

Online Magazin IAB-Forum

Alternative Daten – ein neuer Rohstoff für die Arbeitsmarktforschung

3. August 2026 | Michael Stops, Maria Sirvent



Klicks, Suchanfragen, Stellenanzeigen, GPS-Signale: Der digitale Alltag hinterlässt Milliarden von Spuren – und die Arbeitsmarktforschung will sie nutzen. Vor allem, um Auswirkungen aktueller Dynamiken schneller zu erfassen, gewinnen solche „alternativen Daten“ an Bedeutung. Doch um sie sinnvoll zu erschließen, müssen Methoden weiterentwickelt sowie technische und rechtliche Voraussetzungen geschaffen werden.

Krisen, Künstliche Intelligenz (KI) und die digitale Transformation verändern den deutschen Arbeitsmarkt grundlegend. Die Forschung muss diese Veränderungen zeitnah und präzise analysieren, stößt dabei aber mit klassischen Datenquellen an Grenzen: Befragungen lassen sich zwar flexibel auf aktuelle Forschungsfragen zuschneiden, sind jedoch zeitaufwändig und kostenintensiv. [Administrative Daten](#) sind dagegen hoch standardisiert und beruhen auf sehr großen Datensätzen. Doch Phänomene, die außerhalb geregelter Prozesse stattfinden, bilden sie

nur selten ab – zum Beispiel Arbeit, die über digitale Plattformen vermittelt wird, oder berufliche Social-Media-Nutzung.

Mit der voranschreitenden Digitalisierung öffnet sich jedoch ein Zugang zu neuen, wertvollen Datenquellen: Jeden Tag erzeugen Menschen Milliarden von Klicks und Suchanfragen, Unternehmen veröffentlichen Stellenanzeigen und Website-Texte, Nutzer produzieren GPS-Signale und Aktivitätsprotokolle. Diese digitalen Spuren sozialer und ökonomischer Prozesse werden nachfolgend als alternative Daten bezeichnet. Sie lassen sich vergleichsweise kostengünstig, schnell und in hoher Frequenz erheben.

Alternative Daten ermöglichen damit granulare Einblicke in organisationsbezogene menschliche Aktivitäten, und das nahezu in Echtzeit. Allerdings müssen Forschende sie aufwändig bereinigen, anreichern und rechtliche Fragen etwa zum Datenschutz klären. Zudem sind die Daten häufig unstrukturiert und decken den Forschungsgegenstand nicht immer repräsentativ ab. Setzt die Arbeitsmarktforschung jedoch innovative methodische Werkzeuge ein und kombiniert alternative mit traditionellen Forschungsdaten, kann sie die drängenden Fragen der heutigen Arbeitswelt [agiler](#) und umfassender untersuchen als je zuvor.

Beispiele für den Nutzen alternativer Daten: Stellenanzeigen, Berufsdatenbanken, Websites

Eine wachsende Zahl an Studien nutzt bereits solche neuen Datenquellen. Im Folgenden werden drei Beispiele näher beschrieben.

Online-Stellenanzeigen schaffen Transparenz über den tatsächlichen Kompetenzbedarf

Stellenanzeigen bilden die gegenwärtige [Arbeitsnachfrage](#) ab und bieten eine Fülle an Informationen über die Berufsbezeichnung hinaus: Sie enthalten Kompetenzanforderungen, Aufgaben und Arbeitsbedingungen. Der [IAB-Kompetenz-Kompass](#) zeigt das exemplarisch: Das Projekt von Michael Stops und anderen zielt darauf ab, Kompetenzanforderungen aus Stellenanzeigen valide und möglichst vollständig zu extrahieren und einen entsprechenden Datensatz für die Arbeitsmarktforschung bereitzustellen. Darüber hinaus lassen sich viele weitere Jobmerkmale sehr spezifisch herauslesen, etwa Arbeitsbedingungen oder nicht lohnbezogene Leistungen.

