DEUTSCHER MUSEUMS BUND

MUSEUMSKUNDE

FACHZEITSCHRIFT FÜR DIE MUSEUMSWELT

Die Fachzeitschrift *Museumskunde* bietet vertiefende, vielseitige Positionen zu aktuellen museumsspezifischen Themen. Die Zeitschrift wurde 1905 als Ausdruck der Zusammengehörigkeit von Museumsfachleuten gegründet und setzt sich seitdem mit relevanten Themen für das Museumswesen auseinander. Die *Museumskunde* wird seit 1917 vom Deutschen Museumsbund herausgegeben.

www.museumsbund.de

ISSN 0027–4178

MUSEUMSKUNDE **Band 89 Doppelausgabe 2024** MUSEEN UND KI / MUSEEN DURCH KRISEN NAVIGIEREN

MUSEUMSKUNDE
2024
FACHZEITSCHRIFT FÜR DIE MUSEUMSWELT

Museen und KI

Museen durch Krisen navigieren

DEUTSCHER MUSEUMS BUND

Vom Sammlungsobjekt zum Datenobjekt für den Umweltschutz

KI-GESTÜTZTE EXTRAKTION VON BIODIVERSITÄTSDATEN AUS NATURHISTORISCHEN SAMMLUNGSETIKETTEN

Von CHRISTIAN BÖLLING, MARGOT BELOT, FRANZISKA SCHUSTER, ANIKA GEBAUER, UTE KISSLING-BRENNER und THERESE REINKE

ABB. 1 — **Objektetiketten aus der Sammlung des Museums für Naturkunde Berlin.** Foto: Margot Belot, © Museum für Naturkunde Berlin.

Methoden der Künstlichen Intelligenz haben immenses Potenzial für die Erschließung naturhistorischer Sammlungen und die Generierung sammlungsbasierter, FAIRer[1] Forschungsdaten. Die Entwicklung entsprechender Workflows und ihre Integration in den Prozess der Sammlungserschließung erfordern geeignete Förderinstrumente und die interdisziplinäre Bündelung von Know-how. In einem KI-Pilotprojekt entwickeln die KI-Ideenwerkstatt für Umweltschutz, eine Initiative des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV) und das Museum für Naturkunde Berlin (MfN) eine Anwendung zur Extraktion biodiversitätsrelevanter Informationen aus Sammlungsetiketten.

Schriftliche Überlieferungen im Kontext naturhistorischer Sammlungen, wie Objektetiketten, Eingangskataloge oder Expeditions- und Reiseberichte, stellen eine aufgrund ihrer teils bis ins 18. Jahrhundert zurückreichenden Historizität eine herausragende Datenquelle dar. So sind die aus dem Sammlungsmaterial für historische Zeiträume ableitbaren Daten zu Vorkommen, Abundanz und Habitaten biologischer Arten von wesentlicher Bedeutung für die Biodiversitätsforschung bei der Analyse langfristiger Trends und als Indikatoren für den Zustand von Landschaftsräumen und Ökosystemen in der Vergangenheit. Bislang sind diese Daten weitestgehend nicht verfügbar, da die manuelle Erschließung der Quellen und die Überführung der Daten in gemäß der FAIR-Prinzipien strukturierte Datenformate aufgrund der Menge, der inhaltlichen und strukturellen Heterogenität sowie Komplexität der Dokumente ein Prozess ist, für den in der Regel keine ausreichenden kuratorischen Kapazitäten vorhanden sind. Verfahren der Künstlichen Intelligenz mit der Fähigkeit zur Erkennung und Kategorisierung komplexer Muster und der Prozessierung großer Datenmengen können diesen Prozess erheblich beschleunigen und im Hinblick auf die Skalierbarkeit wirtschaftlich gestalten. Als initialer Schritt zur Etablierung KI-gestützter Prozesse im Kontext der digitalen Sammlungserschließung wurde am MfN ein teilautomatisierter Workflow zur Informationsextraktion aus gedruckten Sammlungsetiketten entwickelt.[2] In einem in Zusammenarbeit mit der KI-Ideenwerkstatt durchgeführten Projekt soll die KI-basierte Informationsextraktion nun robuster und funktionell erweitert werden.

ERSCHLIESSUNG UND ENTWICKLUNG DER SAMMLUNGEN AM MUSEUM FÜR NATURKUNDE

Die mehr als 30 Millionen Objekte umfassende und in ihren Ursprüngen bis ins frühe 18. Jahrhundert zurückreichende wissenschaftliche Sammlung des MfN ist Gegenstand einer Transformation, die die Realisierung von Zugängen und Services für die Nutzung der Sammlung im digitalen, physischen und hybriden Raum durch technologische Innovation zum Ziel hat. Das Projekt *Sammlungserschließung und -entwicklung*[3] steht dabei als Teil des seit 2020 am MfN umgesetzten Zukunftsplans für eine Weiterentwicklung der Sammlung, bei der durch die digitale Erfassung der Objekte das auf ihrer Basis generierte Wissen in einem digital referenzierten Katalog bereitgestellt wird, der internationalen Datenstandards Rechnung trägt, eine Vernetzung mit anderen Datenquellen erlaubt und barrierearme Zugänge für möglichst viele Nutzungsszenarien bereitstellt (**ABB. 2**). Das Leitbild dieser integrierten Wissensinfrastruktur beinhaltet, dass die Repräsentationen der physischen Objekte die Möglichkeit zur multiperspektivischen Exploration von

ABB. 2 — **Leitbild des Projekts *Sammlungserschließung und -entwicklung* ist die Transformation der Sammlungen in eine integrierte Wissensinfrastruktur.**[4] Grafik: Christine Oymann, © Museum für Naturkunde Berlin.

Objekten, ihren Eigenschaften und Bezügen eröffnen, insbesondere auch im Hinblick auf aktuelle gesellschaftliche Herausforderungen wie zum Beispiel den Biodiversitätsverlust und die Analyse seiner anthropogenen Treiber. Die Sammlungserschließung vollzieht sich am MfN in Projekten mit sammlungs- und objektspezifischen Workflows und Erschließungszielen (**ABB. 3**). Der Begriff *Sammlungsdigitalisierung* steht dabei sowohl für die digitale Erfassung von Objekteigenschaften in Sammlungsdatenbanken (etwa zur Objektprovenienz, taxonomischen Einordnung, wissenschaftlichen Nutzung) als auch für bildgebende Verfahren, die digital kodierte, bildliche 2D- oder 3D-Repräsentationen von Objekten erzeugen. Bildgebende Verfahren werden, auch aufgrund ökonomischer und Nachhaltigkeitsüberlegungen, bei besonders frequentierten Beständen sowie *on demand* für spezifische Nutzungsanfragen eingesetzt und beinhalten auch Spezialverfahren wie 3D-Modellierung, Computertomografie (CT) oder Whole-slide-Scanning. Im Projekt *Digitize!*[5] wurde in Kooperation mit externen Technologieanbietern ein fließbandgestützter Workflow zur High-throughput-Digitalisierung von Insektenpräparaten entwickelt, der auch in die Ausstellung des Museums integriert war (**ABB. 4**). Innerhalb des Projekts wurden von mehr als 500.000 kuratorischen Einheiten aus der Hautflügler-Sammlung Bilddigitalisate der Präparate und der mit ihnen assoziierten Etiketten erzeugt und über persistente, global eindeutige Identifier verknüpft (**ABB. 5**). Um eine hohe Effizienz beim manuellen Handling der physischen Sammlungsobjekte zu erreichen, ist die digitale Erfassung der Objektdaten in diesem Ansatz ein zeitlich und organisatorisch nachgelagerter Prozess,[6] der auf der Auswertung der zuvor erzeugten Bilder fußt.

UNTERSTÜTZUNG DURCH DIE KI-IDEENWERKSTATT FÜR UMWELTSCHUTZ

Das Projekt *Collection Mining* am MfN bündelt Aktivitäten zur KI-gestützten Informationsextraktion aus Digitalisaten der schriftlichen Überlieferungen im Kontext der Sammlungen. Ein Schwerpunkt liegt dabei auf der Auswertung der digitalisierten Sammlungsetiketten. Die KI-gestützte Informationsextraktion ist als sequenzieller

ABB. 3 — **Die Objekt- und nutzungsspezifische Erschließung der Sammlung am MfN wird flankiert durch ganzheitliche Technologieentwicklung, Kompetenzaufbau und Infrastrukturentwicklung.**[7] Grafik: Christine Oymann, © Museum für Naturkunde Berlin.

Anwendungsworkflow von fünf aufeinander aufbauenden, funktionalen Schritten konzipiert:

(1) Vorprozessierung der Bilder zur Qualitätsverbesserung im Hinblick auf die folgenden Schritte des Workflows,
(2) Layoutanalyse zur Identifikation und Kategorisierung relevanter Bildbereiche wie Textblöcke und Zeilen,
(3) Schriftanalyse zur Texterkennung auf syntaktischer Ebene (Zeichen und Zeichenketten),
(4) Semantisches Parsing detektiert Kategorien semantischer Entitäten innerhalb des erkannten Texts (zum Beispiel taxonomische Namen, Personen, Orte) und
(5) Disambiguierung und Verlinkung, bei der erkannte Entitäten durch Verlinkung mit Normdateneinträgen identifiziert werden.

Die modulare Struktur mit standardisierten Schnittstellen zwischen den Modulen ermöglicht es, die einzelnen Schritte der Informationsextraktion im Hinblick auf variable Eingangsdaten durch verschiedene Implementationen anzupassen (zum Beispiel durch ein neu trainiertes Modell oder Nutzung einer über eine API eingebundene Anwendung Dritter). Die im Prozess extrahierten Daten werden im PAGE-XML-Format[8] für Folgeanwendungen, wie der Generierung von Linked Open Data,[9] zur Verarbeitung und Aggregation der Daten bereitgestellt. Die Zusammenarbeit mit der KI-Ideenwerkstatt[10] seit Mai 2024 ermöglicht eine Weiterentwicklung der Anwendungsarchitektur und des Funktionsumfangs für die Informationsextraktion auf der Basis der zuvor am MfN entwickelten funktionellen Module und der im Projekt *Sammlungserschließung und -entwicklung* erzeugten Digitalisate. Um künftig den Einsatz der KI-gestützten Informationsextraktionspipelines in verschiedenen Erschließungsprojekten mit unterschiedlichen Anforderungen zu ermöglichen, sollen ein Pipeline-Framework spezifiziert und ein frameworkkompatibler Prototyp implementiert werden. Daneben ist die Erstellung der für eine Veröffentlichung der Anwendung notwendigen Ressourcen, die Annotation weiterer Referenzdatensätze und die Anbindung einer Anwendung zur Prüfung und Korrektur der extrahierten Daten als Teil eines *Human-in-the-loop*-Workflows beabsichtigt. Dafür

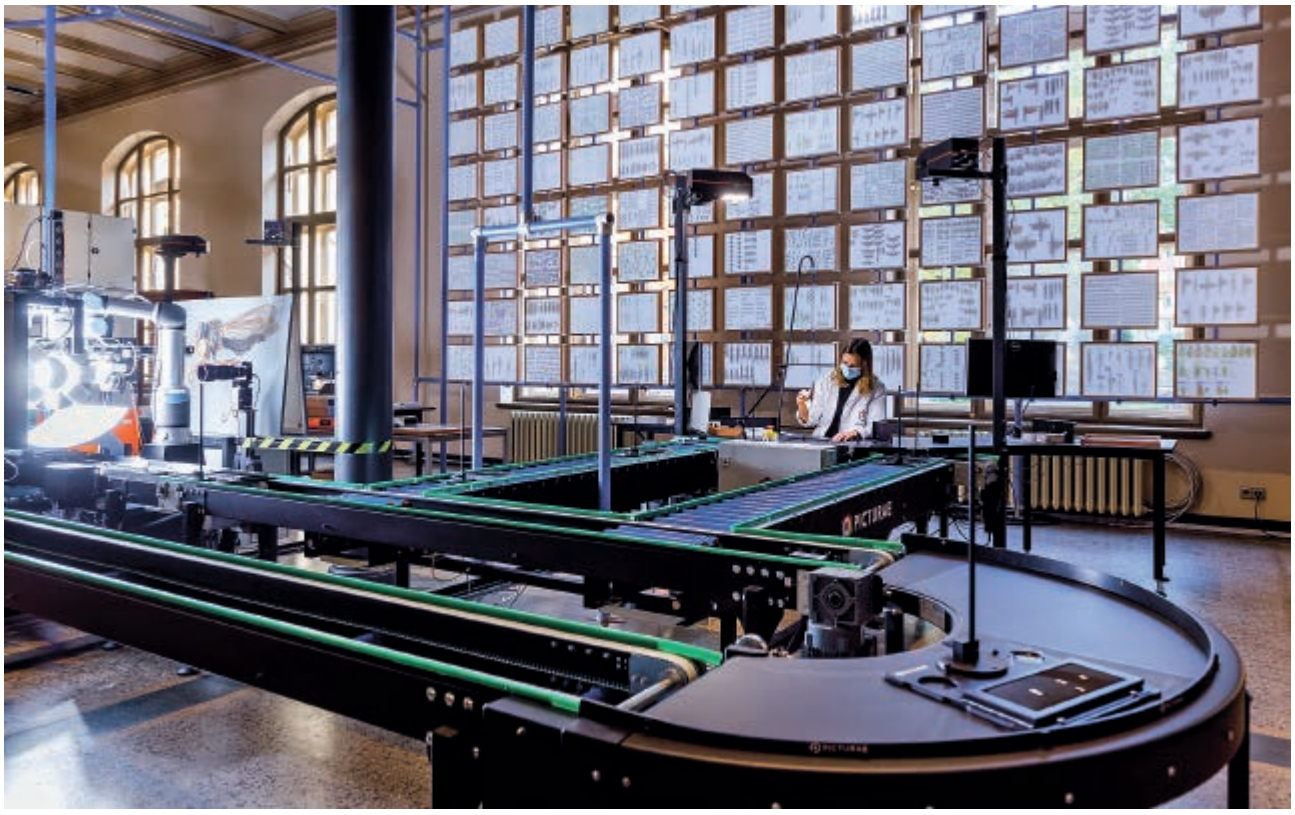

ABB. 4 — **Teilautomatisierte Objektdigitalisierung im Projekt *Digitize!*** Foto und ©: Thomas Rosenthal.

stellt die KI-Ideenwerkstatt entsprechendes Personal und ausreichend Rechenkapazität zur Verfügung. Besonderen Wert legt die KI-Ideenwerkstatt hier auf einen nachhaltigen Einsatz digitaler Technologien. Die Ergebnisse des KI-Pilotprojekts sollen allen Interessierten kostenlos zur Verfügung gestellt werden.

EIN OFFENER ORT FÜR INTERESSIERTE: DIE KI-IDEENWERKSTATT FÜR UMWELTSCHUTZ

Die KI-Ideenwerkstatt für Umweltschutz[11] des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz arbeitet im Impact Hub in Berlin-Neukölln. Sie dient vor Ort und virtuell als Anlaufstelle für alle, die Künstliche Intelligenz gemeinwohlorientiert für den Umweltschutz einsetzen möchten. Die KI-Ideenwerkstatt ist ein offener Ort für NGOs, Initiativen, Wissenschaftler*innen, Start-ups, Privatpersonen und weitere Akteur*innen und deren Austausch untereinander — jede*r kann und soll mitwirken. Gemeinsam und offen für alle entstehen in der KI-Ideenwerkstatt und durch weitere Formate an anderen Orten Ideen für den Natur-, Umwelt- und Klimaschutz. Jedes Jahr unterstützt die KI-Ideenwerkstatt mehrere KI-Pilotprojekte. 2024 spannen diese thematisch einen großen Bogen von der Erfassung historischer Biodiversitätsdaten, dem Unterwasser-Monitoring von Seegraswiesen bis hin zum smarten Schadstoff-Check in Alltagsprodukten und Kosmetika. Bewerben können sich Projekte und Ideen von

ABB. 5 — **Holzbienen der Hautflügler-Sammlung nach Durchlaufen der Digitalisierungsstraße.** Foto und ©: Eran Wolff, Museum für Naturkunde Berlin.

gemeinnützigen, nicht wirtschaftlich tätigen Zusammenschlüssen mehrerer Menschen aus Deutschland.

DIE KI-INITIATIVEN DES BMUV UND CIVIC CODING

Die KI-Ideenwerkstatt für Umweltschutz ist eine von mehreren KI-Initiativen des Programms *Künstliche Intelligenz für Umwelt und Klima*[12] des BMUV, das zum Ziel hat, einer nachhaltigen KI-Gestaltung und der Nutzung ihrer Chancen für Klima und Umwelt näherzukommen. Die *Zukunft—Umwelt—Gesellschaft (ZUG) gGmbH* setzt die KI-Ideenwerkstatt im Auftrag des Bundesumweltministeriums um. Die KI-Ideenwerkstatt ist zudem Teil von *Civic Coding — Innovationsnetz KI für das Gemeinwohl*,[13] einer gemeinsamen Initiative des BMUV, des Bundesministeriums für Arbeit und Soziales (BMAS) sowie des Bundesministeriums für Familie, Senioren, Frauen und Jugend (BMFSFJ).

Dr. Christian Bölling
Wissenschaftlicher Datenkurator
christian.boelling@mfn.berlin

Margot Belot
Scan-Operatorin
margot.belot@mfn.berlin

Franziska Schuster
Erschließungsmanagerin (Text, Bild und audiovisuelle Medien)
franziska.schuster@mfn.berlin

Forschungsbereich Zukunft der Sammlung
Museum für Naturkunde Berlin — Leibniz-Institut für Evolutions- und Biodiversitätsforschung
Invalidenstraße 43, 10115 Berlin

Anika Gebauer
Referentin für Data Science, Machine Learning und Sensorik in der KI-Ideenwerkstatt für Umweltschutz (ZUG)
KI-Ideenwerkstatt für Umweltschutz
c/o Impact Hub
Rollbergstraße 28A, 12053 Berlin
KI-Ideenwerkstatt@z-u-g.org

Ute Kissling-Brenner
Therese Reinke
Referentinnen für Kommunikation und Öffentlichkeitsarbeit,
Zukunft — Umwelt — Gesellschaft (ZUG) gGmbH
Stresemannstraße 69-71, 10963 Berlin
kommunikation-ki-ideenwerkstatt@z-u-g.org

Anmerkungen

1 Mark D. **Wilkinson** u. a., „The FAIR Guiding Principles for scientific data management and stewardship", in: *Sci Data*, Bd. 3, Nr. 1, Art. Nr. 1, März 2016 (DOI: 10.1038/sdata.2016.18).

2 Margot **Belot** u. a., „High Throughput Information Extraction of Printed Specimen Labels from Large-Scale Digitization of Entomological Collections using a Semi-Automated Pipeline", in: *Biodiversity Information Science and Standards*, Bd. 7, S. e112466, Sep. 2023, (DOI: 10.3897/biss.7.112466).

3 **Museum für Naturkunde Berlin**, „Sammlungserschließung und -entwicklung", online unter: museumfuernaturkunde.berlin/de/wissenschaft/sammlungserschliessung-und-entwicklung (letzter Aufruf am 19. Mai 2024).

4 Jana **Hoffmann** u. a., 2022, *Illustration des* Projektes Sammlungserschließung und -entwicklung *am Museum für Naturkunde Berlin*. Publisher: **Museum für Naturkunde Berlin (MfN) — Leibniz Institute for Evolution and Biodiversity Science**. DOI: 10.7479/44ds-qd81.

5 **Museum für Naturkunde Berlin**, „digitize!", online unter: museumfuernaturkunde.berlin/de/presse/pressemitteilungen/digitize (letzter Aufruf am 31. Mai 2024).

6 Christian **Bölling**, „Flexible and scalable label generation workflows in support of collection digitization using persistent and machine-actionable identifiers", Vortrag gehalten auf der *Conference for the Society for the Preservation of Natural History Collections* (SPNHC), International Partner — BHL (Biodiversity Heritage Library) and National Partner — NatSCA (Natural Sciences Collections Association) 2022 (SPNHC2022), Edinburgh, Scotland, UK, 5. Juni 2022. DOI: 10.5281/zenodo.6593465.

7 Jana **Hoffmann** u. a., 2022, *Illustration des Prozesses der Sammlungserschließung am Museum für Naturkunde Berlin*. Publisher: **Museum für Naturkunde Berlin (MfN) — Leibniz Institute for Evolution and Biodiversity Science**. DOI: 10.7479/ryd4-p845.

8 Stefan **Pletschacher** und Apostolos **Antonacopoulos**, „The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", in *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey: IEEE, August 2010, S. 257–260. DOI: 10.1109/ICPR.2010.72.

9 Sabine von **Mering** u. a., „Sharing data, caring for collections. Open data on collection agents affiliated with the Museum für Naturkunde Berlin", in: *Research Ideas and Outcomes*, Bd. 10, S. e118851, Mai 2024, DOI: 10.3897/rio.10.e118851.

10 **KI-Ideenwerkstatt**, „Mit KI verborgene Schätze entdecken", online unter: ki-ideenwerkstatt.de/unterstuetzung-materialien/pilotprojekte/mit-ki-verborgene-schaetze-entdecken (letzter Aufruf am 31. Mai 2024).

11 Informationen über alle Angebote der KI-Ideenwerkstatt finden sich auf der Webseite: ki-ideenwerkstatt.de

12 Weitere Informationen zu den KI-Initiativen des BMUV: bmuv.de/themen/digitalisierung/kuenstliche-intelligenz-fuer-umwelt-und-klima (letzter Aufruf am 31. Mai 2024).

13 Weitere Informationen zu *Civic Coding*: civic-coding.de (letzter Aufruf am 31. Mai 2024).