Digitalgespräch Folge 35

Datenvielfalt handhabbar machen – das Beispiel Biodiversitätsforschung

Mit Barbara Ebert von der Gesellschaft für Biologische Daten e. V., 4. April 2023
https://zevedi.de/digitalgespraech-035-barbara-ebert/

[Der Vorspann mit Musik und Ausschnitten aus dem Gespräch beginnt.]

Marlene Görger [mg]: Frau Ebert, Sie sind Biologin und Wissenschaftsmanagerin und befassen sich besonders mit dem Management von Forschungsdaten und dem Aufbau von Dateninfrastrukturen.

Barbara Ebert [Ebert]: Um wirklich etwas beizutragen zu der Frage, wie schützen wir die Artenvielfalt, wie erhalten wir sie, braucht man komplexe Analysen, man braucht verschiedene Daten. Hier ging es vor allen Dingen darum, ein gemeinsames Portal zu bespielen. Je vielfältiger die Daten, desto schwieriger wird es da, ein gutes Angebot zu machen für die Nutzer.

Petra Gehring [pgg]: Wie sieht es aus mit den stofflichen Originalen, den nicht digitalen Bestandteilen dieser riesigen Sammlungswelt? Sind die von Bedeutung oder kann man die einfach abhängen?

[Ebert]: Es ist tatsächlich ein Praxisproblem, dass die Daten ganz oft für sehr genau definierte Nutzungszwecke erhoben werden und nicht für andere Anwendungen zur Verfügung stehen.

[pgg]: Es liegt dann daran, dass die Fische auch Persönlichkeitsrechte haben – ja wahrscheinlich eher nicht?

[Ebert]: Der Wille, gerade Daten für Forschung bereitzustellen, ist auf jeden Fall da. Ist, glaube ich, klar, dass wir im Forschungsdatenmanagement nur skalieren, wenn man wegkommt von den aufwändigen, händischen Exceltabellen.

[Der Vorspann endet, das Gespräch beginnt.]

[mg]: Wertvolle und wichtige, dabei zugleich heterogene, also sehr unterschiedliche Daten sollen für nützliche Zwecke möglichst breit und einfach zugänglich gemacht werden. Vor allem aber nicht nur der Forschung. Das ist eine häufige Forderung, und man hat oft den Eindruck, dass ihrer Erfüllung vor allem juristische Gründe entgegenstehen, oft zum Beispiel scheinbar der Datenschutz oder Eigentumsrechte. Besonders der Datenschutz wird breit diskutiert und auch wir haben in diesem Podcast schon viel darüber gesprochen. Weitaus weniger Beachtung findet in der Öffentlichkeit aber die Frage, was es eigentlich in der Praxis heißt, so ganz allgemein vorhandene, aber eben vielfältige Daten zu sammeln, zusammenzuführen und in guter Weise bereitzustellen. Dringend benötigt werden Daten, beispielsweise für die Forschungen zum Biodiversitätsverlust, also zum weltweiten Artensterben. Man hat hier viel zu spät damit begonnen, systematisch aufzuzeichnen und auch im Zusammenhang auf vorhandene Daten zu blicken. Geht das aber so einfach, wenn man sich einmal dafür entschieden hat? Wirft man einen Blick in die echte Welt des konkreten Machen-Wollens von Dateninfrastrukturen für die Forschung, wird das sehr

schnell sehr komplex. Wie immer, wenn man aus dem Überlegen ins Handeln kommt, und hinter dem harmlosen, knappen Oberbegriff der Bio- und Umweltdaten, die die Biodiversitätsforschung braucht, versteckt sich ein gigantisches Feld oft völlig unterschiedlicher, kaum vergleichbarer Daten. Welche sind das und wo kommen sie her? Was heißt es praktisch, hier von einzelnen Datensammlungen oder Datenbanken zu integrierten Dateninfrastrukturen zu kommen, sie nachhaltig zu pflegen und möglichst zentral für Forschende verfügbar und sinnvoll nutzbar zu machen? Und welche Ziele motivieren diese großen Anstrengungen? Darüber wollen wir heute im Digitalgespräch reden. Mein Name ist Marlene Görger. Ich bin Physikerin und Technikphilosophin und arbeite am Zentrum verantwortungsbewusste Digitalisierung.

[pgg]: Und ich bin Petra Gehring, Professorin für Philosophie an der Technischen Universität Darmstadt. Zu Gast ist heute im Podcast, wie immer, eine Expertin fürs Thema. Es ist Frau Dr. Barbara Ebert aus Bremen. Herzlich willkommen im Digitalgespräch Frau Ebert! Schön, dass Sie Zeit für uns haben.

[Ebert]: Ja, vielen herzlichen Dank für die Einladung.

[mg]: Frau Ebert, Sie sind Biologin und Wissenschaftsmanagerin und befassen sich besonders mit dem Management von Forschungsdaten und dem Aufbau von Dateninfrastrukturen. Seit 2021 sind Sie Geschäftsführerin der Gesellschaft für Biologische Daten e.V. oder kurz GFBio. Die GFBio widmet sich im Dienst der Forschung dem Management und der Standardisierung von Bio- und Umweltdaten und trägt mit ihrer Arbeit zur Entwicklung von Infrastrukturen, insbesondere für die Biodiversitätsforschung, bei. Wir wollen heute von Ihnen wissen, was das bedeutet, und zwar vor allem auch praktisch in der Umsetzung. Was tut die Gesellschaft für biologische Daten? Was sind ihre Aufgaben?

[Ebert]: Die Gesellschaft für biologische Daten ist ein eigentlich noch recht junger Verein, der sich gegründet hat aus einem mehrjährigen Verbundprojekt, an dem verschiedene Sammlungen, Forschungseinrichtungen und wissenschaftliche IT-Servicezentren in Deutschland mitgewirkt haben. Das Ganze wurde ursprünglich initiiert aus der Wissenschaft heraus von der Senatskommission für die Biodiversitätsforschung in der Deutschen Forschungsgemeinschaft. Die hatten nämlich festgestellt, dass es da doch erhebliche Desiderate gibt in der Verfügbarkeit der vielen Daten, die aus den geförderten Projekten in der Deutschen Forschungsgemeinschaft entstehen. Insgesamt haben sich da knapp 20 Partner zusammengetan und haben angefangen, zum einen systematisch eigene Datensammlungen zu mobilisieren und Daten an ein gemeinsames Portal zu liefern. Zum anderen wurde ein Service Portfolio entworfen, das speziell für die Bedarfe der Wissenschaft gedacht war. Entlang des sogenannten Datenlebenszyklus, also von der Erfassung der Daten über die Analyse und Publikation von Daten bis hin dann zum Zusammenführen bereits vorhandener Daten mit neuen experimentellen oder Feldforschungsdaten, die dann aus weiteren Projekten entstehen. Und die Aufgabe des Vereins ist es zunächst erstmal, die Nachhaltigkeit, also den Weiterbetrieb dieser erreichten Ergebnisse aus den siebeneinhalb Jahren GFBio-Projekt sicherzustellen. Und wir sind seit 2020 auch koordinierender Partner in einem neuen, viel größeren Verbundprojekt innerhalb der nationalen Forschungsdateninfrastruktur, bringen da diese Vorergebnisse und Services ein und entwickeln mit unserem hier sehr stark gewachsenen Mitarbeiterstab jetzt neue Dienste zusammen mit dem auch sehr stark gewachsenen Partnernetzwerk, was sich von 20 auf rund 50 Partner erhöht hat.

