Ein verteiltes Medienarchiv für bioakustische Datenbestände

Rolf Bardeli[1], Michael Clausen[1], Karl-Heinz Frommolt[2], Frank Kurth[1],
[1]{bardeli, clausen, frank,}@cs.uni-bonn.de
[2]karl-heinz.frommolt@rz.hu-berlin.de

Abstract: In diesem Beitrag stellen wir eine Projektinitiative zum Aufbau eines verteilten Medienarchivs für bioakustische Datenbestände vor. Im Zuge dieser von der Deutschen Forschungsgemeinschaft (DFG) geförderten Initiative wurde zunächst aufbauend auf den momentan in Digitalisierung befindlichen Datenbeständen des Tierstimmenarchivs der Humboldt Universität (HU) zu Berlin ein internetbasiertes Informationssystem zur örtlich verteilten Forschungskooperation im Bereich Bioakustik konzipiert und realisiert. Ein wesentliches Ziel der aktuellen Projektphase stellt darauf aufbauend die konkrete Anbindung von deutschlandweit verfügbaren bioakustischen Medienbeständen dar.

1 Einleitung

Zur Zeit existieren weltweit vereinzelt größere Tierstimmenarchive aus analogen Tonaufnahmen. Hinzu kommt eine große Zahl kleinerer, auf bestimmte Arten eingeschränkter Spezialarchive. Trotz begonnener Digitalisierungsbestrebungen gibt es zur Zeit keine internetbasierten Informationssysteme, die die wissenschaftliche Kooperation im Bereich Bioakustik unter Einbeziehung multimedialer Aspekte ermöglichen. In einem von der DFG geförderten Kooperationsprojekt des Instituts für Informatik der Universität Bonn mit dem Tierstimmenarchiv der HU Berlin wurde zunächst aufbauend auf dem momentan in der Digitalisierungsphase befindlichen Datenbestand des Tierstimmenarchivs, ein entsprechendes internetbasiertes Informationssystem entwickelt. Zentrale Ziele des Projekts sind der dauerhafte Einsatz des Systems am Tierstimmenarchiv der HU Berlin, die Realisierung eines Zugriffskonzepts für kooperative Nutzung mit der Möglichkeit zur Integration der Daten externer Archive, die integrierte Darstellung multimedialer Information (akustische Daten, visuelle Daten wie Spektrogramme, textuelle Daten, bibliographische Daten) und insbesondere die Konzipierung des Einsatzes inhaltsbasierter Retrievalfunktionalität.

In diesem Beitrag geben wir einen praxisorientierten Überblick über die bereits im Zuge dieser Projektinitiative entstandene Informationsinfrastruktur, diskutieren die zugrundeliegenden Ziele und beschreiben die angestrebten Funktionalitäten. Im nächsten Abschnitt geben wir hierzu zunächst einen Überblick über die am Tierstimmenarchiv der HU Berlin verfügbaren Mediendaten und die aktuellen Digitalisierungsaktivitäten. Danach fassen wir

die zentralen Ziele der Projektinitiative zum Aufbau der Bioakustik-Infrastruktur zusammen und berichten über den aktuellen Entwicklungsstand. Zum Abschluss stellen wir erste Konzepte und aktuelle Aktivitäten zur Anbindung externer Datenarchive vor.

2 Das Tierstimmenarchiv in Berlin

Das Tierstimmenarchiv an der HU Berlin[1] 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 HU 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 dem Jahr 2002 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, Aufnahmedatum, Aufnahmeort, usw. erfasst werden. Momentan sind ca. 18.000 Datensätze in die Datenbank übernommen, und ca. 1.500 Aufnahmen in Studionorm digitalisiert. Mit dem vorläufigen Abschluss der Digitalisierung wird in ca. fünf Jahren gerechnet. Es wird dann Datenmaterial im Umfang von etwa 5.000 Stunden (ca. 10 Terabytes) digital vorliegen.

3 Ziele

Das wichtigste erste 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 existierende bioakustische Aufnahmen und Metadaten realisieren. Dieser Zugriff soll durch ein webbasiertes Informationssystem ermöglicht werden, das über ein einheitliches Interface Datenverwaltung, Daten- und Benutzersichten zur Verfügung stellt. Ein wichtiges Teilziel ist hierbei die Erweiterbarkeit des Bestands durch Daten externer Nutzer. Die Basis hierzu stellt eine flexibel konfigurierbare 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.

Zur Erzielung einer hohen Akzeptanz des Systems ist dessen Benutzerfreundlichkeit von hoher Bedeutung. Um diese zu erreichen, ist die Verwendbarkeit des Systems auch mit den im vorgesehenen Nutzerkreis im allgemeinen vorauszusetzenden geringen technischen

[1] `http://www.biologie.hu-berlin.de/~tsarchiv/index_ger.html`

Kenntnissen sicherzustellen.

