

Peter Uetz

Alte Bücher - neue Technik

Das Projekt "Historische Literatur" bei der EMBL-Reptiliendatenbank

In der Zoologie ist die historische Literatur schon wegen des Prioritätsprinzips wichtig: die erste Beschreibung ist massgebend für die Definition einer Art, zumindest wenn sie nach 1758 erfolgt ist. In diesem Jahr wurde die zehnte Auflage von Linné's *Systema Naturae* veröffentlicht, die nicht nur die binäre Nomenklatur einführte, sondern als Markstein der zoologischen Systematik überhaupt gilt.

Historische Quellen haben jedoch auch aus weniger offensichtlichen Gründen Bedeutung erlangt. Wegen der zunehmenden Naturzerstörung werden Berichte aus den vergangenen Jahrhunderten zunehmend zu Quellen über ursprüngliche Verbreitungsgebiete, in denen die beschriebenen Arten oft schon lange nicht mehr existieren. Nicht selten wurden Arten aufgrund einzelner Exemplare beschrieben, die dann nie wieder gefunden wurden. Im Falle kleiner Populationen kann nämlich die Abholzung eines kleinen Wäldchens schon zum Austerben der Art führen.

In vielen Fällen droht alten Büchern ein ähnliches Schicksal. Durch kleine Auflagen, säurehaltiges Papier oder Weltkriege sind viele Bücher in ihrer Existenz bedroht und oftmals nur noch in streng behüteten Einzelexemplaren vorhanden. Zum Erhalt dieser Werke würden sich Nachdrucke anbieten, die aber aus Kostengründen oft nicht den grossen Aufwand wert sind. Aus diesem Grund wenden sich viele Bibliotheken jetzt der Digitalisierung zu. In der Tat haben die meisten Industrieländer mittlerweile mehr oder weniger umfangreiche Programme zur elektronischen Speicherung ihrer historischen Quellen aufgelegt. In Deutschland wird ein solches Programm von der Deutschen Forschungsgemeinschaft organisiert und gefördert.

Digitale Bibliotheken

Die Computerspeicherung von alten Arbeiten hat mehrere Vorteile: Zunächst muss das Original nur ein Mal eingescannt werden; danach kann es guten Gewissens in einem Safe oder einem sonstigen sicheren Ort verwahrt werden. Die Originale werden also nach dem Scannen vor weiterer Abnutzung verschont. Die elektronischen Kopien sind danach ohne Qualitätsverlust beliebig vervielfältigbar. Vor allem können sie aber leicht verfügbar gemacht werden: während man an viele Originale wegen ihrer Empfindlichkeit schon gar nicht mehr herankommt, kann man elektronische Kopien leicht über Disketten oder CD-ROMs verbreiten. Werden sie im Internet zur Verfügung gestellt, sind sie sogar rund um die Uhr weltweit zugänglich. Elektronische Dokumente werden zumeist in einem Standardformat verbreitet, sodass sie praktisch auf allen Computern gelesen werden können. Beispiele hierfür sind TIFF-Dateien für Bilder und Text"abbildungen" oder ASCII für computerlesbaren Text (werden Textseiten als digitale Photos abgespeichert ist der Text natürlich nicht "computerlesbar" im engeren Sinne). Als Format für kombinierte Dateien hat sich mittlerweile das *Portable Document Format* (PDF) durchgesetzt, ein Format, das die Firma Adobe entwickelt hat. Die Computerspeicherung von Texten, insbesondere wenn sie direkt computerlesbar sind, erlaubt vor allem eine effiziente Indexierung der wissenschaftlichen Literatur. Man erinnere sich daran, dass bis vor wenigen Jahrzehnten alle Register von Hand erstellt wurden! Selbst die alphabetische Sortierung von Stichworten oder Autoren musste von Hand vorgenommen werden. Es ist klar, dass diese mühselige Arbeit nicht selten zu unvollständigen Registern geführt hat. Elektronische Register sind nicht nur leichter zu erstellen, sondern vor allem leichter zu durchsuchen. Das weiss jeder, der noch mit alten Zettelkatalogen

gearbeitet hat: wenn man zum Beispiel den Namen eines Autors nicht mehr genau wusste, war man in der Regel schon verloren. Eine Alternativsuche nach Stichworten war in der Regel vergebens, da Stichwortkataloge extrem aufwendig zu erstellen und schon von daher praktisch immer unvollständig sind. Noch gravierender wird der Nachteil nicht-elektronischer Kataloge, wenn man Suchkriterien kombinieren will: Sie müssen dazu nur in einem Zettelkatalog nach allen Arbeiten über die Afrikanische Herpetofauna vor 1900 suchen; ein beinahe hoffnungsloses Unterfangen. In einer ordentlichen Datenbank ist das eine Frage von Sekunden.

