

Balanced Five-City Municipal Publication Audit

A Pilot Study on Local Institutional Exposure and Long-Term Aggregability

Author: Benjamin Metzиг

Affiliation: Wissenschaftswelle.de

Date: June 12, 2026

Corpus: 125 public municipal-publication PDFs, five German cities

Publication-ready pilot-study manuscript for integration with The Private Palantir research report.

Balanced Five-City Municipal Publication Audit

A Pilot Study on Local Institutional Exposure and Long-Term Aggregability

Author: Benjamin Metzig

Affiliation: Wissenschaftswelle.de

Status: Publication-ready pilot-study manuscript for integration with *The Private Palantir* research report

Date: June 12, 2026

Recommended citation: Metzig, B. (2026). *Balanced Five-City Municipal Publication Audit: A Pilot Study on Local Institutional Exposure and Long-Term Aggregability*. Wissenschaftswelle.de.

Abstract

Public documents often remain socially protected because they are dispersed, contextual, and costly to recombine. This pilot study examines whether recurring municipal publications weaken that protection by making local institutional traces durable and machine-readable. A balanced corpus of 125 official municipal-publication PDFs was assembled from five German cities: Heidelberg, Mannheim, Leipzig, Frankfurt am Main, and Nuremberg. Each city contributed 25 live-validated serial PDFs. The study did not target individuals and did not store names, raw text, raw PDFs, or person-level records. Instead, it computed document-level aggregate indicators for name-like candidate density, contact patterns, date patterns, role terms, school/youth/family-related terms, PDF metadata, and extraction quality. All 125 documents were successfully processed. All contained email and phone-number patterns; 79 contained PDF author metadata; extraction quality was high for every document. The mean aggregability score was 8.92, with 48 documents classified as high and 77 as very high on a heuristic document-level index. The findings do not show illegality, harm, or actual misuse. They support a narrower claim: routine municipal publication can create durable, comparable, machine-readable inputs for later aggregation. This makes publication governance an upstream privacy issue in debates about AI-assisted profiling and practical obscurity.

Keywords: practical obscurity; municipal publications; public data; privacy; AI governance; institutional exposure; PDF metadata; data aggregation; local government; public records.

Contribution

This study contributes to the broader *Private Palantir* research programme in three ways.

First, it provides an empirical pilot for a claim that is otherwise often made conceptually: public availability does not eliminate the privacy relevance of aggregation. Second, it demonstrates a low-risk audit method that avoids creating person-level dossiers while still measuring publication-level exposure.

Third, it improves on earlier exploratory versions of the corpus by using equal city counts and comparable source types.

The study should be read as a methodological and descriptive pilot, not as a prevalence study and not as a legal assessment of individual documents.

Research Question

How aggregable are recurring public municipal PDFs across comparable city-level publication contexts, and which document-level features increase the risk that public traces can later be converted into private intelligence?

Study Design

The study uses a balanced five-city corpus of serial municipal publications. The unit of analysis is the document, not the person. The corpus focuses on recurring official city publications because they are public, durable, institutionally legitimate, and comparable across municipalities.

City	Documents	Source type
Heidelberg	25	Stadtblatt serial PDFs
Mannheim	25	Amtsblatt serial PDFs
Leipzig	25	Amtsblatt and E-Amtsblatt serial PDFs
Frankfurt am Main	25	Amtsblatt serial PDFs
Nuremberg	25	Amtsblatt serial PDFs

The corpus contains 125 publicly accessible PDFs in total. All URLs were live-validated before inclusion. The selection was based on human-defined official source patterns and then checked computationally for reachability and PDF content type. This preserves human source selection while making the validation step reproducible.

Inclusion and Exclusion Criteria

Documents were included when they met all of the following criteria:

- official municipal publication or city bulletin;
- public PDF available through an official municipal domain or static municipal asset host;
- live response at the time of validation;
- suitable for serial comparison within a city publication series;
- document-level analysis possible without storing raw text or names.

Documents were excluded when they were unavailable, non-PDF, duplicate URLs, unrelated to municipal publication practice, or likely to require person-level interpretation.

Ethical Safeguards

- No person was targeted.
- No person profiles were created.
- No names were stored.
- No raw PDF text was stored.
- No raw PDFs were stored by default.
- No facial recognition or image analysis was conducted.
- The output contains only document-level and corpus-level indicators.
- The results are interpreted as institutional exposure patterns, not as assessments of individuals.

These safeguards are part of the study design. They are not merely operational precautions: the aim is to examine publication governance without reproducing the harm mechanism under study.

Measures

The audit script fetched each public PDF in memory, extracted text and PDF metadata, and computed document-level indicators. It did not persist raw PDFs, extracted text, names, or person-level records.

Indicator	Interpretation
Name-like candidate density	A conservative proxy for identifiable human references, measured per 10,000 text characters. It is not a named-entity model and includes false positives.
Contact density	Email and phone-number patterns per 10,000 text characters.
Date density	Date patterns per 10,000 text characters, indicating temporal linkability.
Role density	Terms such as mayor, council, board, chair, coordinator, teacher, trainer, or contact person.
School/youth/family term density	Terms related to children, schools, classes, youth, daycare, parents, or education.
Archive age	Number of years between the document's year hint and 2026.
PDF metadata	Presence of author, creator, producer, or creation-date metadata.
Document type	Used to distinguish different serial-publication formats.
Extraction quality	A low, medium, or high label based on extracted text characters per page.

The resulting aggregability score is a heuristic document-level index. It is not a legal risk score and does not rank people. It estimates how much friction a document removes for later personal aggregation. The index is intentionally transparent: the accompanying PDF data appendix reports document-level aggregate indicators and the final score for each document.

Results

All 125 documents were successfully processed.

Metric	Result
Documents audited	125
Successfully processed	125
Cities	5
Documents per city	25
Documents with manual minor-context flag	0
Documents containing email patterns	125
Documents containing phone-number patterns	125
Documents with PDF author metadata	79
Mean aggregability score	8.92
Median aggregability score	9
Maximum aggregability score	13

Risk levels were distributed as follows:

Risk level	Documents
High	48
Very high	77

Extraction quality was high for all 125 documents. This matters because the pilot measures machine-readable exposure; poorly extractable PDFs may still be readable through OCR or image analysis, but that was outside the current protocol.

City-level descriptive results show meaningful differences between otherwise comparable publication series:

City	Documents	Mean score	Median score	High	Very high	Author metadata
Frankfurt am Main	25	10.16	11	6	19	25
Heidelberg	25	7.68	8	23	2	0
Leipzig	25	9.28	9	0	25	25
Mannheim	25	8.80	9	8	17	25
Nuremberg	25	8.68	9	11	14	4

The strongest aggregate score contributors were name-like candidate density, PDF metadata, contact density, date density, role terms, and school/youth/family-related term density. Although no document was manually flagged as a dedicated minor-related document, ordinary municipal publications frequently contained school, youth, childcare, education, or family-related terms. This is analytically relevant but should not be confused with targeted extraction of children's identities.

Interpretation

The balanced five-city corpus strengthens the main report's mechanism in a narrower but more comparable way than the earlier mixed corpus. It shows that routine municipal serial publications can consistently combine contact channels, roles, dates, name-like references, institutional contexts, and metadata in machine-readable form.

The study does not show that these documents are unlawful, negligent, or harmful in isolation. Most have legitimate public functions. The finding is more subtle: public-interest publication and later machine aggregability can diverge over time. A document can be appropriate at publication and still become a component in future profiling when combined with search, OCR, long-context analysis, memory, and repeated updating.

The city-level differences are also important. They suggest that exposure is not only a function of what municipalities publish, but also of publication workflows: metadata practices, formatting choices, contact conventions, and recurring document structure all affect later aggregability.

Relationship to *The Private Palantir*

The main report argues that AI agents, multimodal systems, long-context analysis, memory, and falling costs can weaken practical obscurity by reducing the effort needed to collect and recombine public traces. This pilot does not test AI agents directly. Instead, it examines an upstream condition: whether public institutional documents already provide machine-readable, durable, and recurring inputs that could later be recombined.

The study therefore supports the report's governance framing. The problem is not only downstream misuse by malicious actors. It is also a publication-governance problem: local institutions can unintentionally produce durable inputs for private intelligence automation.

Limitations

- The sample is balanced across five cities but not representative of all German local institutions.
- The corpus focuses on serial municipal publications and does not capture the broader diversity of schools, associations, event programmes, sports clubs, cultural programmes, and administrative forms.
- The indicators are heuristic and include false positives.
- Name-like candidates are not verified names.
- The study did not analyze images, scans, photo captions, or image-text coupling.
- The study did not test search-engine indexing, robots.txt behavior, platform caching, or web-archive persistence.
- The study did not measure whether AI agents improve extraction compared with ordinary search.
- No causal link to actual misuse is claimed.
- No inter-coder validation or external audit of the scoring model has yet been conducted.

Recommendations for Institutional Practice

- Review recurring municipal bulletins and archives for long-term aggregability.
- Treat school, youth, family, and childcare references as sensitive contexts even when individual children are not named.
- Avoid unnecessary personal contact details where role-based addresses are sufficient.
- Remove unnecessary PDF metadata before publication.
- Add publication dates, context labels, and review dates to public documents.
- Create retention or review policies for downloadable archives.
- Prefer institutional role contacts over named personal contacts where accountability does not require naming.
- Keep a public record of why personal information is published, when it should be reviewed, and who is responsible for removal requests.

Data Availability

The public source URL list, corpus summary, city-level results, and document-level aggregate indicators needed to assess the study are provided in the accompanying PDF data appendix, *Balanced Five-City Municipal Publication Audit: PDF Data Appendix*. The appendix contains aggregate document indicators only. Raw PDFs, raw extracted text, names, and person-level records are not included or stored as publication materials.

The computational helpers used for validation and aggregation are not required to interpret the PDF publication package. The methods section describes the corpus-construction logic, indicator definitions, and ethical safeguards needed to evaluate the study.

Ethics Statement

This study did not involve human subjects research in the sense of interaction, intervention, recruitment, or person-level data collection. It examined public institutional documents at document level. The protocol was deliberately designed to avoid creating lists of names or profiles. Documents were analyzed only as publication artifacts.

Author Note

This pilot study was developed by Benjamin Metzиг for Wissenschaftswelle.de as part of the *Private Palantir* research report. AI tools were used to support structuring, language refinement, code drafting, and consistency checks. Editorial responsibility, final judgement, source selection, interpretation, and any publication decision remain with Benjamin Metzиг.

References

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-](https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf)

[10/edpb_guidelines_202401_legitimateinterest_en.pdf](https://www.edpb.europa.eu/system/files/2024-10/edpb_guidelines_202401_legitimateinterest_en.pdf)

German Data Protection Conference. (2024). Orientierungshilfe KI und Datenschutz.

<https://www.datenschutzkonferenz->

[online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf](https://www.datenschutzkonferenz-online.de/media/oh/20240506_DSK_Orientierungshilfe_KI_und_Datenschutz.pdf)