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 der angebundenen Datenbanken an ihren eigenen Beständen von zentraler Bedeutung. Der Zugriff sowohl auf Metadaten als auch auf Audioaufnahmen muss hier über ein geeignetes Zugriffskonzept erfolgen.

Besonderer Wert wird auf die Modularität des Systems gelegt. Durch eine offene Schnittstelle zur Einbindung externer Such- und Analyse-Algorithmen wird dabei eine weitreichende Erweiterbarkeit angestrebt und eine Nutzung des Systems im Bereich der Bioakustikforschung und der bioakustischen Mustererkennung unterstützt. Neben den traditionellen im System zur Verfügung zu stellenden Funktionalitäten zur metadatenorientierten Datenbanksuche können hierdurch langfristig Methoden zur inhaltsbasierten akustischen Suche (wie etwa der Suche anhand von Gesangsbeispielen) in das System integriert werden.

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

4 Systemüberblick

Für die Umsetzung des in diesem Projekt zu erstellenden verteilten Informationssystems wurde das OpenSource Datenbankmanagement-System MySQL zugrundegelegt, wodurch eine Web-Anbindung des Systems problemlos zu gewährleisten ist. Für die vorhandene Access-Datenbank wurden geeignete Filter erstellt, so dass ein Export der bei der Digitalisierung erfassten (Meta-) Daten in das Informationssystem jederzeit möglich ist.

Problematisch stellt sich der große Datenumfang der anfallenden digitalen Audiodateien (Studionorm) dar. Kann die Langzeitarchivierung zentral auf Systemen des Rechnezentrums (CMS) der HU Berlin erfolgen, so sprengen Datenvolumen im Terabytebereich zur Zeit noch die am Informationssystem lokal verfügbaren Resourcen. Diesem wird in der Anfangsphase dadurch begegnet, dass die über das Informationssystem unmittelbar zugreifbaren Audiodaten verlustbehaftet komprimiert werden. Dies lässt sich rechtfertigen, da auch Versionen mit geringerer Datenrate aber ohne signifikante perzeptuelle Unterschiede zum Original für übliche Rechercheszenarien sowie viele wissenschaftliche Aufgaben durchaus hinreichend sind.

Wir wollen nun einen Überblick über die bereits realisierten Funktionalitäten des System geben. Dies sind vor allem Suchfunktionalitäten, eine Möglichkeit zum Einstellen von Daten und die Verwaltung von Benutzern und Benutzerrechten.

Für die Suche nach Metadaten wurde einerseits eine Standardsuche, die eine typische Anfrage an ein Tierstimmenarchiv widerspiegelt implementiert, andererseits eine erweiterte

Suche, die eine im Wesentlichen freie Suche in der Metadatenbank erlaubt[2]. Eine typische Suche in den Metadaten hat sich hierbei als eine Anfrage nach einer bestimmten Tierart herausgestellt, wobei möglicherweise eine Einschränkung der Suche nach Lauttyp sinnvoll ist. In der erweiterten Suche ist eine disjunktive Termsuche in allen Datenbankfeldern, die für den Nutzer relevant sind, möglich. Zudem ist eine Einschränkung der Suche auf einen bestimmten Aufnahmezeitraum möglich. Wie in vielen Recherchesystemen üblich, werden Suchergebnisse in Form einer Ergebnisliste dargestellt. Aus dieser lässt sich zu jedem Datensatz eine Detailansicht mit allen verfügbaren Metadaten-Einträgen aufrufen. Für diese Ansicht können zusätzlich Plugin-Module entwickelt werden, die spezielle Funktionalitäten zur Darstellung eines Datensatzes zur Verfügung stellen. Exemplarisch werden hier derzeit zwei Plugin-Module zur Wiedergabe von Audiodateien und für eine Google-Bildsuche nach dem Artnamen zur Verfügung gestellt.

Unter Verwendung eines jeweils lokal zu installierenden Programm-Moduls (Java-basiert) können berechtigte Nutzer eigene Daten zur Einstellung in das Informationssystem vorschlagen. Hierzu können über eine Eingabemaske zunächst lokal Metadatensätze eingegeben und bearbeitet werden. Dabei sollen eine Reihe von Eingabehilfen die Gefahr von Fehleingaben minimieren. Konkret ist hier zum Beispiel eine automatische Vervollständigung von Art- und Ortsnamen, sowie die automatische Anreicherung von Ortsnamen mit weiteren Metadaten, wie geographischen Koordinaten implementiert. Eingegebene Datensätze lassen sich lokal speichern und bei Bedarf zusammen mit den zugehörigen Audiodateien über das Netz an eine zentrale Datenbank übermitteln. Die so vorgeschlagenen Daten werden nicht ungeprüft in die Datenbank übernommen, sondern zunächst in eine Vorschlagsliste transferiert. Die Vorschlagsliste kann von autorisierten Nutzern eingesehen und bearbeitet werden, um so eine Qualitätssicherung aller eingestellten Daten zu gewährleisten.