Berufsdatenbanken sind Indikatoren für den strukturellen Wandel

Berufsdatenbanken wie das deutsche BERUFENET liefern zur gesamten Berufslandschaft detaillierte und aktuelle Informationen, die in klassischen Statistiken oder Taxonomien oft unsichtbar bleiben, zum Beispiel zu Tätigkeiten, Arbeitsorten oder Werkzeugen. Aus diesen Daten lassen sich Indikatoren ableiten, die strukturellen Wandel innerhalb von Berufen messbar machen – also Veränderungen, die nicht zwingend als Berufswechsel in den Statistiken erscheinen.

Ein Beispiel ist der Indikator zur Regulierung des Zugangs in berufliche Tätigkeiten, den Basha Vicari im Jahr 2014 entwickelt hat. Katharina Dengler und andere haben im selben Jahr gezeigt, wie sich der Anteil von Routine- oder kognitiven Tätigkeiten in Berufen messen lässt – eine Grundlage, auf der Katharina Grienberger und weitere 2024 aufbauten, um die Substituierbarkeit von Berufen durch Maschinen und Computer zu untersuchen. Dazu kommen themenspezifische Indizes wie 2018 der „Greening Of Jobs“-Index von Markus Janser oder 2025 der „Home Office Potential“-Indikator von Franziska Bruns und anderen.

Websites geben Auskunft über Unternehmensaktivitäten

Unternehmenswebsites liefern Informationen über Partnerschaften, Vernetzungen und technologische wie strategische Ausrichtungen, die in klassischen Datenquellen verborgen bleiben. Sie können mittels Web-Scraping, dem automatisierten Auslesen von Daten aus Websites, beispielsweise nach Schlüsselbegriffen oder Mustern durchsucht werden, die darauf hindeuten, dass Unternehmen digitale Technologien wie künstliche Intelligenz systematisch nutzen. Das zeigt eine [Studie](#) von Johannes Dahlke und einem internationalen Forschungsteam vom ZEW Mannheim aus dem Jahr 2025.

Methoden und Werkzeuge: Wie man alternative Daten sprechen lässt

Von Online-Stellenanzeigen bis zu Website-Texten: Digitale Spuren sind in ihrer Rohform in der Regel für die wissenschaftliche Analyse ungeeignet. Denn sie werden nicht gezielt für die Forschung erhoben, sondern fallen als Nebenprodukt von Online-Aktivitäten an. Aus dieser Grundlage hochwertige Datensätze für die Arbeitsmarktforschung zu destillieren, ist eine technische wie methodische Herausforderung. Dieser kann mit Verfahren aus Disziplinen wie der Informatik und den Computational Social Sciences, bei der computergestützte Methoden Einblicke in menschliches Verhalten liefern, begegnet werden.

Die Datengewinnung: Flüchtige Spuren im Netz aufspüren und einsammeln

Die Erhebung beginnt direkt an der digitalen Quelle. Flüchtige Informationen auf Webseiten, in Daten verarbeitenden Unternehmenssystemen oder in sozialen Netzen müssen eingesammelt und abgespeichert werden. Offizielle Schnittstellen, so genannte APIs, oder Kooperationen mit den Datenanbietern stehen nicht immer zur Verfügung. Forschende müssen deshalb oft Web-Scraping-Algorithmen entwickeln, die wie digitale User automatisch durchs Netz navigieren und Inhalte einsammeln.

Neben Textdaten können auf diese Weise ebenso andere Datentypen heruntergeladen werden. Dies können etwa Sensor- und Mobilitätsdaten aus GPS- oder Mobilfunksystemen sein, aber auch Interaktions- und Transaktionsdaten aus sogenannten Klickströmen oder Logfiles sowie Bild- und Videodaten.

