und GFBio im April diesen Jahres unterzeichnet. Durch die Verwendung internationaler Datenbankstandards, Datensatzbeschreibungen (Ontologien) und entsprechenden Vernetzungstechnologien ist eine breite Interoperabilität auf internationaler Ebene auch in Bezug zu anderen Infrastrukturprojekten gewährleistet.

Wie sieht die Finanzierung aus? Wer soll die Kosten für GFBio übernehmen und wie lange?

Gemeinholzer: Momentan wird GFBio von der DFG im Förderprogramm „Infrastrukturen für Forschungsdaten“ finanziert dies ist jedoch explizit eine Anschubfinanzierung zur Infrastrukturförderung der Biodiversitätsforschung in Deutschland. Ein Bestandteil der Förderung ist die Etablierung eines tragfähigen Geschäftsmodells, welches den Kernbereich der Infrastruktur auf der Basis verschiedener Finanzierungsquellen auch langfristig aufrechterhält. Eine mögliche Säule dieser Finanzierung könnten projektspezifische Nutzungsgebühren sein, die in Analogie zu Gerätezentren in regulären Drittmittelprojekten mitbeantragt und bewilligt werden können

Was sind die nächsten Schritte?

Gemeinholzer: GFBio arbeitet auf Hochtouren an der Vernetzung deutscher Akteure im Bereich biologischer Daten und Umweltdaten. Parallel wird auch auf internationaler Ebene am Thema der Datenvernetzbarkeit und Interoperabilität gearbeitet. Mitglieder des *International Council for Science (ICSU)* und der *International Group of Funding Agencies for Global Change Research (IGFA)* haben sich im Belmont Forum zusammengeschlossen, um länderübergreifende Forschungsvorhaben im Bereich Global Change Research (GCR) zu ermöglichen. In diesem Rahmen wurde die e-Infrastrukturförderung und -vernetzung auf internationaler Ebene als prioritär erkannt. Eine Arbeitsgruppe des [Belmont Forums](#) erarbeitet gerade Empfehlungen zur Forschungsinfrastrukturförderung, um auch die internationale Datenverfügbarkeit und den Datenaustausch im Rahmen des globalen Wandel bestmöglich gewährleisten zu können.

Das Interview führte Sebastian Tilch



IMPRESSUM

Das Netzwerk-Forum zur Biodiversitätsforschung Deutschland (NeFo) ist ein Projekt gefördert durch das Bundesministerium für Bildung und Forschung (BMBF). Das Projekt wird maßgeblich durchgeführt vom Helmholtz-Zentrum für Umweltforschung – UFZ Leipzig und dem Museum für Naturkunde Berlin (MfN). Weitere Informationen und Hinweise zum NeFo-Projekt und Team unter www.biodiversity.de.

