Policy Brief

Daten als Schlüsselressource:

Die Innovationskraft annotierter Daten im Umweltbereich

Executive Summary

Künstliche Intelligenz (KI) spielt eine bedeutende Rolle in der doppelten Transformation zu Digitalisierung und Nachhaltigkeit. Insbesondere im Umweltbereich sind Daten eine der wertvollsten Ressourcen. Durch den Einsatz von KI und Machine Learning (ML) können Umweltprobleme effizienter analysiert und nachhaltige Lösungen entwickelt werden. Annotierte Daten sind dabei ein zentraler Aspekt.

Das Projekt *LabelledGreenData4All* untersuchte das Potenzial annotierter Umweltdaten und entwickelte strategische Handlungsempfehlungen für deren optimale Nutzung. Die Analysen zeigten, dass die Verfügbarkeit und Qualität von Daten oft durch technische, rechtliche und infrastrukturelle Barrieren eingeschränkt sind.

Fehlende Standardisierung, mangelnde Interoperabilität und unklare rechtliche Rahmenbedingungen behindern den effizienten Austausch und die Nachnutzung umweltrelevanter Daten.

Die Forschungsergebnisse verdeutlichen, dass eine verbesserte Verfügbarkeit und Qualität von Umweltdaten entscheidend sind, um das Innovationspotenzial von KI im Umweltbereich voll auszuschöpfen. Der Aufbau interoperabler Datenräume, klare rechtliche Rahmenbedingungen und eine gezielte Förderung der Datenkompetenz sind wesentliche Schritte, um KI-gestützte Umweltforschung nachhaltig und effizient zu gestalten.

Zusammenfassung der Empfehlungen

- Förderung der Datenkompetenz und Implementierung von Datenmanagement-Richtlinien
- Etablierung von Datenräumen und Datentreuhändern
- Standardisierung als Grundlage für das Datenteilen
- Rechtlicher Rahmen für Risikobewertung
- Verpflichtendes Teilen von Forschungsdaten
- Etablierung von Anreizstrukturen

Einleitung

Digitalisierung und die digitale Transformation führen dazu, dass Daten zu den größten Vermögenswerten gezählt und als eine wesentliche Ressource in allen Bereichen angesehen werden. Umweltdaten und umweltrelevante Daten sind essenziell für inter- und transdisziplinäre Forschung und haben ein enormes Innovationspotential. Durch die rasche Entwicklung von Methoden der künstlichen Intelligenz (KI) ergeben sich innovative Möglichkeiten, um die dringendsten gesellschafts- und umweltpolitischen Herausforderungen unserer Zeit zu analysieren und adäquate Lösungen zu finden. KI-Technologien haben das Potenzial, die Forschung in den Bereichen Umweltschutz und Nachhaltigkeit maßgebend voranzutreiben und zu verändern (BMUV, 2023).

Annotierte Daten bilden die Grundlage für eine gute Modellbildung und dienen als treibende Kraft für die Weiterentwicklung KI-gesteuerter Umweltforschung. Der Einsatz von KI und ML-Ansätzen im Umweltbereich bietet sektorübergreifend signifikante Vorteile. KI ermöglicht es, große Datenmengen auszuwerten und ein besseres Verständnis komplexer Umweltsysteme zu entwickeln, wodurch insbesondere Prozesse für Umweltüberwachung und -schutz optimiert werden können (Thompson, 2023; Branco et al., 2023). Folglich sind sie essenziell für eine Beschleunigung der Politikgestaltung und tragen dazu bei, dass politische Entscheidungen auf einer fundierteren, datenbasierten Grundlage getroffen werden können (Höchtl et al., 2016).

Im wirtschaftlichen Kontext führen KI-Verfahren zu Kostensenkungen und Prozessoptimierung sowie besserer Ressourcenverwaltung. In der Landwirtschaft zum Beispiel unterstützen Deep Learning-Algorithmen die Auswertung von Sensordaten und verbessern den Einsatz von Düngemitteln und Pestiziden durch präzise Entscheidungsvorschläge zu Bewässerung, Nährstoffen und Erträgen (Xu, et al., 2021).

In sozialer und ökologischer Hinsicht fördert KI die Erreichung der Nachhaltigkeitsziele (Sustainable Development Goals - SDGs) durch eine effizientere Analyse und strategische Planung von Maßnahmen im Bereich des Umweltschutzes und der Nachhaltigkeit. So kann ML etwa das Wassermanagement und das Monitoring gefährdeter Arten unterstützen und damit einen direkten Beitrag zu Erfüllung der SGDs leisten (Vinuesa, et al., 2020).

