

Balanciertes Fünf-Städte-Audit kommunaler Publikationen

Pilotstudie zu lokaler institutioneller Exposition und langfristiger Aggregierbarkeit

Autor: Benjamin Metzиг

Projekt: Wissenschaftswelle.de

Datum: 12. Juni 2026

Publikationsform: Deutsche PDF-Fassung

Deutschsprachige Pilotstudie zur Ergänzung des Forschungsberichts Das private Palantir.

Balanciertes Fünf-Städte-Audit kommunaler Publikationen

Pilotstudie zu lokaler institutioneller Exposition und langfristiger Aggregierbarkeit

Autor: Benjamin Metzig

Projektkontext: Wissenschaftswelle.de

Status: Publikationsfähige deutschsprachige Pilotstudie zur Ergänzung des Forschungsberichts *Das private Palantir*

Datum: 12. Juni 2026

Empfohlene Zitierweise: Metzig, B. (2026). *Balanciertes Fünf-Städte-Audit kommunaler Publikationen: Pilotstudie zu lokaler institutioneller Exposition und langfristiger Aggregierbarkeit*.

Wissenschaftswelle.de.

Abstract

Öffentliche Dokumente bleiben häufig sozial geschützt, weil sie verstreut, kontextgebunden und aufwendig zu rekombinieren sind. Diese Pilotstudie untersucht, ob wiederkehrende kommunale Publikationen diesen Schutz schwächen, indem sie lokale institutionelle Spuren dauerhaft und maschinenlesbar machen. Dafür wurde ein balancierter Korpus von 125 offiziellen kommunalen PDF-Publikationen aus fünf deutschen Städten aufgebaut: Heidelberg, Mannheim, Leipzig, Frankfurt am Main und Nürnberg. Jede Stadt steuerte 25 live-validierte serielle PDFs bei. Die Studie untersuchte keine einzelnen Personen und speicherte keine Namen, Rohtexte, Roh-PDFs oder personenbezogenen Datensätze. Stattdessen wurden dokumentbezogene Aggregatindikatoren berechnet: Dichte namensähnlicher Kandidaten, Kontaktmuster, Datumsangaben, Rollenterme, schul-/jugend-/familienbezogene Begriffe, PDF-Metadaten und Extraktionsqualität. Alle 125 Dokumente wurden erfolgreich verarbeitet. Alle enthielten E-Mail- und Telefonnummernmuster; 79 enthielten PDF-Autorenmetadaten; die Extraktionsqualität war bei allen Dokumenten hoch. Der mittlere Aggregierbarkeitsscore lag bei 8,92; 48 Dokumente wurden als hoch und 77 als sehr hoch eingestuft. Die Befunde zeigen keine Rechtswidrigkeit, keinen konkreten Schaden und keinen realen Missbrauch. Sie stützen eine engere Aussage: Routinemäßige kommunale Publikation kann dauerhafte, vergleichbare und maschinenlesbare Inputs für spätere Aggregation erzeugen. Damit wird Publikationsgovernance zu einer vorgelagerten Datenschutzfrage in Debatten über KI-gestützte Profilierung und praktische Obskürität.

Schlagwörter: praktische Obskürität; kommunale Publikationen; öffentliche Daten; Datenschutz; KI-Governance; institutionelle Exposition; PDF-Metadaten; Datenaggregation; Kommunalverwaltung; öffentliche Dokumente.

Beitrag

Diese Studie ergänzt das Forschungsprogramm *Das private Palantir* in drei Punkten. Erstens liefert sie einen empirischen Pilot zu einer häufig konzeptuell formulierten These: Öffentlichkeit beseitigt nicht die Datenschutzrelevanz von Aggregation. Zweitens demonstriert sie eine risikoarme Auditmethode, die keine personenbezogenen Dossiers erzeugt und dennoch Publikationsexposition messbar macht. Drittens verbessert sie frühere explorative Korpusfassungen durch gleiche Stadtzahlen und vergleichbare Quellentypen.

Die Studie ist ein methodischer und deskriptiver Pilot. Sie ist keine Prävalenzstudie und keine rechtliche Bewertung einzelner Dokumente.

Forschungsfrage

Wie aggregierbar sind wiederkehrende öffentliche kommunale PDFs über vergleichbare städtische Publikationskontexte hinweg, und welche dokumentbezogenen Merkmale erhöhen das Risiko, dass öffentliche Spuren später in private Intelligenz überführt werden können?

