

# Ein verteiltes Informationssystem zur Forschungs Kooperation in der Bioakustik

Rolf Bardeli<sup>1</sup>, Michael Clausen<sup>1</sup>, Karl-Heinz Frommolt<sup>2</sup> und Frank Kurth<sup>1</sup>

<sup>1</sup> Institut für Informatik III, 53117 Bonn, Deutschland, Email: {bardeli, clausen, frank}@iai.uni-bonn.de

<sup>2</sup> Museum für Naturkunde, 10115 Berlin, Deutschland, Email: karl-heinz.frommolt@rz.hu-berlin.de

## Einleitung

In einem zweijährigen DFG-Projekt wird in einer Kooperation zwischen dem Tierstimmenarchiv der Humboldt-Universität zu Berlin und dem Institut für Informatik der Universität Bonn eine webbasierte Plattform zum Austausch bioakustischer Daten geschaffen. Diese soll sowohl einen komfortablen Zugriff auf die digitalisierten Aufnahmen aus Tierstimmenarchiven erlauben, als auch die langfristige Archivierung und Verfügbarkeit von Feldaufnahmen ermöglichen. Eine wichtige Rolle spielt dabei die verteilte Eingabe und Haltung von Daten.

## 1 Das Tierstimmenarchiv in Berlin

Das Tierstimmenarchiv an der HU Berlin [2] ist weltweit eine der ältesten und umfangreichsten Sammlungen an Lautäußerungen von Tieren. Es wurde im Oktober 1951 von Prof. Günter Tembrock am damaligen Zoologischen Institut der Humboldt-Universität zu Berlin begründet.

Das Datenmaterial besteht derzeit aus ca. 110.000 Tonaufnahmen auf mehr als 4.500 Tonbandspulen, DAT-Kassetten oder CDs. Dabei wird mit 1.800 Vogelarten, 580 Säugetierarten, mehr als 150 Arthropodenarten sowie einigen Fisch-, Amphibien- und Reptilienarten ein breites Artenspektrum abgedeckt. Dieser Datenbestand wird seit nunmehr 3 Jahren digitalisiert. Zur Erhaltung höchstmöglicher Qualität wird die Digitalisierung mit Studionorm (96 kHz, 24 bit) vorgenommen. Die textuelle Annotation der digitalisierten Aufnahmen wird innerhalb einer MS-Access Datenbank vorgenommen, in der zu jeder Aufnahme Informationen über Tierart, Aufnahme datum, Aufnahmeort, usw. erfaßt werden. Momentan sind ca. 17.000 Datensätze in die Datenbank übernommen, und ca. 1.500 Aufnahmen in Studionorm digitalisiert. Mit dem vorläufigen Abschluß der Digitalisierung wird in ca. 5 Jahren gerechnet. Es wird dann Datenmaterial im Umfang von etwa 5.000 Stunden (ca. 10 Terabytes) digital vorliegen.

## 2 Motivation

Bei der Suche nach Tierstimmenmaterial ergeben sich in der Praxis ganz allgemein zahlreiche signifikante Probleme. Zum einen existiert bisher keine zentrale Erfassung, an welchem Ort welches Material vorhanden ist. Entsprechend steht auch keine allgemeine Suchfunktionalität (etwa per Internet) zur Verfügung. Zum anderen geschieht die Informationsbeschaffung häufig über persönliche Kontakte mit den wenigen vorhandenen Experten und Kontaktpersonen der Archive. Erschwerend

kommt hinzu, daß der vorhandene Datenbestand nur zum Teil digitalisiert vorliegt, was den zügigen Austausch stark behindert und auch den Zugriff der Archive auf die eigene Sammlung erschwert. Das analoge Material, etwa auf Tonbändern, nutzt sich durch wiederholten Gebrauch ab und droht mit der Zeit verloren zu gehen. Zahlreiche Aufnahmen werden außerdem ausschließlich für spezielle Projekte aufgenommen und sind für die Allgemeinheit weder auffindbar noch zugänglich.

Daher besteht der Bedarf nach einem Informationssystem, das es erlaubt sich schnell einen Überblick über vorhandenes Tonmaterial zu verschaffen und dieses möglichst leicht zu beziehen.

## 3 Ziele

Das wichtigste Ziel der hier vorgestellten Initiative ist die Schaffung einer Infrastruktur zur *kooperativen Datennutzung*. Diese soll einen schnellen Zugriff kooperierender Forschergruppen in Bioakustik und Naturschutz auf gemeinsame Aufnahmen und Metadaten realisieren. Dieser Zugriff soll durch ein *webbasiertes Informationssystem* ermöglicht werden, das über ein Webinterface Datenverwaltung, Daten- und Benutzersichten einheitlich zur Verfügung stellt.

Ein weiteres Ziel ist die *Erweiterbarkeit* durch Daten externer Nutzer. Die Basis hierzu stellt eine *Benutzerverwaltung* dar, die Benutzergruppen mit unterschiedlichen Zugriffsberechtigungen erlaubt.

Um die wissenschaftliche Verwendbarkeit der verwalteten Daten zu sichern, ist die *Qualitätskontrolle* der Daten von großer Bedeutung. Eingehende Datensätze sollen deshalb zunächst einer Integritäts- und Qualitätssicherung durch Experten unterzogen werden, bevor sie in den Datenbestand aufgenommen werden.

Um eine hohe Akzeptanz des Systems zu erzielen ist die *Benutzerfreundlichkeit* von hoher Bedeutung. Um diese zu erreichen ist die Verwendbarkeit des Systems auch mit wenigen technischen Kenntnissen sicherzustellen. Wichtig sind in dieser Hinsicht auch kurze Wartezeiten bei der Anfragebearbeitung.

Die *Datenverknüpfung* mit zugehörigem Material wie etwa Literatur, Bildern oder Videos soll die Einordnung der Daten in den wissenschaftlichen Kontext erleichtern.

Um möglichst leicht möglichst viele Datenbestände über die Infrastruktur erreichbar zu machen, wird eine *verteilte Nutzung* angestrebt. Import- und Exportmöglichkeiten

sollen die Anbindung externer Datenbanken bei verteilter Datenhaltung ermöglichen. Dabei ist die Wahrung der Rechte an eigenem Material häufig wünschenswert und soll auf diese Weise vereinfacht werden.

Traditionelle *Suchfunktionalitäten* anhand von Metadaten sollen zur Verfügung gestellt werden und langfristig durch eine offene Schnittstelle für Erweiterungen (z.B. für akustische Mustererkennung) flankiert werden.

Besonderer Wert wird auf die *Modularität* des Systems gelegt. Durch eine offene Schnittstelle für Such- und Analyse-Algorithmen wird dabei eine weitreichende Erweiterbarkeit angestrebt.

Schließlich sollen *Signalverarbeitungs- und Visualisierungskomponenten*, wie zum Beispiel die Spektrogrammdarstellung und Audio-Wiedergabe, Filterung, sowie halbautomatische Extraktion relevanter Aufnahmeausschnitte, die schnelle Bewertung des vorhandenen Materials erlauben.

## 4 Zwischenstand

Nach dem Anlauf des Projekts im Sommer 2004 ist inzwischen eine Testversion des Informationssystems im Netz verfügbar und stellt unter anderem folgende Funktionalitäten zur Verfügung:

- Das Modul zur Benutzerführung ist im wesentlichen fertig gestellt. Es können über das Webinterface Benutzer verwaltet werden und zum Beispiel Rechte für verschiedene Aufgaben, sowie Dateibenutzungsrechte vergeben werden.
- Eine umfassende Metadatensuche über das Web wird durch eine MySQL-Datenbank ermöglicht. Zu allen Suchergebnissen lassen sich (abhängig von Benutzerrechten) Details einsehen, Aufnahmen anhören, herunterladen usw.
- Über ein Webinterface lassen sich bequem Metadaten eingeben und als Vorschlag in die Datenbank aufnehmen. Die Audiodaten können auf einem beliebigen angeschlossenen Dateiserver abgelegt werden.

## 5 Einsatz-Szenarien

Anhand von drei Szenarien soll nun kurz auf Möglichkeiten zur Nutzung der entstehenden Infrastruktur eingegangen werden.

Das Monitoring biologischer Lebensräume und der darin vorkommenden Tierarten ist eine wichtige Informationsquelle für den Naturschutz und ist in internationalen Verträgen als Berichtspflicht festgeschrieben. Die Durchführung solcher Aufgaben basiert häufig auf Zählungen akustisch aktiver Tierarten und deren Identifikation meist ausschließlich anhand ihrer charakteristischen Lautäußerungen. Zur sicheren Identifikation ist leicht verfügbares Referenzmaterial eine unschätzbare Hilfe [4].

Bei der Beschreibung neuer Tierarten spielen Lautaufnahmen häufig eine wichtige Rolle [3]. Mitunter werden

neue Unterarten sogar erst durch Merkmale ihrer Lautgebung erkannt. Ein leichter Zugriff auf Referenzaufnahmen untersuchter Tierarten ist deshalb wichtig. Dieser kann durch das vorgestellte Informationssystem deutlich erleichtert werden.

Die vorgestellte Plattform erlaubt schließlich auch den Einsatz in der Lehre. Sie ermöglicht den leichten Zugriff auf Materialien für die Ausbildung in Zoologie und Bioakustik.

## 6 Ausblick

Mit Fertigstellung der Grundfunktionen der Infrastruktur können komplexere Funktionen in Angriff genommen werden. Ein interessanter Punkt ist hier die Realisation inhaltsbasierter Suche. Es soll ermöglicht werden anhand von durch Benutzer vorgegebenen „Beispielaufnahmen“ alle Aufnahmen, die Lautäußerungen der gleichen Tierart oder strukturell ähnliche Lautungen enthalten, zu finden. Hierzu sollen bestehende Techniken aus dem Musikretrieval [1] und der Mustererkennung auf das vorliegende Szenario übertragen werden.

Weiterhin ist die Verwendung des Audiomaterials zur Tierstimmenerkennung angedacht. Die Aufnahmen sollen durch die Benutzer so annotiert werden können, daß daraus automatisch Trainingsmengen für Klassifikatoren gebildet werden können. Diese ließen sich dann zum Beispiel unterstützend im Tiermonitoring verwenden.

Andersherum soll die Annotation der Audiodaten durch Klassifikationsalgorithmen unterstützt werden können. Beispielsweise könnten automatisch Positionen gesprochener Kommentare und von Tierstimmenaktivität gefunden werden.

Zunächst ist jedoch die Anbindung erster externer Datenbestände geplant.

## Literatur

- [1] M. Clausen and F. Kurth. A unified approach to content-based and fault tolerant music recognition. *IEEE Transactions on Multimedia*, 6(5):717–731, October 2004.
- [2] Humboldt-Universität zu Berlin. Tierstimmenarchiv. <http://www.biologie.hu-berlin.de/~tsarchiv/>.
- [3] J. Martens. Vocalizations and speciation of palearctic birds. In D.E. Kroodsma and E.H. Miller, editors, *Ecology and evolution of acoustic communication in birds*, pages 221–240, Ithaca and London, 1996. Cornell University Press.
- [4] R. Rempel, K. Hobson, G. Holborn, S. Van Wilgenburg, and J. Elliott. Bioacoustic monitoring of forest songbirds: interpreter variability and effects of configuration and digital processing methods in the laboratory. *Journal of Field Ornithology*, 76(1):1–11, 2004.