Qualitätsgesicherte und interoperable Daten: Ein Engpass für KI-Anwendungen im Umweltbereich

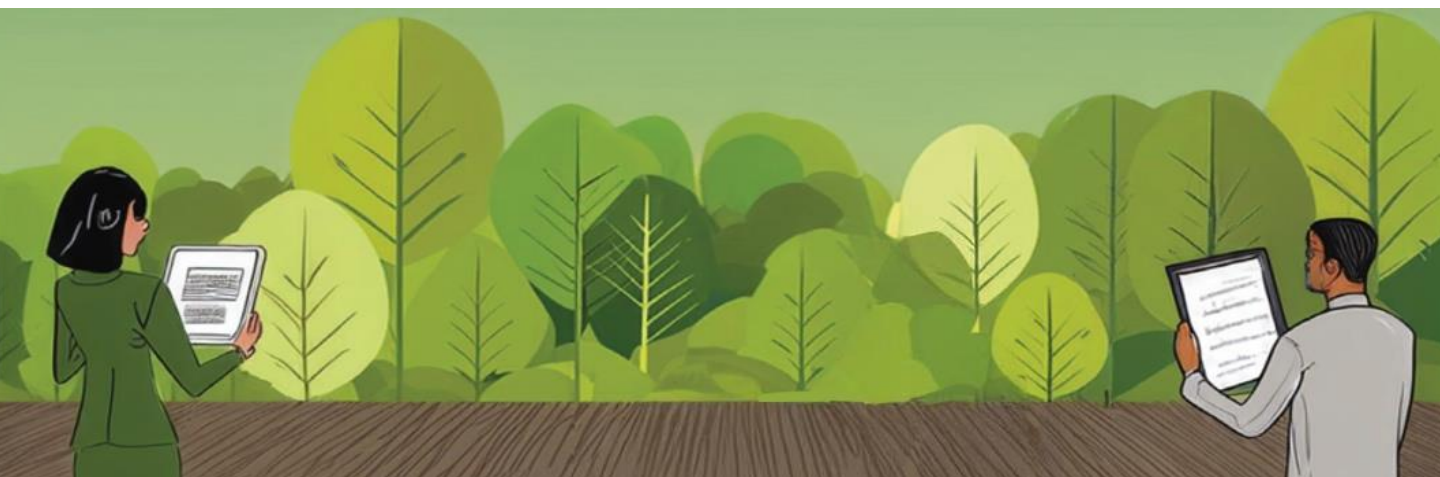
KI-Werkzeuge bieten weitreichende Möglichkeiten, jedoch müssen ihre Ergebnisse zuverlässig sein. Neben erklär- oder interpretierbaren Modellen ist das Vorhandensein von qualitativ hochwertigen, harmonisierten Trainingsdaten essenziell (Halevy et al., 2009). Es besteht der Bedarf nach qualitätsgesicherten und leicht zugänglichen Daten, um Umweltprobleme wie Biodiversitätsverlust und Klimawandel effizient zu adressieren. Beispielsweise in der Forstwirtschaft sind umfassende, standardisierte Inventurdaten erforderlich, die oft aus betrieblichen Gründen nicht zugänglich sind. Im Bereich Biodiversität besteht ein dringender Bedarf an aktuellen, interoperablen Daten mit einer hohen räumlichen Repräsentativität. In der Landwirtschaft wird vor allem eine verbesserte Datenverfügbarkeit für die Erkennung von Pflanzenarten und Schaderregern angestrebt.

Die derzeit bestehenden technischen, rechtlichen und infrastrukturellen Barrieren beschränken die Nutzung und den Zugang zu diesen. Herausfordernd gestalten sich dabei insbesondere Defizite in den Bereichen Datenqualität, Interoperabilität und Datenstandardisierung. Technische Barrieren, wie beispielsweise komplexe Schnittstellen, die Verwendung proprietärer Dateiformate und unzureichende Standards für Trainingsdaten erschweren die effektive Nutzung. Häufig werden existierende Standards, welche die Nachnutzung und automatisierte Weiterverarbeitung vereinfachen würden, nur bedingt beachtet.

Besonders in sensiblen Bereichen wie Forstwirtschaft und Biodiversität fehlen oft standardisierte, qualitätsgesicherte und zugängliche Trainingsdaten. Ein Mangel an Metadaten zur Qualität und Provenienz dieser erschweren zudem die Eignungsbewertung für die weitere Verwendung als Trainingsdaten. Zudem fehlt in vielen Sektoren eine standardisierte Semantik als Schlüsselfaktor zur Verbesserung der Interoperabilität.

Problematisch ist die Verteilung der Daten auf verschiedene Speicherorte, Plattformen und projektzentrische Repositories. Diese Plattformen stellen häufig keine automatisiert nutzbaren Schnittstellen bereit, was die Datennutzbarkeit und -zugänglichkeit einschränkt. Weitere Schwächen sind das Fehlen von Metadaten und die eingeschränkte Auffindbarkeit von Datensätzen sowie das Nichtvorhandensein von Mechanismen für die gemeinsame Nutzung sensibler Daten.