[mg]: Wer ist da alles so dabei, bei diesen 50 Partnern?

[Ebert]: Ja, das Konsortium, dieser neue große Verbund innerhalb der nationalen Forschungsdateninfrastruktur, da wird mit Forschungseinrichtungen, die zum Teil wiederum andere große Forschungsverbünde koordinieren, wir haben die großen naturkundlichen Sammlungen in Deutschland als Mitglieder, dann eine ganze Reihe von Expertenorganisationen für bestimmte Arten – das war auch ein bisschen Neuland für mich – es gibt große Fachgesellschaften von Vogelliebhabern, Vogelexperten, von Spinnenexperten, von Libellenfreunden, und die erheben sehr, sehr viele Daten, bringen die mit in unser Netzwerk – sehr spannend. Und wir haben IT-Servicezentren aus der Wissenschaft und Informatiklehrstühle dabei, weil es auch ein ganzes Stück weit eben um den Aufbau von IT-Servicestrukturen geht, von der Software bis hin zum Betrieb und Hardware.

[pgg]: Klingt jetzt alles noch ein bisschen abstrakt für mich. Ich habe verstanden: Also Sammlungen kommen zu Sammlungen und Forschungsergebnisse und vielleicht beim Forschen erhobene Daten, aber auch so aus Liebhabergesellschaften entstandene Sammlungen von Daten kommen zusammen. Aquatisch war jetzt ein Stichwort gewesen, also was mit Wasser, und verschiedene Tiersorten. Jetzt liegt die Herausforderung ja in der Heterogenität der Daten, also in der Verschiedenheit dieser Daten. Jetzt könnte man erst mal sagen: Naja, ob ich jetzt Libellen abbilde oder ob ich Spinnen abbilde oder ob ich Vögel abbilde, das ist ja doch nicht so besonders heterogen. Wo liegt die Quelle der Vielfalt?

[Ebert]: Ja, die Quelle der Vielfalt liegt tatsächlich in den Datentypen. Ich nenne vielleicht mal die beiden Forschungsfelder, um die es da konkret geht, auf die wir spezialisiert sind. Das ist zum einen die biologische Systematik, also die Kunde vom Stammbaum der Arten, vom Stammbaum des Lebens: Welche Arten gibt es, wie sind sie verwandt? Dort werden verschiedene Methoden eingesetzt, unter anderem molekularbiologische Methoden. Da haben wir es also zum Beispiel mit molekularen Sequenzdaten zu tun. Dann ist ein weiteres großes Forschungsfeld die Ökosystemforschung, also: Wie interagieren Arten mit ihrer Umwelt miteinander? Dort haben wir Datentypen, zum Beispiel aus Erhebungen: Welche Art finde ich zu welchem Zeitpunkt, an welchem Ort? Also räumlich zeitliche Informationen über das Vorkommen von Arten. Das kann verbunden sein mit einem gesammelten Tier oder einer gesammelten Pflanze, also einem physischen Objekt, können aber auch rein digital übermittelte Beobachtungsdaten sein, zum Beispiel bei den Vogelliebhabern, die Aufzeichnungen führen da drüber. Ja, es gibt eine ganz bunte Methodenvielfalt in der Biodiversitätsforschung, auch was Laboranalysen angeht, Studien am lebenden Objekt: Was sind physiologische Eigenschaften von Pflanzen oder Mikroorganismen? Auch das wird in den Datenbanken mit hinterlegt. Wir haben auch Experimente oder Feldexperimente im Freiland, wo man untersucht, wie verändern sich Lebensgemeinschaften, wenn man bestimmte Interventionen vornimmt, also neue Arten einführt oder Störungsexperimente durchführt, um zu schauen, wie regeneriert sich der Pflanzenbestand, nachdem man gerodet hat oder Pflanzen entnommen hat. Das sind vielfältige, ja räumlich zeitlich organisierte Daten aus solchen Experimenten und Untersuchungen, die für das Verständnis der Ökosysteme insgesamt natürlich unheimlich wichtig sind. Also: Warum nimmt zum Beispiel die Artenvielfalt wieder zu? Bringen Renaturierungsmaßnahmen was? Oder: Welche Faktoren führen zu einer Verarmung im Bereich der Artenvielfalt? Also: Warum reduziert sich die Vielfalt von

Arten zum Beispiel in der Umgebung von Siedlungen oder nach bestimmten Eingriffen in die Natur? Das heißt, um wirklich etwas beizutragen zu der Frage, „Wie schützen wir die Artenvielfalt, wie erhalten wir sie?“, braucht man komplexe Analysen, man braucht verschiedene Daten zu dem konkreten Vorkommen von Arten auf einem Zeitstrahl: Früher, jetzt, vielleicht extrapoliert, wie könnte das in Zukunft aussehen? Und welche Umweltbedingungen sind dafür vielleicht entscheidend?

[pgg]: Und gleichzeitig ganz, ganz klein, sage ich jetzt mal, was weiß ich, Biochemie, DNA sowas in der Art. Das ist ja nicht nur klein, sondern auch sehr chemisch, aber eben winzigst. Und dann aber auch die einzelnen Exemplare bis hin zu den ganzen Systemen.

[Ebert]: Ja. Und das ist eben das Interessante daran, Daten aus verschiedenen Quellen zusammenzuführen. Ganz großes Interesse in der Forschungscommunity ist zum Beispiel die Verbindung von Vorkommensdaten mit Umweltdaten, also Niederschlag, Bodenfeuchtigkeit, Temperatur, Bodenbeschaffenheit, aber auch zum Beispiel Siedlungsdichte oder Informationen über die landwirtschaftliche Nutzung in bestimmten Gebieten. Da reden wir dann über Satellitendaten, da reden wir über Karten, die von ganz anderen Institutionen erstellt werden, die nicht biologisch sind, sondern eher geowissenschaftlich zum Beispiel ausgerichtet. Und auch solche Forschungseinrichtungen sind jetzt in dem neuen großen Verbund, im Konsortium NFDI4Biodiversity, Partner. Und wir haben einzelne Anwendungsfälle, in denen wir Daten zum Beispiel aus dem Institut für Raumentwicklung zusammenbringen mit den Daten des Libellenatlas von Deutschland oder des Fischartenatlas von Deutschland, um zu schauen: Wie kann man da sinnvolle Korrelationen herstellen? Macht es überhaupt Sinn? Sind diese Datensätze wissenschaftlich sinnvoll kombinierbar? Denn letztlich können Sie ja nur eine gute Aussage treffen, wenn nicht nur die Orte zusammenpassen, das heißt, Sie legen den Libellenatlas über den Atlas der Siedlungsdichte in Deutschland, sondern wenn auch die Zeit passt. Es muss ja auch der passende Zeitraum sein, den man da vergleicht, den man da übereinanderlegt. Und das ist gar nicht trivial. Da kommt man dann in einen hohen Komplexitätsgrad und braucht auch sehr, sehr gute Informationen zu diesen Datensätzen und eine sehr klare wissenschaftliche Fragestellung.