Das große Aufräumen: Struktur ins Rauschen bringen

Rohdaten aus dem Internet enthalten oft Bestandteile, die die eigentlichen Informationen überlagern oder verzerren: irrelevante Fragmente, Testdatensätze oder störenden HTML-Code. Das Säubern der Daten (Data Cleaning) erfordert eine genaue Auseinandersetzung mit der Quelle und dem Inhalt, um Störfaktoren zu eliminieren. Denn diese können Scheinzusammenhänge erzeugen und das Bild verzerren – etwa, wenn Dubletten in Online-Stellenanzeigen die Nachfrage nach bestimmten Jobs größer erscheinen lassen, als sie in Wirklichkeit ist. Zum Säubern und Strukturieren der Daten gehören auch die Anreicherung mit Metadaten, wie zum Beispiel Zeitstempel, und die Anonymisierung sensibler Inhalte.

Die semantische Entschlüsselung: Signale in Bedeutung verwandeln

Die inhaltliche Operationalisierung ist der Kern der Aufbereitung: Sie entlockt den Rohdaten ihre semantische Bedeutung. Je nach Datentyp stehen hierfür unterschiedliche Verfahren zur Verfügung.

Bei Textdaten kommen Methoden des Natural Language Processing (NLP) zum Einsatz. Dazu gehören sowohl transparente, regelbasierte Wörterbuchmethoden als auch KI-gestützte Verfahren unter Zuhilfenahme großer Sprachmodelle (Large Language Models, LLMs). LLMs ersetzen zwar aufwändige menschliche Codierung, etwa bei der Klassifikation offener Befragungsantworten, sind aber oft (noch) intransparent und erfordern ein tiefes Verständnis der Modellarchitektur.

Auch bei Bild- und [Geodaten](#) setzen Forschende KI-Modelle ein: Diese erkennen Muster in Pixelstrukturen und lesen daraus etwa aus Satellitenbildern menschliche Verhaltens- und Aktivitätsmuster ab – Verkehrsströme zum Beispiel oder Baustellenaktivität. Ähnliche Methoden lassen sich auf strukturierte [Prozessdaten](#) wie Sensordaten, Klicks oder Finanztransaktionen

anwenden, um Milliarden von Datenpunkten zu abstrakten Verhaltensmustern oder Nutzertypen zusammenzufassen und Auffälligkeiten aufzudecken.

Erst komplexe Verfahren übersetzen die Signale in digitalen Spuren, also in wissenschaftlich analysierbare Variablen. Wichtig dabei ist, einschätzen zu können, ob und wie die Wahl bestimmter Aufbereitungsverfahren – etwa eines passenden Auswertungsalgorithmus – die Forschungsergebnisse beeinflusst.

Die intelligente Verknüpfung: Hybride Datensätze sorgen für Verlässlichkeit

Alternative Daten können Verzerrungen enthalten. Dem lässt sich begegnen, indem man sie mit klassischen administrativen Daten oder mit Befragungsdaten verknüpft.

Lässt sich beispielsweise ein durch KI identifiziertes Unternehmensprofil aus Website-Daten mit einem amtlichen Betriebsschlüssel verknüpfen, entsteht ein wertvoller hybrider Datensatz. Die Verknüpfung erweitert nicht nur die analytischen Möglichkeiten, sondern erlaubt auch, die aus digitalen Spuren erzeugten Variablen zu validieren und zu plausibilisieren.

Rechtliche, methodische und technische Hürden bei alternativen Daten

Wer alternative Daten nutzen will, muss erst einige Hürden überwinden. Bevor der erste Datensatz digitaler Spuren aus dem Netz geladen werden kann, muss die Frage der rechtlichen Zulässigkeit geklärt sein. Datenschutz und Urheberrecht setzen enge Grenzen – selbst bei öffentlich einsehbaren Webdaten. Forschende müssen die Wahrung etwaiger Geschäftsgeheimnisse und den Schutz von Persönlichkeitsrechten gegen das Informationsinteresse abwägen – nicht zuletzt, weil Rechtsverstöße die institutionelle Reputation gefährden.

Bei alternativen Daten ist außerdem zu prüfen, ob und für welche Grundgesamtheit sie repräsentative Aussagen erlauben. Weil diese Daten nicht immer kontinuierlich generiert werden, ist diese Überprüfung auf Repräsentativität immer wieder erforderlich. Die oben beschriebenen hybriden Datensätze liefern dafür konkrete Ansätze.

Bei alternativen Daten ist noch ein anderer Aspekt bedeutsam: Neben möglichen Inkonsistenzen bei der Erzeugung kann es auch dazu kommen, dass eine Datenquelle versiegt oder nicht mehr zugänglich ist. Forschende müssen mit einer gewissen Voraussicht mehrere Fragen gegeneinander abwägen: die Beständigkeit und die Qualität der Datenquelle, ihre eigenen verfügbaren Kapazitäten für die Anpassung von Erhebungs- und Auswertungsmethoden sowie die Erreichbarkeit und Qualität ihrer Forschungsziele.

Hinzu kommen infrastrukturelle Grenzen. Die Daten liegen zwar vermeintlich greifbar nah, doch die bestehenden IT-Infrastrukturen klassischer Forschungsumgebungen können deren massenhafte Verarbeitung und Veredelung meist nicht leisten. KI-Verfahren erfordern massive Rechenkapazitäten, und um große, verrauschte Datenmengen aufzubereiten, müssen komplexe Software-Pipelines entwickelt werden. Forschungsinstitute benötigen also nicht nur entsprechende Infrastrukturen, sondern müssen auch deren dauerhafte Wartung vorsehen, wenn sie die Daten nicht nur einmalig erheben wollen.

Außerdem ist die Prozesskette selbst komplex. KI-Methoden versprechen, große Datenmengen massenhaft und effizient zu verarbeiten. Doch Mensch und KI treffen in jedem Prozessschritt weiterführende Entscheidungen, die das Risiko von Informationsverlust oder unbewusster Verzerrung bergen. Dabei ist insbesondere die begrenzte Erklärbarkeit gegenwärtiger KI-Modelle ein Problem.

Die KI ist zwar ein hochproduktiver, aber oft unbesonnener „Forschungsassistent“. Erstens lassen sich Entscheidungswege von neuronalen Netzwerken schwer nachvollziehen. Zweitens können KI-Modelle unbemerkt halluzinieren, also plausible, aber faktisch falsche Artefakte produzieren. Ohne erklärbare KI (explainable AI), deren Konzeption und Umsetzung ein eigenes Forschungsthema ist, lässt sich wissenschaftliche Reliabilität kaum gewährleisten.

Trotz der Hürden und Risiken wäre es kurzfristig, auf das Potenzial alternativer Daten zu verzichten – sowohl methodisch als auch inhaltlich. Die Arbeitsmarktforschung sollte deshalb, wie andere Wissenschaftsdisziplinen, eigene Data-Science-Kompetenzen aufbauen, auch durch einen intensiven interdisziplinären Austausch mit der Informatik. Nur so lassen sich wachsende Rechenleistung und moderne Algorithmen für die eigenen Forschungsbelange voll nutzen.

Entscheidend bleibt dabei ein reflektierter Umgang mit den Werkzeugen. Neue Verfahren eröffnen wertvolle Analyseperspektiven, doch methodische Strenge und nachvollziehbare Ergebnisse müssen stets die Grundlage belastbarer wissenschaftlicher Erkenntnis bleiben.

Fazit

Alternative Daten eröffnen der Arbeitsmarktforschung vielversprechende Wege, um digitale Phänomene oder kurzfristige Veränderungen nahezu in Echtzeit zu erfassen. Um diese zu erschließen, müssen Forschende verschiedene methodische, rechtliche und technische Voraussetzungen erfüllen, die auch der besonderen, nicht immer kontinuierlichen Generierung solcher Daten Rechnung tragen.

Hybride Datensätze aus bewährten Referenzdaten und hochfrequenten digitalen Daten machen schnellere Analysen möglich, ohne wissenschaftliche Prämissen für belastbare Ergebnisse aufzugeben. Arbeiten Mensch und KI zusammen, lassen sich riesige Datenmengen in einem bisher unerreichten Maßstab verarbeiten. Dabei darf die vermeintliche Magie der KI nicht über methodische Lücken hinwegtäuschen. Interdisziplinäre Kollaboration und Kompetenzaufbau erweitern die methodische Bandbreite und schaffen Sicherheit im Umgang mit KI-Verfahren.