Schliesslich sind elektronische Bibliotheken auch eine Frage des Platzes. Auf einer CD-ROM kann man leicht 100,000 eng bedruckte Seiten Text unterbringen. Die EMBL-Reptiliendatenbank enthält bereits ca. 3000 Seiten Text, deren Ausdruck wenig sinnvoll erscheint, da man zusätzliche Register zur Erschliessung mit ausdrucken müsste, was zu einem vielfachen Papierverbrauch führen würde.

Es kann natürlich nicht verschwiegen werden, dass elektronische Bibliotheken auch einige **Nachteile** haben. Nicht jeder kann sich die notwendige Hardware samt Internetanschluss leisten, vor allem Benutzer in Ländern der dritten Welt sind nach wie vor von der technischen Entwicklung abgekoppelt. Allerdings können sich diese Länder in aller Regel auch keine guten Bibliotheken leisten und ein Internetanschluss ist fast immer deutlich billiger als die historische Originalausgabe eines Buches. Der Hauptnachteil elektronischer Dokumente ist freilich, dass man überhaupt ein Gerät braucht, um sie lesen zu können. Gerade bei historischen Arbeiten wird das Problem aber dadurch abgemildert, dass man sie aber sowieso nur an bestimmten Orten einsehen kann und nicht unbedingt in der Strassenbahn oder in der Badewanne liest.

Hingegen wird oft kritisiert, dass elektronische Datenträger und Formate viel vergänglicher sind als Papier. Es ist wahr – niemand weiss, wie lange eine CD wirklich hält und vor allem: wie lange es noch Lesegeräte dafür gibt. Schon seit längerem werden keine 5 1/4-Zoll-Laufwerke mehr in Computer eingebaut. Die entsprechenden Disketten können damit nicht mehr gelesen werden. CD-ROMs werden bereits durch DVDs ersetzt und durch neue Versionen der Computersoftware sind viele ältere Datei-Formate nicht mehr lesbar. Gegen dieses Problem hilft nur das regelmässige Umkopieren auf neue Datenträger, z.B. von CDs auf DVDs. Dateiformate sind meines Erachtens vor allem dann ein Problem, solange sie nicht standardisiert sind. Es wird jedoch noch lange TIFF, PDF und natürlich ASCII-Dateien geben. Nicht zuletzt weil diese Formate gut dokumentiert sind und deshalb auch relativ leicht in neuer Software implementiert werden können. Das ist für bestimmte proprietäre Formate ganz anders. Zum Beispiel ist das "Word"-Format von Microsoft ein recht gut behütetes Geheimnis von Microsoft, das schon wegen des Patentschutzes von kaum einer anderen Software verarbeitet werden kann. Wenn also Microsoft beschliesst, dass Word ältere Dateiversionen nicht mehr lesen können muss, könnte das katastrophale Folgen für Archivare haben.

Das EMBL-Projekt "Historische Literatur"

Die Idee, historische Arbeiten einzuscannen, entstand bei der Literatursuche in der Universitätsbibliothek von Seattle, USA, wo ich als Postdoktorand tätig war. Obwohl die dortige Bibliothek eine exzellente Zeitschriftensammlung hat, die auch am Wochenende und abends frei zugänglich ist, erforderte das Nachsehen einzelner Artikel doch jedesmal einen Fussweg von 10 min von meinem Arbeitsplatz zur Bibliothek. Für jedes Nachschlagen war ich also gut 20 min unterwegs. Photokopien wären zwar eine Möglichkeit, kosten aber auch Geld und brauchen Platz, was ein

nicht zu unterschätzendes Problem ist, wenn man eine grössere oder systematische Sammlung anlegen will.

In den meisten deutschen Universitätsbibliotheken werden alte Schriften oft in Magazinen aufbewahrt, aus denen sie bestellt werden müssen oder erst gar nicht ausgeliehen werden können. Ausserhalb der Öffnungszeiten sind sie gar nicht erreichbar.

Stand der Sammlung

Wegen der erheblichen Anzahl von Originalbeschreibungen konzentrierten sich meine Recherchen zunächst auf 3 Zeitschriften: die *Proceedings of the Academy of Natural Sciences of Philadelphia* (PANSP), die *Annals and Magazine of Natural History* (AMNH) und die *Proceedings of the Zoological Society* (London). Die mehr als 500 Arbeiten über Reptilien aus den ersten beiden Zeitschriften wurden bis zum Jahr 1900 komplett eingescannt und sind im Internet (PANSP) oder auf CD erhältlich (AMNH).

Urheberrechtliche Aspekte

Das Jahr 1900 wurde aus urheberrechtlichen Gründen als Zeitgrenze ausgewählt. Publikationen vor 1900 sind nach internationalen Vereinbarungen nicht mehr urheberrechtlich geschützt, obwohl die Gesetzeslage nicht völlig einheitlich ist. Für die *Proceedings* erhielt ich allerdings die offizielle Erlaubnis, die Arbeiten im Internet zu veröffentlichen. Der Verlag Francis und Taylor in London, der die *Annals* publizierte, hat bisher nicht auf meine Anfragen reagiert, was allerdings vor allem für die Artikel nach 1900 wichtig wäre.

Von der Bibliographie zur Faktendatenbank

Obwohl die Digitalisierung von Büchern und Zeitschriften ihre Benutzung und Erschliessung in vielerlei Hinsicht erleichtert, hat sie für wissenschaftliche Inhalte weniger eine erhaltende Funktion. In der wissenschaftlichen Literatur ist Papier ja nur der Träger für Informationen. Wichtiger als der Träger ist aber die Information selbst, vor allem ihre Wiederauffindbarkeit und ihre Verknüpfbarkeit. So werden in der zoologischen Literatur zumeist einzelne Arten beschrieben. Wenn man aber alle Arten einer bestimmten Region oder solche mit bestimmten Merkmalen sucht, kommt man nicht darum, viele Artikel durchzusehen (wenn man nicht das Glück hat, eine Monographie zum Thema zu finden). Aus diesem Grund werden gut organisierte Datenbanken immer unerlässlichere Hilfsmittel: Sie lassen sich nicht nur gezielt und schnell nach Daten durchsuchen, sondern sind auch immer aktualisierbar. Die Aktualisierbarkeit ist besonders wichtig. Wenn z.B. eine Schlangengattung 3 Arten enthält, kann eine neubeschriebene vierte Art alle vorherigen Artdiagnosen über den Haufen werfen. Das heisst, es wird nicht nur eine neue Art hinzugefügt, sondern die älteren Artdefinitionen müssten auch präzisiert und damit aktualisiert werden.

In der Tat ist es ein Fernziel der EMBL-Reptiliendatenbank alle Artbeschreibungen zu sammeln und in einer durchsuchbaren Datenbankstruktur zu speichern. Eine solche Datenbank könnte dann leicht als Bestimmungstool verwendet werden. Die gedruckte Literatur muss dazu jedoch in strukturierte, computerlesbare Datensätze überführt werden. Es ist offensichtlich, dass dies einen riesigen Aufwand erfordern wird. Aus diesem Grund habe ich mich in der ersten Phase der Datenbankentwicklung darauf konzentriert, alle Originalzitate und einige weitere wichtige Literaturzitate zu sammeln, z.B. Revisionen und Checklisten (wegen der Verbreitungsangaben). In der

nächsten Stufe kann jetzt versucht werden, diese Arbeiten einzuscannen bzw. die Daten daraus zu extrahieren.

Wohlweislich hat sich die EMBL-Reptiliendatenbank auf die taxonomische Literatur konzentriert. Die Aufnahme von Daten zum Verhalten, zur Ökologie, oder zur Terrarienhaltung würden die Kapazitäten bei weitem sprengen. Es sei deshalb nochmals betont, dass die Reptiliendatenbank nie eine Bibliographie sein wollte (dazu gibt es den Zoological Record), sondern anstrebt, eine Faktendatenbank zu sein, die nicht nur Zitate sondern Daten aus denselben enthält.

Aufruf zur Mitarbeit

Es ist klar, dass das Arbeitsprogramm für die Taxonomie-Datenbank nicht von einer einzelnen Person geleistet werden kann. Trotz der zahlreichen "Gelegenheits"mitarbeiter fehlen der Datenbank feste Mitarbeiter mit klar definierten Aufgaben. Dieser Mangel soll in naher Zukunft durch die Einstellung von Diplomanden oder Doktoranden gelindert werden.

Trotzdem will ich Sie hiermit daran erinnern, dass die Mitarbeit von Freiwilligen notwendig ist. Insbesondere sind wir auf der Suche nach historischer Originalliteratur, die sich zum Scannen eignet. Leihgeber solcher Literatur erhalten im Gegenzug Kopien der bisher eingescannter Schriften einschliesslich einer Kopie der Datenbank, die diese Arbeiten indiziert und somit erschliesst. Ich bin gespannt auf Ihre Rückmeldung!

Literatur

Zimmer, D.E.(2000) Die Bibliothek der Zukunft. Text und Schrift in den Zeiten des Internet. Hoffmann und Campe, Hamburg.

EMBL-Reptiliendatenbank: <http://www.reptile-database.org>

Autor: Peter Uetz, Institut für Genetik, Forschungszentrum Karlsruhe, Postfach 3640, 76021 Karlsruhe, e-mail: peter@uetz.de