Studiendesign

Die Studie nutzt einen balancierten Fünf-Städte-Korpus serieller kommunaler Publikationen. Untersuchungseinheit ist das Dokument, nicht die Person. Der Fokus auf wiederkehrenden offiziellen Stadtpublikationen wurde gewählt, weil sie öffentlich, dauerhaft, institutionell legitim und zwischen Kommunen vergleichbar sind.

Stadt	Dokumente	Quellentyp
Heidelberg	25	Stadtblatt, serielle PDFs
Mannheim	25	Amtsblatt, serielle PDFs
Leipzig	25	Amtsblatt und E-Amtsblatt, serielle PDFs
Frankfurt am Main	25	Amtsblatt, serielle PDFs
Nürnberg	25	Amtsblatt, serielle PDFs

Der Korpus umfasst insgesamt 125 öffentlich zugängliche PDFs. Alle URLs wurden vor Aufnahme live validiert. Die Auswahl beruhte auf menschlich definierten offiziellen Quellenmustern und wurde anschließend rechnerisch auf Erreichbarkeit und PDF-Inhaltstyp geprüft. Dadurch bleibt die Quellenauswahl menschlich verantwortet, während der Validierungsschritt reproduzierbar ist.

Ein- und Ausschlusskriterien

Aufgenommen wurden Dokumente, die alle folgenden Kriterien erfüllten:

- offizielle kommunale Publikation oder Stadt-/Amtsblatt;

- öffentliches PDF auf einer offiziellen kommunalen Domain oder einem statischen kommunalen Asset-Host;
- Live-Antwort zum Validierungszeitpunkt;
- Eignung für seriellen Vergleich innerhalb einer städtischen Publikationsreihe;
- dokumentbezogene Analyse möglich, ohne Rohtexte oder Namen zu speichern.

Ausgeschlossen wurden nicht erreichbare Dokumente, Nicht-PDFs, doppelte URLs, nicht einschlägige Dokumente und Dokumente, die personenzentrierte Interpretation erfordert hätten.

Ethische Schutzmaßnahmen

- Keine Person wurde gezielt untersucht.
- Es wurden keine Personenprofile erstellt.
- Es wurden keine Namen gespeichert.
- Es wurden keine Rohtexte gespeichert.
- Es wurden keine Roh-PDFs als Publikationsmaterial gespeichert.
- Es wurde keine Gesichts- oder Bildanalyse durchgeführt.
- Die Ergebnisse enthalten nur dokument- und korpusbezogene Aggregatindikatoren.
- Die Resultate werden als institutionelle Expositionsmuster interpretiert, nicht als Bewertung von Personen.

Diese Schutzmaßnahmen sind Teil des Studiendesigns. Ziel ist, Publikationsgovernance zu untersuchen, ohne den kritisierten Schadensmechanismus selbst zu reproduzieren.

Messgrößen

Die Auswertung rief jedes öffentliche PDF zur Verarbeitung ab, extrahierte Text und PDF-Metadaten und berechnete dokumentbezogene Indikatoren. Roh-PDFs, extrahierte Rohtexte, Namen und Personenprofile wurden nicht als Publikationsmaterial gespeichert.

Indikator	Interpretation
Dichte namensähnlicher Kandidaten	Konservativer Proxy für identifizierbare menschliche Referenzen pro 10.000 Textzeichen. Kein Named-Entity-Modell; Fehlpositive sind erwartbar.
Kontaktdichte	E-Mail- und Telefonnummernmuster pro 10.000 Textzeichen.
Datumsdichte	Datumsangaben pro 10.000 Textzeichen, als Hinweis auf zeitliche Verknüpfbarkeit.
Rollendichte	Begriffe wie Bürgermeister, Rat, Vorstand, Vorsitz, Koordination, Lehrkraft, Trainer oder Kontaktperson.
Schul-/Jugend-/Familienbegriffe	Begriffe zu Kindern, Schule, Klassen, Jugend, Kita, Eltern oder Bildung.
Archiv-/Jahreskontext	Abstand zwischen Jahrhinweis und 2026.
PDF-Metadaten	Vorhandensein von Autor-, Creator-, Producer- oder Erstellungsdatumsmetadaten.

Indikator	Interpretation
Dokumenttyp	Unterscheidung serieller Publikationsformate.
Extraktionsqualität	Niedrig, mittel oder hoch auf Basis extrahierter Textzeichen pro Seite.