[mg]: Man merkt dann ja schon daran, wie Sie das schildern, also, Sie haben einerseits ganz unterschiedliche Gruppen von Akteuren, die Daten beisteuern. Das sind nicht mal alles unbedingt digitale Daten, sondern manchmal muss man da überhaupt erst dafür sorgen, dass man das irgendwie in einen digitalen Zusammenhang kriegt. Man muss da doch bestimmt, ich will jetzt nicht sagen, Abstriche machen, aber Entscheidungen treffen mit einer gewissen Tragweite dann auch, wie man vorgeht, wenn man da Standards setzt. Oder wenn ich jetzt zum Beispiel mir vorstelle, ich bekomme, wie Sie es gesagt haben, ein gesammeltes Tier oder ein gesammeltes Pflanzenteil, dann muss ich ja dann entscheiden, was davon ist jetzt wichtig und relevant. Also spielt das so eine große Rolle, wie ich mir gerade vorstelle, oder ist das eigentlich klar für Fachleute, wie man das macht?

[Ebert]: Nein, das sind genau die Aushandlungsprozesse, die da stattfinden im Konsortium. Und deshalb ist es auch so gut, einen Mix zu haben von Wissenschaftlern und Wissenschaftlerinnen, die wirklich in der Forschung drin sind und ihre Schmerzpunkte bei der Datenverfügbarkeit gut kennen. Und eben Einrichtungen, die diese Daten haben, die auch ein Interesse haben, die für die Forschung bereitzustellen

und diese beiden Parteien zusammenzubringen. Das machen wir tatsächlich sehr kleinklein anhand von konkreten Anwendungsfällen. Wir haben 26 sogenannte Use-Case-Projekte, in denen wir einzelne Fragestellungen angehen. Ich greife mal den Faden auf: Mit dem gesammelten Tier oder der gesammelten Pflanze, wo landet die? Und mache das mal am fiktiven Beispiel eines Herbariums. Wir haben einen großen Partner, Botanischen Garten, die nicht nur historische Daten verwahren, sondern auch immer wieder aktuelle Pflanzenfunde aufnehmen, katalogisieren.

[pgg]: Kurz dazwischen: Herbarium, das heißt, das sind tatsächlich gesammelte Blätter oder gepresste Blätter?

[Ebert]: Genau, im Herbarium werden gesammelte Pflanzen, wenn sie denn auf einen Papierbogen gepresst passen, fixiert und beschriftet. Dort werden aber auch zum Beispiel große Früchte gesammelt. Die müssen wieder etwas anders aufbewahrt werden. Aber das klassische Herbarium hat eine große Sammlung von Papierbögen, auf die Pflanzenblätter, Pflanzenteile aufgeklebt sind, zusammen mit Informationen über die Funddaten und die sammelnde Person und die Bestimmung der Art. In meinem Beispiel des Botanischen Gartens haben die ein internes Datenmanagementsystem. Da werden die Informationen zu den Fundstücken eingespeist, und das sind natürlich sehr, sehr viele Parameter. Die richten sich auch danach: Was ist an der Einrichtung üblich? Was hat vielleicht das Projekt noch erfordert an Kontextinformationen? Die haben sehr, sehr viele sogenannte Metadaten zu dieser einen Pflanze, die gesammelt wurde. In einem Netzwerk wie NFDI4Biodiversity oder GFBio, wo wir von vielen solchen Einrichtungen Daten zusammenführen wollen, brauchen wir nicht alles, was in solchen internen Datenmanagementsystemen gesammelt wird. Das heißt, ein Aushandlungsprozess, der aber schon abgeschlossen ist, war die Festlegung eines Kerndatensatzes zu so einem Artvorkommen, zu so einer gefundenen Pflanze. Welche Daten liefere ich damit? Da haben sich die Sammlungen auf einen Satz von Konsensuselementen verständigt, und aus ihren großen Dateninformationssystemen liefern die Auszüge. Die Artvorkommen, die sie haben, und zu diesen Artvorkommen genau die Konsensuselemente, die sie festgelegt haben, liefern sie an das gemeinsame Portal, liefern sie auch an die große internationale Dateninfrastruktur im Bereich der Biodiversitätsforschung. Und diese Daten sind dann sehr gut miteinander kombinierbar, weil sie eben die gleichen Parameter enthalten: Ort, Zeit, Art und noch verschiedene andere Angaben, aber alle einheitlich.

[pgg]: Und die sind dann auch technisch schon standardisiert, also kommen alle in den gleichen Formaten bei Ihnen an?

[Ebert]: Genau. Und maschinenlesbar und zitierbar. Das ist auch wichtig in der Wissenschaft, dass man eben sagen kann, welchen Datensatz, wenn man diese Daten verwendet, welchen Datensatz mit welcher Identifikationsnummer habe ich denn für meine Analyse verwendet, damit das Ganze nachvollziehbar ist. Das war ein großer Meilenstein. Das haben die Datenzentren in GFBio geleistet. Und genau mit diesen Konsensuselementen und der guten Praxis, die dort entwickelt wurde, arbeiten wir jetzt weiter in der nationalen Forschungsdateninfrastruktur, um weitere Datenlieferanten, die solche Typen von Daten haben, auch anzuschließen und weitere Daten nach diesem Schema zu mobilisieren.

[pgg]: Ist nach wie vor da Einigkeit bei den Beteiligten, dass dieser Kerndatensatz, diese, ich sage mal, minimale Zahl von Informationen, die man von allem braucht, dass das ausreicht für die Biodiversitätsforschung? Oder gibt es schon die ersten Versuche, das noch aufzubohren und noch feiner zu erfassen? Das ist ja wahrscheinlich die Tendenz. Man wird ja wahrscheinlich nicht 50 Jahre lang mit demselben Kerndatensatz forschen. Oder muss man das gerade tun?

[Ebert]: Im Moment haben wir die Diskussion nicht im Netzwerk, sondern eher die Nacharbeiten bei einzelnen Datenanbietern, die vielleicht noch nicht alle diese Parameter drin haben in ihren Datensätzen. Aber ich kann mir vorstellen, dass das kommt. Und vielleicht zum Trost auch für alle Forschenden: Wenn es nicht reicht, kann man sich an die Einrichtung auch nochmal wenden und natürlich für das eigene Projekt besprechen, ob nicht die Lieferung eines anderen Datensatzes möglich ist mit anderen Parametern. Hier ging es vor allen Dingen darum, so ein gemeinsames Portal zu bespielen und da möglichst gut definierte Datensätze drin zu haben, die bestimmte Parameter eben verlässlich verfügbar sind.

[mg]: Wir hatten ja ganz zu Beginn auch mal das Datenmanagement angesprochen. Wir haben jetzt ja so ein paar Beispiele gehört, wie man da vorgeht. Ist denn im Grunde, was Datenmanagement jeweils heißt, auch schon geklärt? Oder vielleicht anders gefragt: Was heißt das denn, im Einzelfall Datenmanagement einzuführen, wo es vielleicht noch fehlt?