Auch die unsichere Rechtslage wirkt sich negativ auf die Nachnutzung aus. Dazu zählen unter anderem Datenschutz, Datenurheberschaft, Nutzungs- und Verwertungsrechte, Datensicherheit, Geheimhaltungspflichten sowie Schutz weiterer Verwertungsrechte. Folglich tendieren Dateneigentümer*innen dazu, trotz hohem Nachnutzungspotential, weniger Daten zugänglich zu machen. Unklare und komplexe Zugangs- oder Nutzungsbedingungen stellen für Datennutzende ein großes Problem dar und es werden Daten neu aufgenommen, anstatt bestehende Datensätze wiederzuverwenden.



Datenräume, Standards und Anreize: Politische Handlungsempfehlungen für bessere Umweltdaten

Die politischen Handlungsempfehlungen für den Einsatz und das Management von annotierten Daten im Umweltbereich umfassen mehrere strategische Ansätze zur Verbesserung von Datenkompetenz, Datenzugänglichkeit und -sicherheit. Basierend auf unseren Ergebnissen wurden folgende Handlungsempfehlungen formuliert:

- **Förderung der Datenkompetenz und Implementierung von Datenmanagement-Richtlinien**: Um Daten nachhaltig und sicher zu nutzen, ist eine gezielte Förderung der Datenkompetenz in Verwaltung, Forschung und Industrie erforderlich. Insbesondere die Etablierung von Data Governance schützt wertvolle Daten vor Verlust und Missbrauch und ermöglicht die Einhaltung rechtlicher Standards. Verbindliche Datenrichtlinien sollten diesen Prozess unterstützen, etwa durch Schulungen und einheitliche Datenmanagementpraktiken.

- **Etablierung von Datenräumen und Datentreuhändern**: Offene, verteilte Datenökosysteme nach Standards wie der International Data Spaces Association (IDSA) und Gaia-X bieten eine Lösung zur Überwindung isolierter Datenplattformen und fördern diskriminierungsfreien Zugang und nachhaltige Nutzung. Dabei könnten Datentreuhänder*innen als Vermittelnde fungieren, den Zugang insbesondere zu sensiblen Daten vereinfachen und eine langfristige Nutzung von Daten für Wissenschaft und Gesellschaft sicherstellen. Durch die Bereitstellung solcher Daten können Behörden, Forschende und Unternehmen sich auf Innovationen konzentrieren, anstelle immer wieder 80% oder mehr ihrer Entwicklungszeit im Bereich Datenaufbereitung und -annotation aufwenden zu müssen.

- **Standardisierung als Grundlage für das Datenteilen**: Die Anwendung offener Datenformate und standardisierter Schnittstellen stellt eine zentrale Anforderung dar, um Daten nachhaltig nutzbar zu machen. Eine sukzessive Erhöhung der Interoperabilität (Stufenmodell) sollte schrittweise den Ausbau der Datenkompatibilität fördern und langfristig zu einer vollständigen Interoperabilität führen.

- **Rechtlicher Rahmen für Risikobewertung**: Ein evidenzbasierter Ansatz zur objektiven Bewertung der Risiken beim Datenaustausch ist essenziell, um eine verlässliche Grundlage für die Freigabe von Daten zu schaffen und Vorbehalte gegen die Weitergabe abzubauen.

- **Verpflichtendes Teilen von Forschungsdaten**: Daten aus öffentlich geförderten Projekten sollten nach den FAIR-Prinzipien bereitgestellt werden. Dabei sollten Forschende eine exklusive, aber zeitlich begrenzte Auswertungsphase erhalten, um die Daten und ihre Forschungsergebnisse wissenschaftlich zu publizieren. Projekte sollten bereits bei der Antragstellung einen Datenmanagementplan darlegen und könnten im Falle einer Nichteinhaltung Sanktionen, wie das Zurückhalten von Mitteln, erfahren.

- **Etablierung von Anreizstrukturen**: Neben Verpflichtungen sollten Anreize wie steuerliche Vorteile für Datenspenden und immaterielle Bilanzierungsmöglichkeiten für Unternehmen geschaffen werden. Weitere Anreize könnten sein, den Zugang zu spezifischen Funktionen auf Datenbereitstellende zu beschränken oder ihre Sichtbarkeit in Datenportalen zu erhöhen, um den Austausch und die Nutzung von Umweltdaten aktiv zu fördern.

Fazit

Der bessere Zugang zu Umweltdaten und umweltrelevanten Daten führt insbesondere in den Umweltwissenschaften zu einer Steigerung des Innovationspotentials. Annotierte Daten spielen dabei eine besondere Rolle, da qualitativ hochwertige Daten einen direkten Einfluss auf das Training von KI-Modellen, deren Leistungsfähigkeit und Ergebnisqualität haben.

Das grundlegende Problem ist derzeit noch immer der erschwerte Zugang zu Umweltdaten und umweltrelevanten Daten. Dabei bilden die FAIR-Prinzipien einen zentralen Aspekt. Ein FAIRer Umgang erleichtert das Auffinden (*Findability*), und führt zu einer Verbesserung des Zugangs (*Accessibility*), der Interoperabilität (*Interoperability*) und der Nachnutzung (*Reuse*). FAIR bedeutet dabei nicht notwendigerweise „Open“ – auch FAIRe Daten können immer noch eingeschränkt zugänglich und nutzbar sein.

All dies erfordert eine Kombination aus strukturierter und interoperabler Datenspeicherung mit notwendigen Sicherheits- und Datenschutzmaßnahmen, welche eine Transparenz- und Effizienzsteigerung im Umgang mit Daten und eine Stärkung des Innovationspotentials und des wissenschaftlichen Fortschritts ermöglicht.

Um die Nutzbarkeit der Daten zu verbessern, ist es daher unabdingbar, dass die Bereitschaft zum Datenteilen gesteigert wird. Dies kann gelingen, indem existierende Hürden abgesenkt und die Mehrwerte des Datenteilens klar herausgestellt werden.

Über LabelledGreenData4All

Das Projekt **LabelledGreenData4All** - „Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für ML-Modelle“ – zielte darauf ab, für das Umweltressort strategische Empfehlungen zu erarbeiten, in welchen Anwendungsbereichen und mit welchen Daten die größten Potentiale für den Einsatz von Machine Learning (ML)-Modellen bestehen.

Das Team der wetransform GmbH und des Fraunhofer-Institut für Graphische Datenverarbeitung IGD arbeiteten gemeinsam daran, die Innovationskraft annotierter Umweltdatensätze anhand von Anwendungsfeldern zu erforschen, um den Einsatz von KI im Umweltbereich effizient und nachhaltig zu gestalten. Wetransform leitete das Projekt und fokussierte sich auf den Bedarf, das Potential und die Wirkung annotierter Daten. Ein zentraler Aspekt war, die Verfügbarkeit von annotierten Umweltdaten und umweltrelevanten Daten zu verbessern und den sektorübergreifenden Datenaustausch in Datenräumen zu fördern. Das Fraunhofer IGD konzentrierte sich auf die Evaluierung und Analyse vorhandener Annotationsverfahren sowie ihrer Skalierbarkeit und ihrer Ergebnisqualität. Auf dieser Basis wurde ein Vorgehensmodell für die Datenannotation unter Berücksichtigung verschiedener Anwendungsfälle und am Beispiel von Geodaten entwickelt.

Die Forschungsergebnisse von **LabelledGreenData4All** tragen dazu bei, KI im Umweltbereich effizient und nachhaltig zu nutzen und KI-basierten Klima- und Umweltschutz zu beschleunigen.

Autoren & Kontakt

Franziska Hochenegger ist Lead Project Manager bei wetransform GmbH
E-Mail: fh@wetransform.to,

Thorsten Reitz ist Gründer der wetransform GmbH
E-Mail: tr@wetransform.to

Dr. Eva Klien ist Abteilungsleiterin beim Fraunhofer IGD
E-Mail: eva.klien@igd.fraunhofer.de

Referenzen

BMUV. (2023). *Künstliche Intelligenz für Umwelt und Klima*. Von Künstliche Intelligenz als Chancentreiber: https://www.bmuv.de/WS5881 abgerufen

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Computer Society*, 8-12.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., . . . Zhang, L. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(100179).

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., . . . Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications, 11*(233), 1-10 https://doi.org/10.1038/s41467-019-14108-y.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. *Journal of Organizational Computing and Electronic Commerce, 26*(1-2), 147-169.

Thompson, T. (2023). How AI can help to save endagered species. *Nature, 623*, 232-233.

Branco, V. V., Correia, L., & Cardoso, P. (2023). The use of machine learning in species threats and conservation analysis. *Biological Conservation, 283*, https://doi.org/10.1016/j.biocon.2023.110091.

Umwelt Bundesamt

Das auf 12 Monate ausgelegte Forschungsprojekt „Nachhaltigkeitspotentialanalyse für die Zweckmäßigkeit und den Aufwand von Datenannotationen für Machine Learning-Modelle“ aus dem Ressortforschungsplan des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz wurde im Auftrag des Umweltbundesamtes durchgeführt. Forschungskennzeichen: FKZ 3723 11 602 0.