Die Autorisierung von Nutzern des Informationssystems erfolgt über die Zuteilung eines Benutzernamens zusammen mit spezifischen Rechten für verschiedene Aufgaben wie Metadaten-Recherche, das Herunterladen von Audiodaten oder die Benutzerverwaltung. Zusätzlich werden für jede Datei und jeden Benutzer Dateirechte verwaltet, mittels derer die Zugriffsrechte auf Audiodaten eingeschränkt werden können.

Insgesamt stehen damit derzeit die wichtigsten Funktionen zur Verfügung, um mit dem bereits vorhandenen Datenbestand arbeiten zu können und im kleineren Umfang externe Daten zur Verfügung stellen zu können. Die Anbindung ganzer Archive stellt derzeit eines der wichtigsten offenen Arbeitspakete dar. Auf Konzepte zur Realisierung dieser Aufgabe wird im nächsten Abschnitt eingegangen.

5 Anbindung externer Tierstimmenbestände

Um größere externe Datenbestände über ein zentrales System recherchierbar zu machen sind eine Vielzahl von Lösungen denkbar. Für den vorliegenden Fall wurden drei Konzepte

[2] Unter `http://audentify.iai.uni-bonn.de/guest.php` ist ein Test des Systems über einen zugriffsbeschränkten Gastzugang möglich.

entwickelt, die unterschiedlich starke Voraussetzungen an die verfügbaren Systemresourcen auf Seiten des angebundenen Archivs stellen und eine unterschiedlich starke Kontrolle über die eigenen Daten erlauben.

Das technisch einfachste Konzept ist hierbei die vollständige Eingliederung von Daten in das bestehende System. Hierbei werden sowohl Audiodaten als auch zugehörige Metadaten auf einem zentralen Server gespeichert. Die Kontrolle über die Daten wird dadurch weitgehend dem System überlassen. Ein zweites Konzept sieht vor, zwar die Metadaten weiterhin in eine zentrale Datenbank zu überstellen, die Audiodateien aber auf einem eigenen Dateiserver zur Verfügung zu stellen. Hiermit bleibt die vollständige Kontrolle über die Audiodaten beim jeweiligen Besitzer. Schließlich lässt sich eine vollständig verteilte Lösung ins Auge fassen, bei der sowohl Metadaten als auch Audiodaten auf einem peripheren Server vorliegen und Anfragen von der zentralen Infrastruktur an diesen weitergeleitet werden. Die Ergebnisse solcher verteilten Anfragen müssen dann kombiniert und geeignet dargestellt werden. Diese Lösung ermöglicht die vollständige Kontrolle über alle zur Verfügung gestellten Daten.

Um eines der beiden ersten Konzepte zur Anwendung zu bringen ist es nötig, neu anfallende Daten an eine zentrale Datenbank weiterzuleiten bzw. geänderte Datensätze dort zu aktualisieren. Da in verschiedenen Institutionen unterschiedliche Metadaten-Formate vorliegen ist hierzu eine einheitlich Metadatenstruktur notwendig, auf die sich alle lokalen Lösungen abbilden lassen. In diesem Punkt müssen auch kritische fachspezifische Fragen, wie die Einigung auf eine gemeinsame zoologische Systematik, geklärt werden. Bei jüngsten Gesprächen mit potentiellen Bereitstellern von Daten hat sich hier herauskristalliersiert, dass das zweite vorgestellte Konzept allgemein favorisiert wird. Insbesondere besteht vielfach aufgrund zwingender externer Randbedingungen der Wunsch, den eigenen Datenbestand in einem lokal bereits vorhandenen System (z.B. Inventarisierungssysteme in Museen) weiter zu pflegen und in regelmäßigen Abständen mit einer zentralen Datenbank abzugleichen.

6 Ausblick

Die wichtigsten Komponenten einer Informations-Infrastruktur zur verteilten Kooperation in der Bioakustik stehen durch das vorgestellte Projekt derzeit zur Verfügung. Über die Arbeit mit bereits bestehenden Daten und die Einstellung aktuell anfallender Tierstimmen-Aufnahmen hinaus ist das primäre Ziel für den weiteren Ausbau des Systems nun, erste externe Bestände (Staatliches Museum für Naturkunde Stuttgart, Institut für Zoologie Mainz, Tierärztliche Hochschule Hannover) an das System anzubinden. Die hierzu notwendige Abstimmung eines gemeinsamen Metadatenformats ist weit fortgeschritten und es herrschen i.w. übereinstimmende Vorstellungen über ein konkretes Konzept. Die Anbindung externer Dateiserver ist bereits möglich, so dass nun die Entwicklung konkreter Datenkonverter den wichtigsten Schritt auf diesem Weg darstellt. Insbesondere kann hiermit eine intensive Testphase des Systems gestartet werden.