[Ebert]: Die Bandbreite der Systeme oder Softwarehilfen, die die einzelnen Akteure haben, um ihre Daten zu speichern, zu sortieren und Qualität zu sichern, die ist ganz groß. Die geht wirklich von der Exceltabelle bis hin zu einem komplexen Data Warehouse nach neueren Standards, die tolle Datenexportformate haben, die automatische Qualitätssicherung haben, die helfen, fehlende Felder zu identifizieren oder nichtplausible Informationen aufspüren. Da ist die Bandbreite sehr groß. Und ich denke, das ist auch ein Mehrwert von der Mitarbeit in diesem Netzwerk zwischen denen, die noch sehr einfache Systeme für ihr Datenmanagement benutzen und keine elektronischen Hilfen haben für die Qualitätssicherung. Denen können wir helfen, mit den bereits etablierten Tools bei anderen Einrichtungen, die sich gut damit auskennen, ein bisschen Starthilfe zu geben oder anzubieten, deren Daten vielleicht sogar zu hosten.

[mg]: Das heißt, Sie machen auch ganz konkret die Umsetzung von Datenmanagement und auch für einzelne Akteure. Es geht nicht nur sozusagen immer darum, in die große Infrastruktur reinzuarbeiten, sondern auch das Management, das darüber hinausgeht, wo dann vielleicht auch noch mehr Informationen vorgehalten sind, das spielt da auch eine Rolle?

[Ebert]: Genauso ist es. Es ist ja immer ein Geben und Nehmen in so einem Netzwerk. Also, wir haben auch Anwendungsfälle, ich sprach ja über diese Use-Case-Projekte. Da geht es zum Teil darum, die Datenmobilisierung dadurch zu unterstützen, dass man der Einrichtung hilft, erst mal ein internes Datenmanagementsystem zu implementieren und die internen Prozesse so zu organisieren, dass die Exporte nicht so viel Arbeit sind. Es ist insbesondere auch etwas, wo ich denke, dass das Forschungsdatenmanagement insgesamt nur vorankommen wird, wenn wir da ansetzen. Im Moment ist es für Einzelwissenschaftler ja erheblich aufwendig, am Ende eines Projekts dann einen publizierbaren Datensatz hervorzubringen, wenn die ganzen

Transformationen, die die Daten vorher durchlaufen haben, die ganze Aufnahme der Daten in vielleicht einem 2- bis 5-jährigen Forschungsprojekt mit verschiedenen Hilfsmitteln gemacht wurden, die hinterher nur händisch noch ausgewertet werden können, um die Datenpublikation hervorzubringen. Und wenn man sich da anschaut, was es zum Teil an Datenmanagementsystemen gibt, die solche Exportdatenpakete viel, viel leichter produzieren, in denen man schon während des Forschungsprozesses alles einpflegen kann an Informationen, was man braucht, dann ist glaube ich klar, dass wir im Forschungsdatenmanagement nur skalieren können, wenn solche Hilfsmittel auch zum Alltag werden und man wegkommt von den aufwändigen, händischen Exceltabellen. Gerade bei großen Datenmengen.

[pgg]: Ist es letztlich so, dass dann die Biodiversitätsforschung auch zusammenwächst? Also letztlich die gut geordneten, gut gemanagten Systeme, denen anderer Domänen oder anderer Wissenschaftsdisziplinen ähneln? Oder ist das am Ende dann schon so, dass das sehr spezifisch auf die Biodiversitätsforschung zugeschnitten sein wird?

[Ebert]: Ich würde mal sagen, die Datenbanktechnologien sind bestimmt nicht biodiversitätsspezifisch. Aber die Zuschnitte der Datenbanken müssen natürlich zu den Methoden passen und zu den Arbeitsabläufen in der Biodiversitätsforschung.

[pgg]: Das heißt, es überwiegt dann schon das Spezifische. Das Portal wird eins bleiben für die Biodiversitätsforscher:innen. Und es geht eher darum, wirklich maßgeschneidert zu werden mit all den Angeboten?

[Ebert]: Ich denke schon. Die Biodiversitätsforschung oder die Biodiversitätscommunity, sage ich jetzt mal, weil es nicht nur um Forschung geht, sondern eben ganz viel auch um hoheitliche Aufgaben der Überwachung von Ökosystemen oder des regelmäßigen Monitorings von Arten, zum Beispiel in Deutschland. Das sind Akteure, die haben einfach individuelle Bedürfnisse. Und wenn man die auf einen großen Datensee loslässt mit Daten verschiedenster Art, Umweltdaten, Vorkommensdaten, Sequenzdaten, dann haben die je nach Anliegen oder Teilcommunity, suchen die ja bestimmte Informationen. Und wenn ich versuche, das alles in einem Portal zu machen, wird man sehr schnell merken, dass die Teilcommunities mit keinem der Suchangebote so richtig zufrieden sind. Deswegen ist unser Ansatz eher, dass wir eine Infrastruktur in der Cloud schaffen, wo man die Daten gut speichern kann, wo es auch gute Technologien gibt, um die semantisch zu erschließen. Aber dass man nach außen zum Nutzer hin die Möglichkeit schafft, auch für Teilcommunities gute Angebote zu machen. Zum Beispiel die biologischen Systematiker – die Erfahrung haben wir gemacht mit dem GFbio Portal, das ist ja so ein gemischtes Portal mit Umwelt- und Vorkommensdaten von Arten –, die gehen oft lieber dann zu GBIF, zu der internationalen Biodiversity Information Facility, weil sie dort speziell die Vorkommensdaten haben und mit Such- und Filterfunktionen, die speziell die Parameter betreffen für die Vorkommensdaten. Solche detaillierten Filterfunktionen haben wir in dem GFBio Portal allgemein nicht, weil es dort eben auch viele andere Datentypen gibt. Wenn ich also so eine spezialisierte Fragestellung habe in der biologischen Systematik, suche ich mir vielleicht lieber ein Portal, was eben auch genau die Metadaten und Parameter erschließt, die mir wichtig sind. Je vielfältiger die Daten in einer Sammlung, desto schwieriger wird es, da ein gutes Angebot zu machen für die Nutzer.

[pgg]: Das heißt, Sie haben im Grunde eigentlich zwei Blickrichtungen? Einerseits die Datengeber, die kriegen ein Portal, in das sie hineingeben, und auf der anderen Seite die Datennutzer, die dann wiederum, aber vielleicht sogar unterschiedliche Portale nutzen, um auf die Bestände zuzugreifen.

[Ebert]: Genau. Und da sind wir tatsächlich ein bisschen Experimentierfeld gerade, von GFBio kommend, was noch relativ homogen war, starker Fokus auf die biologische Systematik und die Sammlungen. Jetzt mit dem sehr viel breiteren Feld von Datengebern und auch Interesse von anderen Fachgebieten an diesen Daten muss man dann tatsächlich schauen, wie bekommen wir das hin, dass die Datengeber gut liefern können und wissen, was sie zu tun haben, und wir auch sicher sein können, dass sie die Verfügungsrechte haben. Und den Datennutzenden Angebote zu machen, wo sie möglichst in einer Umgebung, die auf ihre Methoden und Fragestellungen gut zugeschnitten ist, diese Daten dann abrufen können. Aber das ist ja das Schöne an der Digitalität, das bekommt man ja einigermaßen hin. Man kann die Daten mehrfach ausspielen an verschiedene Datendienste oder in verschiedenen Datenprodukten benutzen, vorausgesetzt, man hat sie eben einigermaßen erschlossen und ist auch sicher, dass man sie benutzen darf. Das ist tatsächlich im Moment ein Praxisproblem, dass die Daten ganz oft für sehr genau definierte Nutzungszwecke erhoben werden und gar nicht für andere Anwendungen zur Verfügung stehen.

[mg]: Das heißt, ich frage jetzt nochmal nach, Sie hatten ja die verschiedenen Anliegen der unterschiedlichen Nutzergruppen oder Communities erwähnt: Was sind das denn für Anliegen? Sind das, sage ich mal, trotzdem immer noch alles Forschungsanliegen, oder sind da auch noch andere dabei? Wie stark unterscheiden die sich?

[Ebert]: Also da wir in der nationalen Forschungsdateninfrastruktur finanziert werden, sind unsere Nutzerinnen und Nutzer, die wir so ins Auge fassen, tatsächlich eher die wissenschaftlich motivierten Projekte oder Datennutzer. Eine Sache, die vielen am Herzen liegt, ist zum Beispiel, dass sie die Artenfunde, die sie selber machen und dokumentieren, referenzieren können, auf Artenfunde, die in dieser Region schon vorher gemacht wurden, und dass sie die Artenlisten und die Bestimmung der Arten, die sie finden, abgleichen können mit anderen Artenlisten. Und diese Informationen sind zum Teil für die maschinelle Anwendung im Moment überhaupt nicht gut verfügbar. Also ein ganz großes Ziel ist es, die Referenzlisten für Arten, für verschiedene Familien und Gattungen so verfügbar zu machen, dass man seine eigenen Ergebnisse schnell damit abgleichen kann.

[pgg]: Ganz kurz noch – Sie haben eben gesagt, der Erhebungszweck war unter Umständen sehr spezifisch, als die Daten erhoben wurden, und deswegen hat man jetzt ein Nutzungsproblem. Das kann ich mir jetzt gerade gar nicht vorstellen. Also irgendwo wurden mal Tiere gezählt, ich sage mal jetzt, vielleicht in einem Gewässer, und dann ist das vor 20 Jahren gewesen. Vielleicht kann man die Leute gar nicht mehr erreichen und die Tiere wurden vielleicht zu Zwecken des Angelns gezählt. Wieso ist das jetzt ein Problem, die zu verwenden? Oder liegt das Beispiel völlig falsch?

[Ebert]: Das Beispiel geht schon in die richtige Richtung. Ich nehme mal einen Prozess, der relativ verbreitet ist: Es gibt so, wenn Sie mal ein bisschen schauen, im Bereich der biologischen Systematik sogenannte Atlanten, also den Atlas der Fischarten, den Libellenatlas oder Atlanten zum Vorkommen bestimmter Vogelarten. Atlanten werden vielleicht alle zwei bis fünf Jahre publiziert und die Daten für die jeweils neueste

Ausgabe, die werden zusammengesammelt aus verschiedenen Quellen. Das können Erhebungen von Landesämtern sein, die mal publiziert worden sind. Das kann kombiniert werden mit Beobachtungsdaten von Expertenorganisationen, verschiedene Quellen, die also herangezogen werden, um diesen Atlas zusammenzustellen, den zu veröffentlichen. Der zugrunde liegende Datensatz zu diesem Atlas, zu dieser Publikation ist aber dann nicht unbedingt öffentlich für andere Nachnutzungen verfügbar.

[pgg]: Es liegt dann daran, dass die Fische auch Persönlichkeitsrechte haben? Ja, wahrscheinlich eher nicht. Ist das dann ein Verlag, der die Rechte hat, oder wer genau bremst da?

[Ebert]: Da bremsen zum Teil die ehrenamtlichen Experten, die mit hohem persönlichem Einsatz diese Daten sammeln und nicht möchten, dass sie zum Beispiel unautorisiert für Zwecke verwendet werden, die sie nicht gutheißen. Das heißt, da geht es um die Verfügungsrechte an den Daten, die nur für einen bestimmten Zweck zur Verfügung gestellt wurden. Und teilweise sind andere Nutzungen eben nicht vereinbart worden mit den Erzeugern, sag ich jetzt mal, der Daten. Und da Sie als Herausgeber eines solchen Atlas natürlich auf die Mitwirkung all Ihrer Quellen und Kooperationspartner angewiesen sind, können Sie nicht ohne deren Zustimmung den zugrunde liegenden Datensatz dann im Open Access publizieren. Dann springen die Ihnen beim nächsten Mal ab.

[pgg]: Führen Sie dann auch Verhandlungen um diese Rechte? Ist das eine der Aufgaben von GFBio?

[Ebert]: Das ist tatsächlich eine der Aufgaben, die sich aus diesen Anwendungsfällen ergeben, nämlich aus der Mobilisierung von Daten, hier die Praxis zu ändern. Und da sind wir in der Wissenschaft. Genau wie bei den anderen Datenerzeugern ist da noch sehr viel Überzeugungsarbeit zu leisten. Ist aber auch noch viel Sicherheit zu schaffen. Wie kann man diesen ja auch sehr nachvollziehbaren Bedenken einiger Datenerzeuger Rechnung tragen? Ich sag jetzt mal, das eine sind vielleicht unerwünschte Nutzungen oder dass da Sorge besteht: Jemand verdient Geld mit dem, was man in seiner Freizeit mit großem Aufwand erhoben hat, oder mit dem wissenschaftlichen Herzensprojekt. Man weiß da nichts von. Die Daten werden verwendet für einen Zweck, mit dem man nicht einverstanden ist. Das sind vielleicht individuelle Sorgen, die man auch durch Positivbeispiele ausräumen kann. Eine ganz praktische Sorge ist tatsächlich im Bereich der seltenen Arten, dass man deren Standorte nicht genau preisgeben möchte, damit eben diese Arten weiter gut geschützt bleiben. Das andere ist ebenfalls ein bisschen ein Datenschutzproblem, dass teilweise ja die Informationen über das Vorkommen von Arten nicht nur auf öffentlichem Gelände gesammelt werden, sondern auch auf Landstrichen in Privatbesitz. Und stellen Sie sich vor, da ist eine seltene, geschützte Art, die wurde beobachtet auf dem Acker oder dem Gelände, das sich in Privatbesitz befindet. Das kann zurückgeführt werden auf den Eigner dieses Grundstücks. Es können ihm irgendwie Nachteile entstehen. Das heißt, ein Teil dieser Aushandlungsprozesse betrifft die Granularität der Daten, die zur Verfügung gestellt werden für die Nachnutzung. Und da kommen wir auch wieder in so einen Bereich, wo Standardisierung verhandelt wird, das heißt, auf einer Karte, in welcher Auflösung, in welcher Rastergröße liefere ich die Vorkommensdaten? In Rastern von 100 mal 100 Metern, in Rastern von 500 mal

500 Metern oder 1000 mal 1000 Metern? Das muss auch vereinheitlicht werden, damit man die Daten hinterher gut miteinander kombinieren kann.

[pgg]: Biodiversität ist ja ein super wichtiges Thema. Ich stelle mir gerade vor, wenn alle Landwirte und Großgrundstücksbesitzer mauern würden, was die Vorkommensdaten angeht, da hätte man ja überhaupt gar keine Chance für Deutschland wirklich was zu erheben.