Der Aggregierbarkeitsscore ist ein heuristischer Dokumentindex. Er ist kein rechtlicher Risikoscore und bewertet keine Personen. Er schätzt, wie viel Reibung ein Dokument für spätere personenbezogene Aggregation entfernt. Der begleitende PDF-Datenanhang berichtet die dokumentbezogenen Aggregatindikatoren und den finalen Score für jedes Dokument.

Ergebnisse

Alle 125 Dokumente wurden erfolgreich verarbeitet.

Kennzahl	Ergebnis
Geprüfte Dokumente	125
Erfolgreich verarbeitet	125
Städte	5
Dokumente pro Stadt	25
Dokumente mit manuellem Minderjährigen-Kontext-Flag	0
Dokumente mit E-Mail-Mustern	125
Dokumente mit Telefonnummernmustern	125
Dokumente mit PDF-Autorenmetadaten	79
Mittlerer Aggregierbarkeitsscore	8,92
Median	9
Maximum	13

Risikostufe	Dokumente
Hoch	48
Sehr hoch	77

Die Extraktionsqualität war bei allen 125 Dokumenten hoch. Das ist wichtig, weil die Studie maschinenlesbare Exposition misst. Schlecht extrahierbare PDFs könnten zwar über OCR oder Bildanalyse lesbar sein, dies lag aber außerhalb des aktuellen Protokolls.

Stadtbezogene deskriptive Ergebnisse:

Stadt	Dokumente	Mittelwert	Median	Hoch	Sehr hoch	Autor-Metadaten
Frankfurt am Main	25	10,16	11	6	19	25
Heidelberg	25	7,68	8	23	2	0

Stadt	Dokumente	Mittelwert	Median	Hoch	Sehr hoch	Autor-Metadaten
Leipzig	25	9,28	9	0	25	25
Mannheim	25	8,80	9	8	17	25
Nürnberg	25	8,68	9	11	14	4

Die stärksten aggregierten Score-Beiträge kamen von namensähnlicher Kandidatendichte, PDF-Metadaten, Kontaktdichte, Datumsdichte, Rollentermen und schul-/jugend-/familienbezogener Begriffsdichte. Obwohl kein Dokument manuell als dedizierter Minderjährigenkontext markiert wurde, enthielten gewöhnliche kommunale Publikationen häufig Begriffe zu Schule, Jugend, Kinderbetreuung, Bildung oder Familie. Das ist analytisch relevant, darf aber nicht mit gezielter Extraktion von Kinderidentitäten verwechselt werden.

Interpretation

Der balancierte Fünf-Städte-Korpus stärkt den Mechanismus des Hauptberichts enger, aber vergleichbarer als frühere gemischte Korpora. Er zeigt, dass routinemäßige serielle kommunale Publikationen konsistent Kontaktkanäle, Rollen, Daten, namensähnliche Referenzen, institutionelle Kontexte und Metadaten in maschinenlesbarer Form kombinieren können.

Die Studie zeigt nicht, dass diese Dokumente für sich genommen rechtswidrig, fahrlässig oder schädlich sind. Die meisten erfüllen legitime öffentliche Funktionen. Der Befund ist subtiler: Öffentliches Interesse an Veröffentlichung und spätere maschinelle Aggregierbarkeit können zeitlich auseinanderfallen.

Die Unterschiede zwischen Städten sind ebenfalls wichtig. Sie deuten darauf hin, dass Exposition nicht nur davon abhängt, was Kommunen veröffentlichen, sondern auch von Publikationsworkflows: Metadatenpraxis, Formatierung, Kontaktkonventionen und wiederkehrende Dokumentstruktur beeinflussen spätere Aggregierbarkeit.

Beziehung zum Forschungsbericht *Das private Palantir*

Der Hauptbericht argumentiert, dass KI-Agenten, multimodale Systeme, lange Kontextfenster, Speicher und fallende Kosten praktische Obskurität schwächen können, weil sie den Aufwand für Sammlung und Rekombination öffentlicher Spuren reduzieren. Diese Pilotstudie testet KI-Agenten nicht direkt. Sie untersucht eine vorgelagerte Bedingung: ob öffentliche institutionelle Dokumente bereits maschinenlesbare, dauerhafte und wiederkehrende Inputs bereitstellen, die später rekombiniert werden könnten.