Unter diesen Voraussetzungen bereichert die Erschließung alternativer Datenquellen die Arbeitsmarktforschung gleich doppelt: Sie hält Schritt mit der rasanten Transformation der Arbeitswelt und liefert zudem fundierte Erkenntnisse für ihre Gestaltung.

Literatur

Bruns, Franziska; Matthes, Britta; Stops, Michael (2025): [HOP – An Occupational Indicator for the Potential of Work](#). Jahrbücher für Nationalökonomie und Statistik.

Dahlke, Johannes et al. (2025): [The WebAI paradigm of innovation research: Extracting insight from organizational web data through AI](#). ZEW-Discussion Paper Nr. 25–019.

Dengler, Katharina; Matthes, Britta; Paulus, Wiebke (2014): [Berufliche Tasks auf dem deutschen Arbeitsmarkt. Eine alternative Messung auf Basis einer Expertendatenbank](#). FDZ-Methodenreport Nr. 12 (de).

Grienberger, Katharina; Matthes, Britta; Paulus, Wiebke (2024): [Folgen des technologischen Wandels für den Arbeitsmarkt: Vor allem Hochqualifizierte bekommen die Digitalisierung verstärkt zu spüren](#). IAB-Kurzbericht Nr. 5.

Janser, Markus (2018): [The greening of jobs in Germany: First evidence from a text mining based index and employment register data](#). IAB-Discussion Paper Nr. 14.

Stops, Michael; EzELDin, Asma Ahmed.; Heckel, Marie-Christine; Heß, Pascal; Klevinghaus, Hauke; Malfertheiner, Verena; Metzger, Lina-Jeanette; Peede, Lennert (2025): [„IAB-Kompetenz-Kompass“ Bereitstellung von Stellentext-Daten als strukturierter Datensatz](#). IAB-Forschungsbericht Nr. 1.

Vicari, Basha (2014): [Grad der standardisierten Zertifizierung des Berufs. Ein Indikator zur Messung institutioneller Eigenschaften von Berufen \(KldB 2010, KldB 1988\)](#). FDZ-Methodenreport Nr. 4 (de).

In aller Kürze

- Alternative Daten lassen sich aus „digitalen Spuren“ gewinnen, die von Organisationen und Betrieben durch digitale Prozesse und Aktivitäten hinterlassen werden.
- Aufgrund ihrer Granularität, Aktualität, schnellen Verfügbarkeit und geringen Erhebungskosten im Vergleich zu klassischen Erhebungen sind sie interessant für die Forschung.
- Im Rohzustand sind solche Daten häufig unstrukturiert und müssen daher bereinigt und semantisch entschlüsselt werden. KI-Methoden, die wissenschaftlichen Ansprüchen genügen, können hierzu beitragen.
- Für den Zugang sind spezifische rechtliche Grundsätze, insbesondere Fragen des Datenschutzes, zu beachten und technische Hürden zu überwinden.
- Die Kombination mit klassischen Daten bietet Validierungsmöglichkeiten und erweitert gleichzeitig das Analysepotenzial erheblich.

Bild: HONGWEI/stock.adobe.com

DOI: [10.48720/IAB.FOO.20260803.01](https://doi.org/10.48720/IAB.FOO.20260803.01)

Zitationshinweis

Michael Stops; Maria Sirvent (2026): Alternative Daten – ein neuer Rohstoff für die Arbeitsmarktforschung , In: Online Magazin IAB-Forum 3. August 2026, <https://iab-forum.de/alternative-daten-ein-neuer-rohstoff-fuer-die-arbeitsmarktforschung/>, Abrufdatum: 3. August 2026

Lizenzhinweis

Diese Publikation ist unter folgender Creative-Commons-Lizenz veröffentlicht: Namensnennung – Weitergabe unter gleichen Bedingungen 4.0 International (CC BY-SA 4.0):
<https://creativecommons.org/licenses/by-sa/4.0/deed.de>