[Ebert]: Ja, das ist tatsächlich auch so, dass diese Akteurs- und Systemlandschaft im Moment sehr zersplittert ist und schon seit Jahren Konsolidierungsansätze diskutiert. Wir haben in dem, was wir in unserem Netzwerk leisten können, den Ansatz, dass wir anhand der Anwendungsfälle zeigen wollen, wie es gehen kann. Also so eine Art Blaupausen entwickeln, die andere kopieren können. Gute Praxis entwickeln, zum Nachahmen anregen, Wege vereinfachen. Und das ebnet vielleicht dann auch einen Weg in eine schnellere Einigung der Akteure über solche Datenzusammenführungen oder über die Bereitstellung offener Daten. Das ist nicht nur ein rechtliches Problem, das ist auch ein Problem, glaube ich, von Kapazitäten. Das sehen wir auf der Länderebene. Das ist generell bei den offenen Daten in den Ländern ein Problem. Offene Daten bereitzustellen, ist viel Arbeit. Man braucht die Ressourcen dafür. Wenn es gute Blaupausen gibt, wenn man das Rad nicht neu erfinden muss, ist das schon mal ein bisschen einfacher zu machen. Der Wille, gerade Daten für Forschung bereitzustellen, ist auf jeden Fall da. Was, glaube ich, fehlt, in dieser speziellen Szene, ist so eine Art Bund-Länder-Zusammenwirken, wie wir es jetzt in der Wissenschaft haben. Die nationale Forschungsdateninfrastruktur wird ja gemeinsam finanziert von Bund und Ländern, weil es eine gesamtstaatliche, überregionale Aufgabe ist. Das ist Paragraph 91 B Grundgesetz. Das ist wahnsinnig hilfreich, weil Bund und Länder schon verschiedene Anliegen und Vorhaben im Bereich der Wissenschaft haben, die sie gemeinschaftlich finanzieren. Nehmen wir jetzt mal das Monitoring im Bereich vom Naturschutz: Das ist Ländersache und es gibt kein Äquivalent zu dem Paragraph 91 B, dass Bund und Länder in diesen Fragen, die ja auch gesamtstaatlich und überregional sind, zusammenwirken. Das heißt, in der Praxis ist das relativ langwierig nach meinem persönlichen Eindruck, hier zu guten Vereinbarungen zu kommen. Und da wird immer noch viel einzeln entschieden. Und für einzelne Arten wird bundesweit schon sehr gut zusammengearbeitet, und bei anderen muss man mit jedem Bundesland einzeln verhandeln, ob die Daten verfügbar sind und wenn ja, in welchem Format. Wir haben jetzt seit zwei Jahren ein nationales Monitoringzentrum zur Biodiversität. Das ist auf Bundesebene finanziert unter Federführung des Umweltministeriums. Dieses Monitoring-Zentrum hat auch die Aufgabe, Harmonisierung mit den Ländern zu besprechen und zu verhandeln, auch im Bereich der Datenbereitstellung. Aber es ist nach meinem Eindruck für ein solches Zentrum viel schwieriger, als es zum Beispiel ist im Bereich der Wissenschaft, wo wir die Gemeinsame Wissenschaftskonferenz haben, wo die Minister von Bund und Ländern gemeinsam solche Dinge entscheiden können.

[mg]: Ich würde noch mal nachfragen, ob ich da auf der falschen Fährte bin. Wenn ich jetzt den Eindruck habe, in dem Fall, wo es um Forschung geht, also Biodiversitätsforschung zum Beispiel, gibt es eine gemeinsame Anstrengung von Bund und Ländern, aber dann, wenn es um Naturschutz geht, dann wieder nicht mehr? Kann man das so trennen, also dass die Ausrichtung oder die dahinter liegende Zielrichtung, sage ich mal, Naturschutz, dann Hürden aufbaut, die für die Forschung nicht bestehen?

[Ebert]: Bund und Länder wirken ja bei der nationalen Forschungsdateninfrastruktur zusammen, finanzieren gemeinschaftlich große Verbundvorhaben, die eben für verschiedene Wissenschaftsbereiche Datendienste gemeinschaftlich entwickeln wollen und auch diese Daten mobilisieren wollen. Das heißt, die bringen da zusammen die Universitäten, die ja Ländersache sind, und die vom Bund finanzierten Forschungseinrichtungen in einer gemeinsamen Finanzierung. In unserem Konsortium geht die Partnergruppe über die reinen Forschungseinrichtungen hinaus. Das heißt, wir haben auch Partner, die Landesämter sind, die Expertenorganisationen im Bereich des Naturschutzes sind und die auch über solche Daten verfügen. Letztlich haben alle Akteure, ob sie nun Naturschutz betreiben oder Forschung betreiben, im Bereich der Biodiversität doch ein ähnliches Interesse an Daten und nutzen auch ähnliche Methoden. Also da gibt es einen ganz großen Überlapp und immer wieder kooperative Projekte. Die Einigung von Bund und Ländern, da im Bereich der Forschungsdaten, die im Wissenschaftsbereich entstehen, zusammenzuwirken und auch darauf hinzuwirken, dass die Daten eben in gemeinsamen Infrastrukturen verwahrt und für eine Nachnutzung erschlossen werden, erleichtert natürlich unser Geschäft. In dem Bereich, wo die Finanzierung aber eine andere ist, wo sie aus anderen Ressorts kommt, können wir nur durch Überzeugung wirken, diese Daten für die Forschung zu mobilisieren. Würde es da ein ähnliches Zusammenwirken geben von Bund und Ländern für gemeinsame Dateninfrastrukturen, würde es vielleicht ein bisschen schneller gehen, sage ich jetzt mal so.

[mg]: Ist da was absehbar? Ist das ein Thema, woran gearbeitet wird oder wo verhandelt wird?

[Ebert]: Ja. Die Akteure, wie gesagt, diskutieren seit Jahren Konsolidierungsansätze. Man sieht, dass diese Global Biodiversity Information Facility schon viel vorangebracht hat, auch mit Bundesförderung. Die haben viel beigetragen dazu, dass gemeinsame Standards entstehen, dass ein Toolkit entsteht, was man benutzen kann, um Daten zu erschließen und zu liefern. Das hat auch der Datenmobilisierung in Deutschland sehr geholfen. Und jetzt eben das neue Monitoringzentrum zur Biodiversität. Das hat natürlich auch die Verfügbarkeit von Daten als eine ganz große Aufgabe, neben der Harmonisierung der Monitoringprogramme, also der fachlichen Seite: Welche Parameter werden erhoben, in welchen Abständen, für welche Tier- und Pflanzenarten usw., also die Harmonisierung der Programme selbst, aber eben auch die Durchlässigkeit der Daten und Dateninfrastrukturen.

[pgg]: Sind dann auch die rechtlichen Fragen in der Diskussion, oder lassen sich diese Hürden wegräumen? Oder wäre das dann noch mal ein anderes Kapitel?

[Ebert]: Es lässt sich nicht alles mit rechtlichen Vorgaben lösen, so viel kann man, glaube ich, sagen. Manches ist durchaus berechtigt. Also auch die Frage nach der Granularität veröffentlichter Daten. Welche Rückschlüsse lassen die zu? Gerade weil es ja auch raumbezogene Daten sind, mit Geoinformationen. Da muss man, denke ich, eine gewisse Unschärfe einfach hinnehmen, weil die Daten schutzwürdig sind. Und auch gesetzliche Verpflichtungen zum Liefern von Daten werden nur funktionieren, wenn die Behörden und die sammelnden Einrichtungen auch über die Ressourcen verfügen, diese Daten dann tatsächlich bereitzustellen.

[mg]: Das heißt, wir landen dann am Ende, vielleicht auch wenig überraschend, dann doch am Ende bei juristischen Überlegungen. Daten, die dann doch schützenswert

sind, tauchen auf einmal auf. Es ist jetzt sozusagen nicht nur eine Frage der Umsetzung und der Ressourcen und des Machenkönnens, sondern da schließt sich der Kreis so ein Stückweit.

[Ebert]: Auch dafür gibt es in der Praxis durchaus Lösungen. Also wenn wir von offenen Daten reden, das ist ja das, was wir gerne möchten. Für die Wissenschaft ist, dass die Daten tatsächlich öffentlich, online und kostenfrei zur Verfügung stehen, sodass man relativ schnell an Datensätze herankommt. Die können dann gern ein bisschen grober sein. Es gibt durchaus Modelle, auch sensitive Daten, auch sogenannte Mikrodaten heißt das in den Wirtschaftswissenschaften, für die Forschung zur Verfügung zu stellen, dann aber eher in geschützten Räumen. Die wären dann nicht offen, sondern die werden in Datenzentren bereitgestellt, wo man einen Antrag stellen muss, wo man sein Forschungsvorhaben schildert, wo man dann aber Zugang bekommt unter einigen Auflagen, die dann eben zum Beispiel die Datensubjekte schützen. Da gibt es durchaus Modelle für, das behindert eigentlich Forschung nicht, aber auch solche Modelle muss man dann erstmal etablieren. Die sind in dem Bereich der Biodiversitätsforschung noch nicht implementiert.

[pgg]: Ich springe noch mal zu einer anderen Frage. Sie hatten vorhin ja auch das Herbarium schon als Beispiel erwähnt, mit der getrockneten Pflanze, die da gesammelt ist – ganz klassisch, so auf Papier. Wie sieht es überhaupt aus mit den stofflichen Originalen, also den nicht digitalen Bestandteilen dieser riesigen Biodiversitätssammlungswelt? Sind die von Bedeutung oder kann man die einfach abhängen?

[Ebert]: Ja, die sind schon von Bedeutung. Zum einen sind es ja biologische Materialien. Also wenn man noch mal Analysen der Erbsubstanz machen möchte, dann braucht man das biologische Material, um das zu extrahieren. Die werden zum Teil digitalisiert, aber nur Bruchteile der vorhandenen Sammlungen. Und für eine ganze Reihe von Untersuchungen braucht man Lebendmaterial. Das heißt, wir haben auch Sammlungen, zum Beispiel die Deutsche Sammlung von Mikroorganismen und Zellkulturen. Die vermehren laufend ihren Bestand und geben dann lebendes Material ab für weitere Forschungsfragestellungen oder für Untersuchungen. Und das kann man natürlich auch nicht durch digitale Abbildung oder Reproduktion ersetzen.

[mg]: Würden Sie dann auch eine Plattform bieten, um dann solchen Austausch auch zu vermitteln? Also ist das auch was, was dann so eine Dateninfrastruktur leistet? Ich stelle mir vor, das wäre wie so ein Onlinehandel: Ich brauche meine digitalen Daten oder ich brauche irgendwie eine Zellkultur oder ich brauche vielleicht auch ein Stück von einem Präparat. Wird das zusammengedacht, oder sind das dann andere Angebote?

[Ebert]: Also die Netzwerke dafür gibt es zum Teil schon. Und das wäre, glaube ich, dann wieder etwas, was man bei dem Zuschnitt der Endnutzerdienste berücksichtigen muss. Ist das eine Forschungscommunity, die eben auch Bedarf hat, Materialien auszuleihen? Also es gibt da richtig auch einen Leihverkehr wie bei Büchern. Oder Teile der Originalmaterialien zu erhalten, um dann eben auch die entsprechenden Informationen, die Gateways, zu schaffen, zu den Sammlungen, wo man diese Materialien eben anfragen kann. Zum Teil sind die Informationen bei einem Datensatz zu einem Artenfund dabei. Da steht dabei: Es gibt ein physisches Exemplar. Und dann

steht auch dabei, welche Sammlung das hält, so dass man dort anfragen kann. Also das ist in den Metadaten durchaus drin.

[pgg]: Dann ist es eigentlich schon so, dass der Hinweg sozusagen auch zum echten – echt ist jetzt falsch gesagt – also zum physischen Objekt oder vielleicht sogar zum lebenden Objekt, der Hinweg jedenfalls ist dann auch schon angegeben und den Pfad könnte man beschreiten?

[Ebert]: Ja. Also die GFBio-Datenzentren, die haben eine Typologie der biologischen Daten entwickelt, aus der man relativ schnell entnehmen kann: Gibt es dazu ein physisches Objekt zu diesem Datensatz oder nicht? Bei den Biodiversitäts- und Vorkommensdaten, das sind sogenannte Typ 1 Daten in unseren Datenzentren. Und Typ 1A sind Daten aus Sammlungen, die eine Referenz zu einem physischen Objekt haben, und Typ 1B Daten sind Beobachtungsdaten ohne eine solche Referenz, also zum Beispiel eine Vogelbeobachtung, die einfach per App gemeldet wurde.

[pgg]: Das leuchtet ein, dass man die Vögel nicht irgendwo abrufen kann als Vögel.

[Ebert]: Dass man die nicht alle einsammeln kann, genau. Dann gibt es taxonomische Daten. Das sind Kataloge, rote Listen oder Artenchecklisten. Und Typ 3 Daten, das sind dann die umweltbezogenen oder ökologischen Daten. Physiologische Informationen zu Bakterien, zum Beispiel zu Stoffwechselwegen. Und Typ 5 Daten – ich springe hier mal –sind die molekularen Sequenzdaten. Und mit dieser Typologie kann man relativ schnell sich dann orientieren in diesen Datenkatalogen. Dazu kommen dann allerdings noch Bildmaterialien aller Art bis hin zu Drohnenaufnahmen und Infrarotsensoren, Bildern aus Infrarotsensorik, wenn mit Infrarotkameras Tiere fotografiert werden. Auch Modelle, Modellrechnungen gibt es in der Biodiversitätsforschung. Es ist ein breites Feld – also es kommt einfach mit der Methodik, die da angewandt wird.

[mg]: Gibt es irgendeine Möglichkeit zu sagen, wie weit man ist, mit vielleicht auch dem Aufholen vom Integrieren bestehender Sammlungen? Also, es kommt ja immer noch was dazu. Man kann jetzt ja nicht sagen, das ist irgendwann mal fertig. Aber es gibt vielleicht irgendwie einen Moment, wo man sagt, jetzt sind wir relativ nah dran. Kann man das irgendwie messen, wie der Fortschritt ist?