Damit stützt die Studie die Governance-Perspektive des Berichts. Das Problem liegt nicht nur im nachgelagerten Missbrauch durch böswillige Akteure. Es ist auch ein Publikationsgovernance-Problem: Lokale Institutionen können unbeabsichtigt dauerhafte Inputs für Private Intelligence Automation erzeugen.

Grenzen

- Die Stichprobe ist zwischen fünf Städten balanciert, aber nicht repräsentativ für alle deutschen lokalen Institutionen.
- Der Korpus konzentriert sich auf serielle kommunale Publikationen und erfasst nicht die Vielfalt von Schulen, Vereinen, Veranstaltungsprogrammen, Sport, Kultur und Formularen.
- Die Indikatoren sind heuristisch und enthalten Fehlpositive.
- Namensähnliche Kandidaten sind keine verifizierten Namen.
- Bilder, Scans, Bildunterschriften und Bild-Text-Kopplung wurden nicht analysiert.
- Suchmaschinenindexierung, robots.txt, Plattform-Caching und Webarchivpersistenz wurden nicht getestet.
- Es wurde nicht gemessen, ob KI-Agenten die Extraktion gegenüber normaler Suche verbessern.
- Es wird kein kausaler Zusammenhang zu realem Missbrauch behauptet.
- Externe Validierung oder Intercoder-Reliabilität des Score-Modells stehen noch aus.

Empfehlungen für Institutionen

- Wiederkehrende Amts- und Stadtblätter sowie Archive auf langfristige Aggregierbarkeit prüfen.
- Schul-, Jugend-, Familien- und Kinderbetreuungskontexte sensibel behandeln, auch wenn keine einzelnen Kinder genannt werden.
- Unnötige personenbezogene Kontaktdaten vermeiden, wo Rollenkontakte genügen.
- Unnötige PDF-Metadaten vor Veröffentlichung entfernen.
- Veröffentlichungsdatum, Kontextlabel und Überprüfungsdatum ergänzen.
- Review- und Retention-Prozesse für herunterladbare Archive einführen.
- Rollenkontakte gegenüber personenbezogenen Kontakten bevorzugen, wo Verantwortlichkeit dies erlaubt.
- Dokumentieren, warum personenbezogene Informationen veröffentlicht werden, wann sie überprüft werden und wer für Lösch- oder Korrekturbitten zuständig ist.

Datenverfügbarkeit

Die öffentliche Quellenliste, Korpuszusammenfassung, stadtbezogene Ergebnisse und dokumentbezogene Aggregatindikatoren, die zur Beurteilung der Studie nötig sind, werden im begleitenden PDF-Datenanhang bereitgestellt: *Balanciertes Fünf-Städte-Audit kommunaler Publikationen: PDF-Datenanhang*. Der Anhang enthält nur aggregierte Dokumentindikatoren. Roh-PDFs, extrahierte Rohtexte, Namen und personenbezogene Datensätze sind nicht Teil des Publikationsmaterials.

Die zur Validierung und Aggregation genutzten rechnerischen Hilfen sind nicht erforderlich, um das PDF-Publikationspaket zu interpretieren. Die Methodik beschreibt Korpuslogik, Indikatoren und ethische Schutzmaßnahmen.

Ethikerklärung

Die Studie beinhaltete keine Human-Subjects-Forschung im Sinne von Interaktion, Intervention, Rekrutierung oder personenbezogener Datenerhebung. Untersucht wurden öffentliche institutionelle Dokumente auf Dokumentenebene. Das Protokoll wurde bewusst so gestaltet, dass keine Namenslisten oder Profile entstehen.

Autorennotiz

Diese Pilotstudie wurde von Benjamin Metzиг für Wissenschaftswelle.de als Teil des Forschungsberichts *Das private Palantir* entwickelt. KI-Werkzeuge wurden unterstützend für Strukturierung, Sprachfassung, Codeentwürfe und Konsistenzprüfungen genutzt. Redaktionelle Verantwortung, Schlussbewertung, Quellenauswahl, Interpretation und Veröffentlichungsentscheidung liegen bei Benjamin Metzиг.

Quellen

Hartzog, W. & Stutzman, F. (2010). The case for online obscurity. SSRN.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1597745

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1), 119-158.

<https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/>

European Union. (2016). Regulation (EU) 2016/679, General Data Protection Regulation. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>

European Data Protection Board. (2024). Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR. https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf

Datenschutzkonferenz. (2024). Orientierungshilfe KI und Datenschutz.

https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf