ÜBERSEE
MUSEUM
BREMEN

Aufbereitung der Datenquellen der naturkundlichen Sammlung des Übersee-Museums Bremen

Inhalt

I. Ziele des Projekts

Im Rahmen des drittmittelgeförderten Projekts „Verfügbarmachung der naturkundlichen regionalen und in kolonialen Kontexten erworbenen Sammlungen des Übersee-Museums Bremen für Forschung, Bildung und kulturelle Teilhabe“ (siehe dazu auch unter https://www.uebersee-museum.de/ueber-uns/projekte-positionen/digitalisieren-und-teilen-von-naturkundlichen-sammlungsinformationen/) bereitet das Übersee-Museum Bremen seit einiger Zeit die Optimierung der digitalen Erfassung seiner Sammlungen vor. Die naturkundlichen Sammlungen werden derzeit in verschiedenen Dateien (Access, Excel, dBase) organisiert. In Kooperation mit dem Museum für Naturkunde Berlin plant das Übersee-Museum derzeit eine Überführung der Daten in das europäische Sammlungsmanagementsystem DINA ("DIgital Information System For NAtural History Data"), um u.a. eine gemeinsame Online-Stellung aller Sammlungen (ethnographische, handelsgeschichtliche, historische Fotografie, naturkundliche) sowie einen Datenaustausch zwischen verschiedenen Datenbanksystemen über Schnittstellen zu ermöglichen.

Die naturkundlichen Sammlungen des Übersee-Museums umfassen über 1 Million zoologische und botanische Objekte, darunter mehr als 2.000 besonders wertvolle Sammlungsstücke wie Typenmaterial und Belege ausgestorbener Organismen. Es existieren bereits ca. 271.000 digitale Datensätze zu den o.g. Objekten, die in verschiedenen Datenformaten z.T. sehr kleinteilig und nicht einheitlich erfasst sind. Datenfelder und Erfassungsrichtlinien in diesen Datensätzen variieren stark. Für die geplante Datenmigration müssen diese historisch-bedingt äußerst heterogen erfassten digitalen Daten der naturkundlichen Sammlungsobjekte in einer einheitlichen Datenstruktur zusammengeführt werden.

Das Übersee-Museum sucht dafür eine*n Dienstleister*in für die Planung und Umsetzung der Arbeitsschritte zur Vorbereitung der naturkundlichen Sammlungsdaten für die Migration in ein gemeinsames Datenbanksystem aus diversen Datenquellen. Planung und Umsetzung der Überarbeitung der Datensätze erfolgt in enger Absprache mit den Mitarbeiter*innen des Übersee-Museums.

II. Die Vorarbeiten

Die existierenden Metadaten sowie die geplanten Datenfelder wurden für die Naturkunde-Datenquellen (NK-Datenquellen) zusammengestellt und bewertet. Für die insgesamt ca. 160 geplanten Datenfelder wurden erste Schreibanweisungen formuliert (datenfelder_uebersicht.pdf). Es existieren zudem erste Überlegungen zum Mapping der NK-Datenquellen auf die geplanten Datenbankfelder. Die Informationen der Datenfelder der NK-Datenquellen wurden den Ziel-Datenfeldern aus der Datenfeldtabelle zugeordnet. Daraus resultieren erste Umpflegeanweisungen, wie die Daten neu verteilt oder zusammengeführt werden sollen (Datenfelder-fuer-Mapping,v5.pdf).

Die o.g. Dokumente wurden und werden im Laufe des Projekts stetig weiterbearbeitet und angepasst. Auch die NK-Datenquellen werden derzeit weiterhin produktiv von den Mitarbeiter*innen des Übersee-Museums genutzt und weiterentwickelt. Die dieser Ausschreibung beigefügten Anlagen entsprechen zwar nicht mehr dem aktuellen Stand, sie vermitteln aber einen Eindruck von der Heterogenität und dem Umfang der zu überarbeitenden digitalen Daten und können darum als gute Grundlage für die Erstellung des gewünschten Angebots dienen.

III. Der Umfang im Einzelnen

Um die Daten der NK-Datenquellen in ein gemeinsames Datenbanksystem überführen zu können benötigt das Übersee-Museum einheitlich aufbereitete und an die geplanten Datenbankfelder angepasste Datentabellen. Die NK-Datenquellen umfassen derzeit rund 50 einzelne Dateien, bestehend aus ca. 80 Tabellenblättern mit 9 bis zu 95 Spalten und ca. 271.000 Datensätzen. Anzahl, Inhalt sowie die Benennung der einzelnen Datenfelder („Spalten") ist von Tabelle zu Tabelle sehr unterschiedlich. Es handelt es sich meist um Excel-Tabellen, sowie um Access-Dateien und einen dBase-Export, der voraussichtlich keine relationalen Daten enthält. Die Tabellen der NK-Datenquellen werden in der täglichen Arbeit produktiv genutzt (d.h. fortlaufende Eingabe neuer Datensätze, Überarbeitung bestehender Datensätze) und durch weitere Spalten ergänzt. Dies gilt es bei der Konzeption zu beachten. Die folgende Aufstellung gibt eine Übersicht über die NK-Datenquellen (NK-Datenquellen_Uebersicht2022-04.pdf) (vorbehaltlich Änderungen):

Botanik:

1. Pflanzen – BREM-Datenbank: 1 Excel-Tabelle, 5.057 Datensätze, 57 Spalten
2. Pflanzen – Schulze: 1 Excel-Tabelle, 26 Tabellenblätter, 10.321 Datensätze, 37 bis 39 Spalten
3. Samen & Früchte: 1 Excel-Tabelle: 1.722 Datensätze, 31 Spalten
4. Schatteburg-Kataloge: 5 Excel-Tabellen, 12.530 Datensätze, 9 bis 10 Spalten
5. Herbarbestand: 1 Access-Datei, 41.621 Datensätze, 27 Spalten

Zoologie:

6. Krebstiere: 1 Excel-Tabelle, 839 Datensätze, 14 Spalten
7. Wirbellose: 1 Excel-Tabelle, 14 Tabellenblätter, 1.096 Datensätze, 16 bis 19 Spalten
8. Insekten: 1 Access-Datei (dBase-Export), 139.811 Datensätze (Sammeleinträge), 77 bis 95 Spalten
9. Weichtiere – Mersch: 1 Excel-Tabelle, 838 Datensätze, 24 Spalten
10. Fische: 4 Excel-Tabellen, 4.597 Datensätze, 21 Spalten
11. Vögel – Realkataloge: 10 Excel-Tabellen, 23.683 Datensätze, 18 bis 26 Spalten
12. Vögel – Aves: 1 Excel-Tabelle, 4.493 Datensätze, 26 Spalten
13. Vögel – Spezialkataloge: 13 Excel-Tabellen, 15.967 Datensätze, 9 bis 14 Spalten
14. Vögel – Eier: 1 Excel-Tabelle, 525 Datensätze, 27 Spalten
15. Vögel – Knochen: 1 Excel-Tabelle, 519 Datensätze, 28 Spalten
16. Säugetiere: 1 Excel-Tabelle, 7.556 Datensätze, 24 Spalten

Die Felder (Spalten) der o.g. Tabellen müssen entsprechend der geplanten Datenbankfelder einheitlich formatiert und ggfs. in mehrere Spalten aufgeteilt oder zusammengeführt werden. Erste Überlegungen zum Datenmapping und zur geplanten Metadatenstruktur liegen vor (s. II. Vorarbeiten). Die Inhalte müssen zudem an auf den vorhandenen Metadaten basierende, kontrollierte Vokabulare sowie an festgelegte Schreibweisen (Patterns), z.B. für Datumsangaben und geografische Koordinatensysteme, angepasst werden. Die taxonomische, geografische und Personen-Daten müssen mit Standardvokabularen/Thesauri (z.B. Taxonomie: Catalogue of Life, Geografie: GeoNames oder Getty TGN, Geokoordinaten; Personen/Körperschaften: Gemeinsame Normdatei, GND, ggfs. Art & Architecture Thesaurus, AAT) abgeglichen werden. Bestimmte Metdatenfelder müssen mit Datenfeldern aus anderen Systemen des Übersee-Museums über Listen abgeglichen werden (Personennamen und Eingangsnummern), sofern diese entsprechend standardisiert werden können. Darüber hinaus sollen Fehler (z.B. Tippfehler) sowie Unregelmäßigkeiten aufgefunden und bereinigt werden. Um den Verlust von Informationen zu reduzieren, soll auch ein Großteil der Altdaten – so wie

sie vor der Formatierung/Überarbeitung in den Tabellen enthalten waren - in Sammelfeldern im Datenbanksystem hinterlegt sein. Darüber hinaus sollen die Sammlungsinformationen zu den Vögeln aus verschiedenen Tabellen, nämlich den Realkatalogen, der Aves-Tabelle und den Spezialkatalogen, in 1 Datensatz pro Objekt zusammengeführt werden. Die Datenstruktur der Spezialkataloge muss zuvor in eine einheitliche, zeilenbasierte Datenstruktur überführt werden. Die Sammeleinträge in der Insektentabelle sollen so gesplittet werden, dass jedes Insekt einen eigenen Datensatz bekommt.

Die Aufgaben im Einzelnen:

III.1 Konzept für die Vorgehensweise und Durchführung der Datenaufbereitung

Zunächst soll ein detailliertes Konzept für die Vorgehensweise und Durchführung der Datenaufbereitung und -transformation der tabellarischen Daten erstellt werden. Die einzelnen, dafür notwendigen Arbeitsschritte sollen benannt und definiert sowie in eine zeitliche Reihenfolge gebracht werden. Die Zuständigkeit für die einzelnen Arbeitsschritte soll ebenfalls benannt werden, d.h. welche Arbeitsschritte können oder müssen von den Mitarbeiter*innen des Übersee-Museums durchgeführt werden, welche Arbeitsschritte übernimmt der/die Dienstleister*in. Das Konzept beinhaltet auch den Abgleich der taxonomischen, geografischen und Personen-Daten mit Standardvokabularen/Thesauri (z.B. Taxonomie: Catalogue of Life, Geografie: GeoNames oder Getty TGN, Geokoordinaten; Personen/Körperschaften: Gemeinsame Normdatei, GND, ggfs. Art & Architecture Thesaurus, AAT), die Zusammenführung der 3 Datenquell-Typen der Vogelsammlung, das Splitten der Sammeleinträge in der Insektentabelle, sowie den Abgleich bestimmter Metdatenfelder mit Datenfeldern aus anderen Systemen des Übersee-Museums (Personennamen und Eingangsnummern, sofern diese entsprechend standardisiert werden können).

III.2 Durchführung der Datenaufbereitung und -transformation

Ausgehend von III.1 soll die eigentliche Aufbereitung und Transformation der oben in diesem Abschnitt aufgelisteten Datenquellen durchgeführt werden. Die Durchführung der Überarbeitung erfolgt in Abstimmung und Zusammenarbeit mit den Mitarbeiter*innen des Übersee-Museums.

III.3 Zusammenführen der Vogeltabellen

Die Sammlungsdaten zu den Objekten der Vogelsammlung sind auf 3 Datenquell-Typen verteilt: Die Realkataloge, die Spezialkataloge und die Aves-Tabelle. Jedes Sammlungsobjekt kann in allen 3 Datenquell-Typen enthalten sein und ist dort über die Quellen hinweg eindeutig identifizierbar durch die Realkatalog-Nummer. Die Informationen zu einem Objekt können von Quelle zu Quelle unvollständig und widersprüchlich sein. Alle Informationen sollen in 1 Datensatz pro Objekt zusammengeführt werden. Die 5 Realkataloge sind am umfangreichsten und vollständigsten und sind darum die Ausgangsbasis. Die Aves-Tabelle enthält Informationen auf Etikettenbasis, inkl. Nachbestimmungen und soll darum in die Realkataloge überführt und mit diesen abgeglichen werden. Im Gegensatz zu den Realkatalogen und der Aves-Tabelle liegen einige Spezialkataloge nicht zeilenbasiert vor, stattdessen erstreckt sich der Datensatz zu einem Objekt über mehrere Zeilen entsprechend der Struktur in den analogen Katalogen. Die Tabellen müssen dahingehend bereinigt werden, dass eine maschinell gut weiterbearbeitbare Tabelle entsteht, die eine zeilenbasierte Datenstruktur hat, d.h. 1 Zeile = 1 Datensatz und 1 Datensatz = 1 Objekt.

III.4 Splitten der Sammeleinträge in der Insektentabelle

Bei der Insektentabelle handelt es sich um einen Export aus dem Datenbankmanagementsystem dBase. Die meisten Datensätze (=Zeilen) in dieser Tabelle stellen einen Sammeleintrag über mehrere Insekten-Objekte dar, die jeweils identische Sammel-Metadaten („Sammelereignis“) aufweisen. Ein solcher Datensatz enthält beispielsweise identische Metadaten zu 24 Insekten, von den 2 Weibchen, 9 Männchen und 13 Arbeiterinnen sind. Das Ziel ist eine Individuen-basierte Erfassung der einzelnen Insekten. Die Sammeleinträge (1 Zeile = n Insekten) sollen so gesplittet werden, dass jedes Insekt einen eigenen Datensatz (1 Zeile = 1 Insekt) hat, im o.g. Beispiel entsprechend 24 Datensätze (=Zeilen) für die 24 Insekten.

III.5 Entwicklung von kontrollierten Vokabularen

Basierend auf den vorhandenen Metadaten sollen für bestimmte Datenfelder Vorschläge für kontrollierte Vokabularen entwickelt werden, die anschließend in der Zieldatenbank (DINA) hinterlegt werden können. Dafür müssen die entsprechenden Datenfelder über die NK-Datenquellen hinweg analysiert und ausgewertet werden. Die Vorschläge werden anschließend von den Mitarbeiter*innen des Übersee-Museums bewertet und mit weiteren Begriffen angereichert.

III.6 Dokumentation

Die zugrunde gelegten Aufbereitungs- und Überarbeitungsroutinen für die (Nach-) Nutzung, z.B. in OpenRefine, sollen so dokumentiert werden, dass die Mitarbeiter*innen des Übersee-Museums die Datenüberarbeitung an weiteren NK-Datenquellen nach Ende des Projekts fortsetzen und eigenverantwortlich durchführen können.

III.7 optionale Leistung: dBase-Export der Insektendaten

Die aktuellste Version der Insektendaten liegt derzeit als dBase-Datenbank vor. Diese soll zur Weiterverarbeitung exportiert und analysiert werden. Diese Leistung ist optional. Wenn sie nicht angeboten werden kann, bedeutet dies kein Ausschluss vom Verfahren.

IV. Rahmenbedingungen

IV. 1 Zeitraum

Die Laufzeit des Auftrags ist Mai 2022 bis Oktober 2022 (ggf. Verlängerung bis Dezember 2022 nach Absprache).

IV. 2 Vorgehensweise

Die beteiligten Mitarbeiter*innen des Übersee-Museums verfügen über Grundkenntnisse in der Arbeit mit MS Office Excel und konnten zudem im November 2021 an einer ersten Einführung in die Open-Source-Software OpenRefine (https://openrefine.org/) in Form eines 1,5 tägigen Workshops teilnehmen. Daher soll bei der Durchführung der Datenaufbereitung und -transformation bevorzugt Excel und OpenRefine eingesetzt werden, sofern es für die Bearbeitungsschritte angemessen ist, da so

die Zuarbeit und Nachnutzung von Prozessdokumentation durch das Team erleichtert wird. Wir wünschen uns ein Angebot über folgende Leistungen:

1. Erstellung eines detaillierten Konzepts für die Vorgehensweise und Durchführung der Datenaufbereitung und -transformation (s. III.1).
2. Durchführung der Datenaufbereitung und -transformation (s. III.2)
3. Zusammenführen der Vogeltabellen (s. III.3)
4. Splitten der Sammeleinträge in der Insektentabelle (s. III.4)
5. Entwicklung von kontrollierten Vokabularen (s. III.5)
6. Dokumentation der zugrunde gelegten Aufbereitungs- und Überarbeitungsroutinen (s.III.6)
7. dBase-Export der Insektendaten (optionale Leistung) (s. III.7)

Die Unterstützung und Anleitung der beteiligten Museumsmitarbeiter*innen während der Datenaufbereitung ist ebenfalls Bestandteil des Auftrags. Für die Bearbeitung des Auftrags ist eine Anwesenheit in Bremen in den Räumen des Übersee-Museums nicht erforderlich. Alle Absprachen können telefonisch und per Zugriff von extern erfolgen.

V. Zum Verfahren

Das Übersee-Museum Bremen sucht eine*n Dienstleister*in mit nachgewiesener Expertise in der Aufbereitung und Transformation großer Datentabellen. Vorerfahrungen mit der Open-Source-Software OpenRefine (https://openrefine.org/) sind hierbei von Vorteil. Die Bereitschaft, OpenRefine einzusetzen, auch ohne Vorerfahrung, ist eine Bedingung.

Bitte senden Sie uns Ihr Angebot bis zum 16. Mai 2022. Der Zuschlag wird erteilt zum 20. Mai 2022. Die Laufzeit des Auftrags hängt vom angebotenen Aufwand ab. Der Auftrag soll spätestens bis Oktober 2022 (längstens jedoch, und erst nach Absprache, bis Dezember 2022) abgeschlossen sein.

Die Kriterien für die Erteilung des Zuschlags sind:

- Preis 40 Prozent (Preisgestaltung, Gesamtpreis)
- Konzept und Vorgehensweise: 30 Prozent (Transparenz der Arbeitsschritte, Angemessenheit der vorgeschlagenen Maßnahmen)
- Vorerfahrung: 30 Prozent (Portfolio, Projekte, OpenRefine)

Bitte erstellen Sie ein Angebot, in dem die einzelnen unter IV. 2 genannten Leistungen einzeln aufgeführt und kalkuliert sind (sowohl ihr kostenmäßiger als auch ihr zeitlicher Umfang). Wir behalten uns vor, einzelne Positionen nicht zu beauftragen – der/die Bieter*in erhält aber die Gelegenheit, das Angebot zu korrigieren, wenn einzelne Positionen nicht beauftragt werden.

Folgende Anlagen stehen für die Ausarbeitung des Angebots zur Verfügung:

1. Übersicht über die geplanten Datenfelder (Anlage1_datenfelder_uebersicht.pdf)
2. Mapping der NK-Datenquellen auf die geplanten Datenbankfelder (Anlage2_Datenfelder-fuer-Mapping,v5.pdf)
3. Übersicht über die NK-Datenquellen (Anlage3_NK-Datenquellen_Uebersicht2022-04.pdf)

Der Zugriff auf die 3 Anlagen erfolgt über https://my.hidrive.com/share/ii.o9krhl6. Interessierte Bieter*innen erhalten das Passwort auf Anfrage. Die NK-Datenquellen können auf Anfrage ebenfalls eingesehen werden. Eine Veröffentlichung sowie die Weitergabe der Tabellen an Dritte ist nicht gestattet!

VI. Ansprechpartnerin

Dr. Michaela Grein
Kuratorin Botanik
Abteilung Naturkunde
Übersee-Museum Bremen
Bahnhofsplatz 13
28195 Bremen
m.grein@uebersee-museum.de

Rückfragen bitte schriftlich formulieren. Bitte beachten Sie, dass Ihre Anfrage (anonym) und unsere Antwort darauf auch anderen Bieter*innen auf https://www.uebersee-museum.de/bieterfragen-und-antworten/ zur Verfügung gestellt werden!