[Ebert]: Ich habe mich das auch gefragt. Ich war in Vorbereitung unseres Gesprächs noch mal in der Global Biodiversity Information Facility, um rauszufinden, wie viel Vorkommensmeldungen haben die eigentlich da? Und das sind über 2.000.000.000 von 2400 meldenden Einrichtungen weltweit. Das klingt erstmal viel, aber wenn man sich vorstellt, dass das zum Teil Artvorkommen sind, die historisch belegt sind, also aus dem 18. Jahrhundert kommen bis heute und das den ganzen Globus umspannt, ist es vielleicht doch nicht so viel. Es ist nach wie vor eigentlich Stückwerk, was wir wissen, welche Art wo vorkommt. Und die Forschenden finden mit neuen, DNA basierten Methoden zum Beispiel immer noch Spuren von Arten in Gewässern, die man da nie gesichtet hat, die man nur anhand ihrer DNA-Spuren in dem Gewässer identifizieren kann. Also da sind einfach noch sehr, sehr, sehr viele Lücken. Was vielleicht ein guter Gradmesser ist, ist: Wie viele Einrichtungen melden denn Vorkommensdaten? Wie viele Einrichtungen schicken Datenpakete an gemeinsame Dateninfrastrukturen? Da finde ich jetzt 2400 weltweit auch nicht so wahnsinnig viel. Da hätten wir in Deutschland wahrscheinlich schon einige 100, die das tun könnten. Das wird für unser Netzwerk zu den guten Nachrichten gehören, wie viele Einrichtungen wir gewinnen

konnten, nach einer abgestimmten Systematik, nach einem abgestimmten Kerndatensatz, Daten zu melden an die gemeinsamen Infrastrukturen.

[pgg]: Sind denn in der Laufzeit des Projekts, also Sie haben den Sprung von 20 zu 50 schon erwähnt, aber sind jetzt im Moment schon neue in Sicht, die mit einsteigen?

[Ebert]: Also wir sind dabei, weitere Datengeber mit diesem Kerndatensatz vertraut zu machen. Die Toolbox, also die Tools, die zur Verfügung stehen, um diese Datenmeldungen zu vereinfachen, auch bei anderen Einrichtungen zu implementieren. Da sind insbesondere die deutschen GBIF-Zentren. Also es gibt benannte Zentren in Deutschland, die sich kümmern um die Mobilisierung von Daten, von Vorkommensdaten, für diese gemeinsame, für diese globale Biodiversitätsinformation Facility. Die sind da sehr aktiv und die sind auch alle Mitglied bei uns im Konsortium, und wir nutzen die einfach als Multiplikatoren, um dieses Wissen auch in andere Einrichtungen zu bekommen und sie zu befähigen, die Daten dann auch an die gemeinsamen Infrastrukturen zu geben.

[mg]: Das ist jetzt nur mal so Neugierde, aber wissen Sie vielleicht von Forschungsfragen, die darauf warten, dass Sie weiterkommen mit bestimmten Schritten? Oder vielleicht auch Umweltschutzanliegen, wo man sagt, da bräuchten wir dringend mehr Daten für.

[Ebert]: Ich kann jetzt nur berichten, was sozusagen in meiner Zeit mir über den Schreibtisch ging in meiner Position. Ich bin ja noch relativ neu dabei. Aber was für mich so ein kleines Schlüsselerlebnis war, was zeigt, wie entkoppelt diese ganzen Bemühungen sind, ist, dass wir Kontakt hatten mit einem großen Biosphärenreservat, bei denen diverse Forschungsprojekte laufen: Forschungsprojekte, die bei Ihnen zu Gast sind, Forschungsprojekte, die sie selber anstoßen. Aber sie haben die Daten nicht gut verfügbar, die aus diesen Projekten entstanden sind. Das heißt, Ihre eigene Inventarisierung in diesem Biosphärenreservat wird durch diese Forschungsprojekte gar nicht befördert. Und das wäre natürlich anders, wenn die Daten aus solchen Projekten systematisch in Dateninfrastrukturen landen. Durch die Mitlieferung der Fundorte kann man ja relativ schnell dann auch rausfiltern, was für das eigene Gebiet relevant ist, und hat dann eben auch die Mehrwerte aus diesen ganzen individuellen Forschungsprojekten für das eigene Schutzgebiet. Das war für mich so ein Schlüsselerlebnis, wo ich dachte: Mensch, das müsste doch eigentlich zu machen sein.

[pgg]: Das heißt, dadurch, dass bundesweit gewissermaßen oder flächendeckend, es ist ja nicht nur der Bund, sondern überhaupt flächendeckend erfasst und kartiert wird, kann dann nachträglich spätestens auch eine bestimmte Region oder ein bestimmtes Reservat oder Naturschutzgebiet oder so seine Daten rausziehen und hat sich damit nachträglich dann kartiert oder kartieren lassen oder erfassen lassen, hat seine Bestände, ohne für die eigene Fläche tätig zu werden, dann schon gezählt bekommen?

[Ebert]: Und kann das kombinieren vielleicht mit eigenen Anstrengungen. Genau. Ich glaube, im Endeffekt kann man noch eine Sache sagen über die Vorteile solcher gemeinsamen Dateninfrastrukturen. Sie sparen einfach unendlich viel Zeit. Also was die Forschenden einheitlich berichten, ist: Man kommt schon auch anders an die Daten ran, aber man muss jeden Datengeber einzeln anschreiben, man muss mit ihm verhandeln. Manchmal ist es Glückssache, ob das klappt, in der Laufzeit des Projekts

alle Daten zusammenzubekommen, die man dafür braucht. Da einfach klarere Zugangswege zu schaffen, auch wenn die Daten vielleicht nicht immer offen sind. Aber dass man gute Regeln hat, wie man an die Daten dieser Datengeber als Forscher mit einem legitimen Interesse herankommen kann, das ist, glaube ich, etwas, was die Nationale Forschungsdateninfrastruktur wirklich gut befördern kann. Und sie kann es auch den Datengebern erleichtern, solchen Bitten entgegenzukommen, denn die verursachen ja auch Arbeit bei den Datengebern und die haben durchaus Interesse daran, die Daten herauszugeben. Aber manchmal scheitert es an den Ressourcen. Also je besser wir das Datenmanagement dort auch voranbringen, desto leichter wird das dann für die Forschung, an die Daten in einer vernünftigen Zeit heranzukommen, damit sie mehr Zeit für die Analysen aufwenden können als für die Jagd nach dem Datensatz, den sie brauchen.

[Der Abspann mit Musik beginnt.]

[mg]: Und damit ist dieses Digitalgespräch zu Ende und wir bedanken uns bei Barbara Ebert von der Gesellschaft für biologische Daten für die spannenden Eindrücke und die interessante Diskussion. Viele Grüße nach Bremen. Und wie immer auch vielen Dank an Sie, liebe Zuhörerinnen und liebe Zuhörer, dass Sie uns wieder Ihr Interesse und Ihre Aufmerksamkeit gewidmet haben. Das Digitalgespräch macht jetzt eine kleine Pause und setzt einmal aus. Wir hören uns dann, wenn Sie mögen, wieder am 16. Mai, zur nächsten Folge des Digitalgesprächs, dem Podcast von ZEVEDI, dem Zentrum verantwortungsbewusste Digitalisierung.