

CLARA: Communication of Local German Party Associations, Recorded Annually (2015–2025)

Data report for the public v1.0.0 release

Hanno Hilbig

2026-08-07

Table of contents

Summary	1
Methods	2
Sampling frame and website discovery	2
Archival reconstruction	3
Content extraction and classification	3
From records to panel measures	3
Data Records	4
Technical Validation	4
Human validation of the extracted measures	4
Independent model-judged validation of the released records	4
Issue attention tracks the national agenda	6
Campaign content tracks the electoral calendar	6
Composition and platform gradients	7
Migration attention and the local far right	8
Coverage by variable family	10
Coverage and selection into observability	11
Usage Notes	12
Code Availability	13
Citation	13
References	13

Summary

CLARA (Communication of Local German Party Associations, Recorded Annually) is a chapter-year panel that measures the observable web presence of 2,635 local chapters of the seven main German parties (CDU, CSU, SPD, Grüne, Linke, AfD, FDP), observed annually from 2015 to 2025 (28,985 chapter-years). The unit is the county-level party association. Of the 2,635 chapters, 2,591 are Kreisverbände, meaning one association per party per county; within the recommended analysis scope, 2,316 of 2,319 chapters are Kreisverbände. The remainder are city- or town-level chapters, identified by `org_level` in the roster. Drawing on roughly 1.03 million archived pages from the Internet Archive’s Wayback Machine, I reconstruct each chapter’s website in each year and extract structured measures: leadership boards and their composition, issue emphasis, the social-media

accounts and mobilization features a site links to, and a classified news file of 237,399 archived articles labeled by policy topic, article function, and geographic focus. All extraction and classification behind the released measures ran on a locally hosted open-weight model. A human coder validated every model-generated measure across five stratified samples (336 records) and, in a sample taken from the released records themselves, confirmed 45 of 48 extractions. The public release contains chapter-year aggregates and coverage metadata only: no personal data and no reproduced website text. To my knowledge, no other dataset follows the communication of local party organizations across all seven parties, every county, and a full decade. CLARA measures observable local party web presence rather than total local party activity.

Methods

The unit of observation is the chapter-year: for each chapter and each year from 2015 to 2025, I select archived snapshots of the chapter’s known URL and extract measures from the pages captured in that year. Whether a chapter-year is observable depends on website existence and archival coverage, so every row also carries explicit coverage metadata. I return to the resulting selection issues in the validation and usage sections.

Sampling frame and website discovery

I build the frame of possible chapters before any website discovery. I restrict the Destatis Gemeindeverzeichnis (the official municipality register) to municipalities with at least 5,000 residents, which yields 2,943 municipalities aggregating to 399 Kreise (counties). Crossing these Kreise with the parties’ organizational scope rules (AfD, FDP, Grüne, Linke, and SPD organize everywhere; the CDU outside Bavaria; the CSU only in Bavaria) produces 2,394 party-by-Kreis cells, one cell per party per county. Each cell is a possible chapter, and I compute all coverage statistics in the release against this frame. I also place each discovered chapter in a single municipality, the town it is based in, which I record in the 8-digit official municipality key `ags`; the 5-digit county key `kreis_id` is its first five digits. The two keys answer different questions, and the codebook says which to merge on. A Kreisverband typically contains subordinate town-level chapters (Ortsverbände, Stadtverbände); these are not separate observations, and I attribute pages that a subordinate chapter hosts under the county association’s domain by explicit rules rather than pooling them.

I then locate chapter websites per cell through five channels, in decreasing order of authority: (i) the parties’ own chapter directories (999 chapters); (ii) the SPD’s published organizational roster (8); (iii) systematic probing of predictable domain patterns such as `cdu-kreisname.de` (885); (iv) search-engine queries with model-classified results (740), where queries contained only party and place names; and (v) archived versions of the party directories for chapters whose sites had already disappeared (3). I record the channel per chapter in the `url_source` variable, so users can condition on discovery quality. Discovery found a site for 98.8 percent of cells (2,365 of 2,394); I found no website for 27 cells (14 Linke, 6 FDP, 3 AfD, 2 SPD, 1 CDU, 1 Grüne), and 2 further chapters are known but have no usable URL. I emphasize one caveat: discovery ran against the live web in early 2026, so chapters whose sites went offline earlier without replacement and without an archived directory trace are missing entirely. Early panel years therefore carry survivorship selection on top of archival selection, a limitation I disclose but have not yet quantified.

Archival reconstruction

Drawing on the Wayback Machine’s CDX index (its index of archived URLs), I query each chapter’s known URL for every year, decide which snapshots to fetch, and download the archived HTML, roughly 1.03 million pages in total. When a chapter-year offers more snapshots than I fetch, I take pages carrying organizational structure first (the homepage, the chapter’s self-description page, contact and imprint pages, board and council-group pages), ahead of issue and news pages. I classify pages by function (homepage, officials, news, issues, contact, and so on) and record the snapshots selected for each chapter-year in `chapter_year_provenance.csv`, so that any aggregate can be audited. Whether a chapter-year is observable at all thus depends on two processes I do not control: the chapter maintaining a site, and the Internet Archive having captured it that year. Every chapter-year therefore carries a status (`observability_status`) recording how far it got along the chain from having a known URL to yielding extractable content, together with three coverage columns: the evidence on which it counts as covered (`coverage_basis`), the specific reason it does or does not (`release_coverage_bucket`), and a coarser summary grade (`release_coverage_tier`).

Content extraction and classification

All content extraction and classification behind the released measures ran on a locally hosted open-weight model with schema-constrained decoding, which forces outputs into the target record structure. The exact checkpoint, for replication, is `Qwen/Qwen3-30B-A3B-Instruct-2507-FP8`, served with vLLM at temperature 0. From the archived pages I extract officials with roles, issue content blocks, and organizational contact information. I deduplicate news pages to one record per unique article. Archived pages also preserve older material, so the deduplicated corpus holds 288,728 articles in total, of which 237,399 carry a date inside the 2015 to 2025 panel window and enter the released counts; the remainder are articles dated before 2015. I classify each article along three dimensions: a policy topic family, an article function (press release, event announcement, position statement, election campaigning, or council work), and a geographic focus (local, regional, or national).

The page-function labels that route pages to extractors combine URL rules with the same local model, benchmarked against a 240-page human-adjudicated gold standard (71.7 percent accuracy; the missingness audit addendum describes the benchmark). The independent validation of the released records likewise used a second locally hosted open-weight model from a different family (Mistral-Small-24B; see Technical Validation). No step behind the released measures involves a commercial model.

Three derived classifications complete the record. I infer gender from first names, a standard method benchmarked at roughly 94 percent accuracy (Santamaría and Mihaljević 2018); the inference is binary, I retain and flag unknown names, and `gender_coverage` reports the classified share per chapter-year. A keyword dictionary assigns issue topics to about 25 families, with the language model deciding the cases the dictionary does not cover, and each family maps to a Comparative Agendas Project (CAP) major topic. I normalize the role titles extracted from the pages into a fixed set of board-role categories. These classifiers ran locally as well.

From records to panel measures

Fixed rules, not a model, aggregate the extracted records into chapter-year measures, and no free text survives into the release. Every substantive feature carries a companion status column that distinguishes four cases: an observed positive value, an observed zero (I read a page of the right type

and found none of the target content), coverage that exists but from which the feature was never extracted, and no coverage at all. An apparent zero is therefore never conflated with a missing site or a missing archive.

Data Records

The release contains nine data files, and identifiers are stable across them. The row unit is the chapter for the roster, the chapter-year for both panels and the provenance file, the chapter-year-family for the two long-format topic files, and the county-election-district pair for the Wahlkreis crosswalk. `README.md` lists the files with their row counts, and `CODEBOOK.md` defines every variable.

The base panel carries identity and geography, the coverage block, and twelve substantive features, each with its status column. The enriched panel adds derived measures: board composition and gender, chair and chapter-head fields, professionalization proxies, web-activity counts, mobilization link flags following the Gibson-Ward functions, and issue-topic presence flags in wide format. The news file reports article volume with label shares by geographic focus and article function, and the two long-format files carry per-family counts for issue blocks and news articles, each joined to its CAP major topic. Full variable definitions are in `CODEBOOK.md`.

Topic family and geographic focus are assigned to the same article, so `chapter_year_news_topic_geo.csv` holds their joint distribution rather than two separate marginals. This file allows topic and geographic focus to be analyzed jointly. Figure 1 ranks topics by the gap between their most and least local party. Community news runs between 96 and 98 percent locally focused across all seven parties, and urban development between 93 and 97. Economy news is 30 percent local for the AfD and 78 percent for the CSU, and agriculture, environment and health show gaps above 39 points. Foreign policy is the one topic no party treats as mostly local, which is consistent with the intended interpretation of the geographic-focus classifier. The AfD has the lowest local-focus share in 17 of the 22 topics; excluding the AfD, the median topic still spans 26 points.

Technical Validation

Human validation of the extracted measures

A human coder validated every model-generated measure by reading extracted values next to source-page excerpts. I drew four stratified samples as the measures were built: (i) a spot check of 48 extracted records from the officials, issues, and about/contact extraction paths; (ii) 63 news articles checked field by field; (iii) 100 issue topic-family assignments; and (iv) 77 news articles with all three content labels. Across all 288 reviewed records the coder agreed with the extraction. The coder then reviewed a fifth stratified sample of 48 records taken from the released records themselves, each against a page excerpt anchored on the extracted value: the coder confirmed 45 extractions, marked 1 incorrect (a role string containing a qualifier not present on the page), and could not decide 2 from the excerpt, an agreement of 45 of 46 decidable records. The public release reports aggregate agreement counts in `coverage_and_quality/validation_human_agreement.md`; the underlying sheets contain source excerpts and remain private.

Independent model-judged validation of the released records

I drew a stratified sample of 315 released records (105 each from the officials, issues, and about/contact paths, balanced across the seven parties) and had every field judged against its source page by an

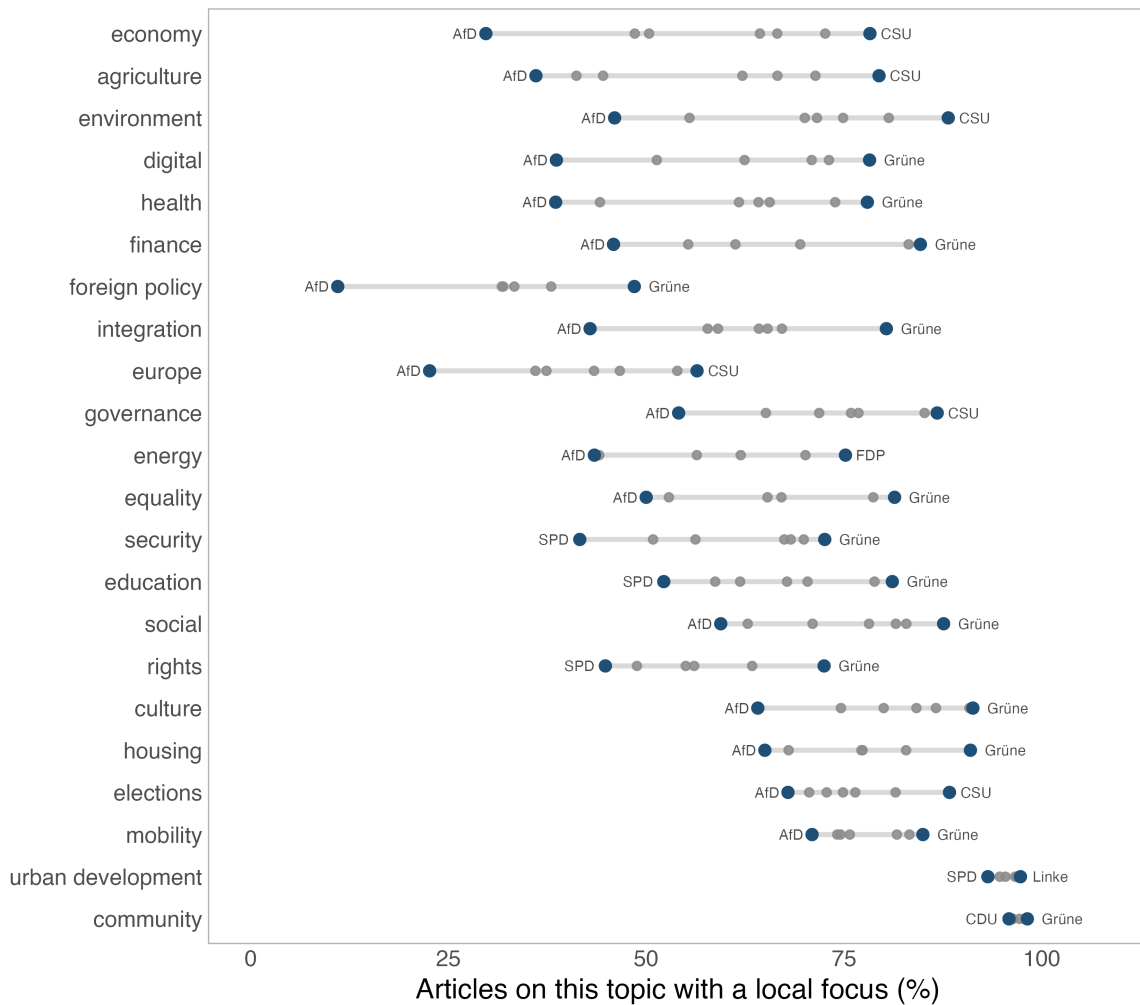


Figure 1: Locally focused share of news articles by topic and party. Each row is one topic family, ordered by the gap between its most and least local party; grey points are the remaining parties. A party needs 100 articles on a topic to appear, and a topic needs five such parties.

independent locally hosted open-weight model from a different family than the extractor (Mistral-Small-24B-Instruct). Of 769 judged non-empty fields, the judge rated 76.9 percent supported by the page and 2.9 percent contradicted. The remainder could not be found in the truncated, navigation-stripped page excerpt the judge was shown, and count as neither confirmed nor refuted rather than as wrong. I then read all 22 contradicted fields myself: in half of them the extracted value appears verbatim in the full page beyond the judge’s excerpt, so the true contradiction rate is lower than the judged one. Full tables are in `coverage_and_quality/validation_report.md`.

Issue attention tracks the national agenda

Figure 2 compares annual topic shares with four major national events. It plots the share of news topics devoted to four issue families. Attention to integration and migration peaks at 9.5 percent of topics in 2015, the year of the refugee crisis, and falls sharply thereafter. Environment and climate attention builds through the climate-protest wave and peaks in 2019, the year of the “climate election” to the European Parliament. Health rises from around 2 percent to over 8 percent in the pandemic years 2020 and 2021, and energy peaks in 2022, the year of the gas crisis. None of these patterns is available to the classifier, which sees article text without dates.

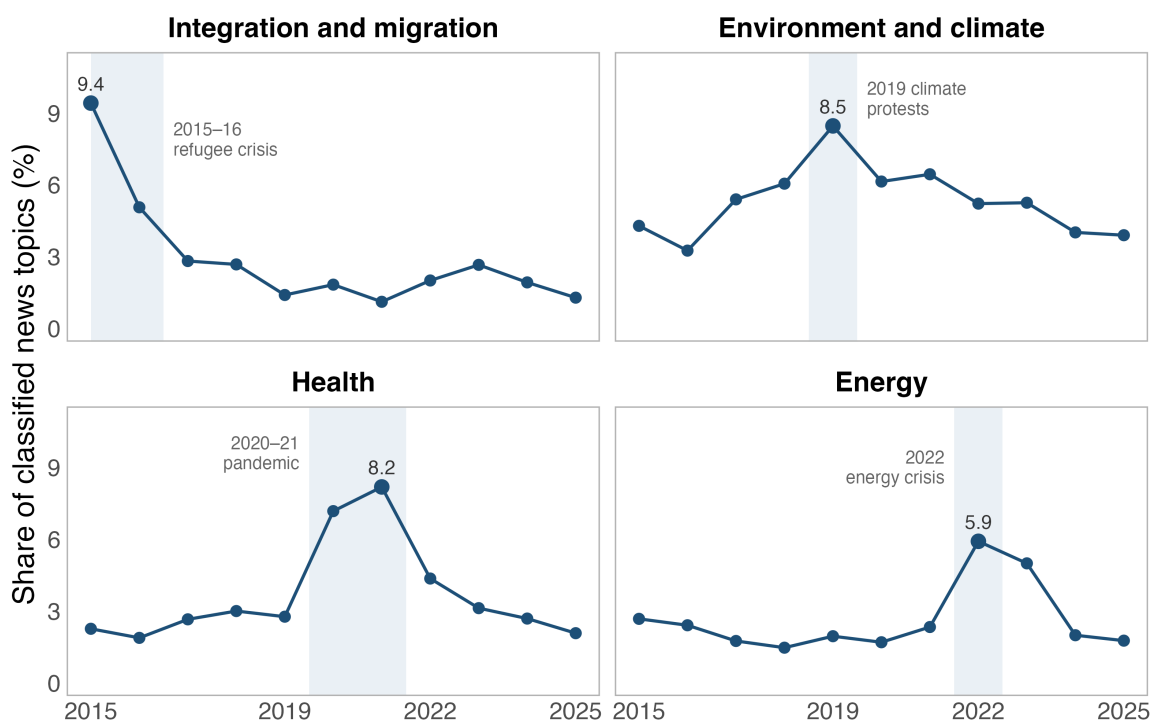


Figure 2: Issue attention in local party news tracks national agenda shocks. Each panel shows one topic family’s share of all classified news topics by year; shaded bands mark the relevant national event.

Campaign content tracks the electoral calendar

Figure 3 shows the share of articles classified as election campaigning: it peaks in federal and European election years and falls in off-years, ranging from 4.8 percent in 2015 to 13.4 percent in 2024.

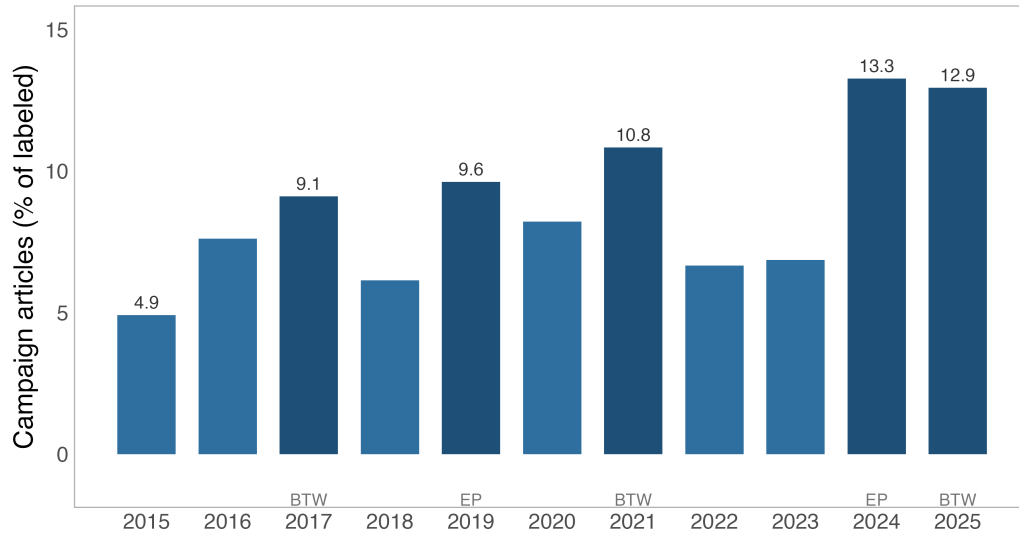


Figure 3: Campaign-content share of archived local party news articles by year. BTW marks federal elections, EP European Parliament elections.

Composition and platform gradients

Mean board female shares differ substantially across parties. Figure 4 orders parties by that share: the Grünen, with a parity quota, lead at 48 percent; Linke and SPD, with formal quotas, follow at 38 and 36 percent; the CDU and CSU, with informal quorums, sit at 27 and 26 percent; and the FDP and AfD, without quotas, trail at 22 and 16 percent. The ordering matches the parties' candidate-quota rules exactly, although I stress that the shares are name-inferred and website-listed rather than administrative counts.

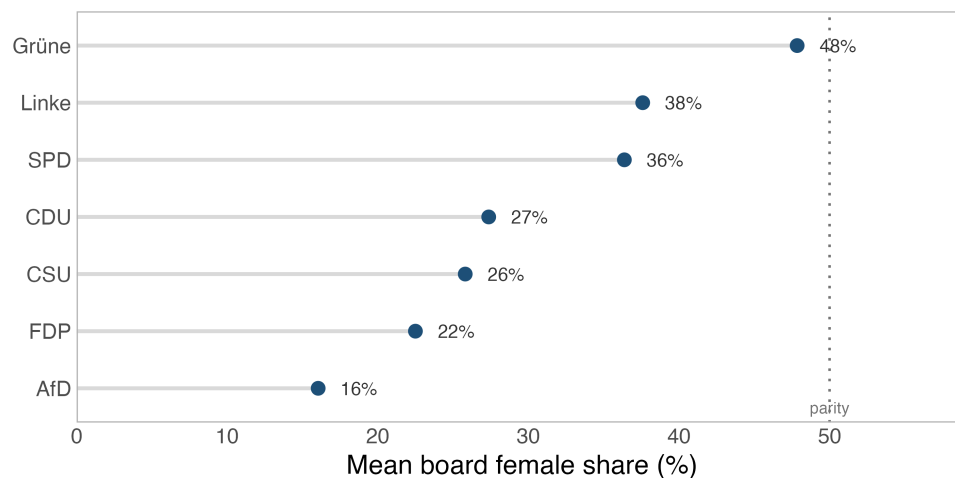


Figure 4: Mean board female share by party, chapter-years with at least 70 percent of board names classified. The dotted line marks parity.

Telegram use is concentrated among AfD chapters. Telegram presence is essentially an AfD phenomenon: it appears after 2018, accelerates in 2021 when mainstream platforms tightened

moderation, and reaches 15 percent of AfD chapters by 2025, while remaining near zero for all other parties (Figure 5). Party identity is not an input to these extraction or classification steps, which are identical for every party.

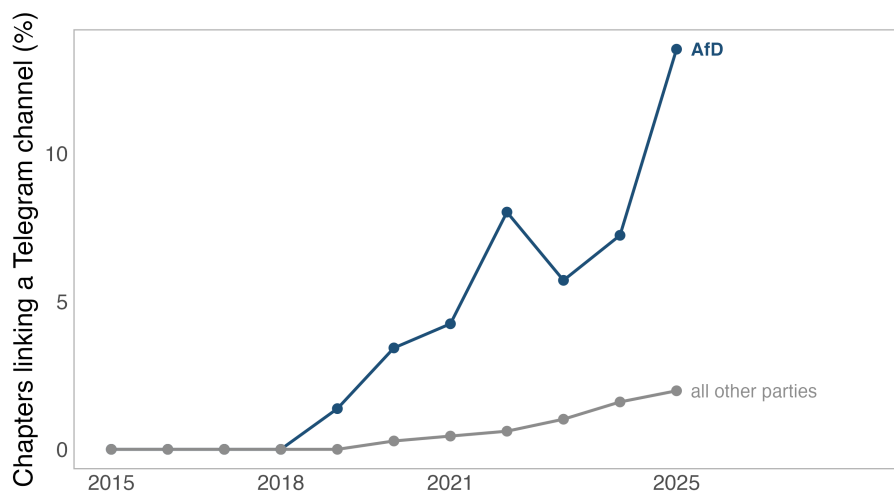


Figure 5: Share of chapters linking a Telegram channel, AfD versus all other parties.

Migration attention and the local far right

Figure 6 to Figure 8 describe migration attention and geographic focus by party. The accommodation literature asks whether mainstream parties win voters back by adopting radical-right issue positions, and most evidence suggests they do not (Krause, Cohen and Abou-Chadi 2023). Figure 6 plots the migration share of each party’s news topics: the AfD leads every mainstream party in every year, mainstream attention falls after 2016 rather than rising to meet the challenger, and only the AfD returns to a high share during the 2023 migration debate. On their websites, mainstream chapters did not increase migration attention as AfD attention increased. This pattern provides little evidence of accommodation from below.

The second pattern speaks to the debate on far-right normalization and local rootedness (Schroeder, Ziblatt and Bochart 2025). Figure 7 compares the geographic focus of AfD news with that of all other parties. The AfD is by a wide margin the most nationally oriented party on the ground: 39 percent of its article content was nationally focused in 2015, against 6 to 18 percent for the others. While that gap persists, the AfD’s local-focus share climbs from 48 percent in 2015 to 73 percent in 2025, with visible interruptions in the federal election years 2021 and 2024.

Figure 8 shows the same pattern across all seven parties: the locally focused share of a party’s articles rises from 53 percent for the AfD through 64 percent for the SPD and 74 percent for the CDU to 83 percent for the Grünen and the CSU, and the national share mirrors it. Party identity was not used in classification, so this gradient was not mechanically imposed by the pipeline.

A further check uses the fact that each article carries two labels assigned in separate passes from different label sets: a geographic focus and an article function. Neither classifier saw the other’s output. If both track how local a chapter’s output is, chapter-years with more locally focused articles should also carry more council-work articles, and Figure 9 shows that they do in every party (Spearman correlations of 0.27 to 0.35, except the CSU at 0.11 on 267 chapter-years). Agreement

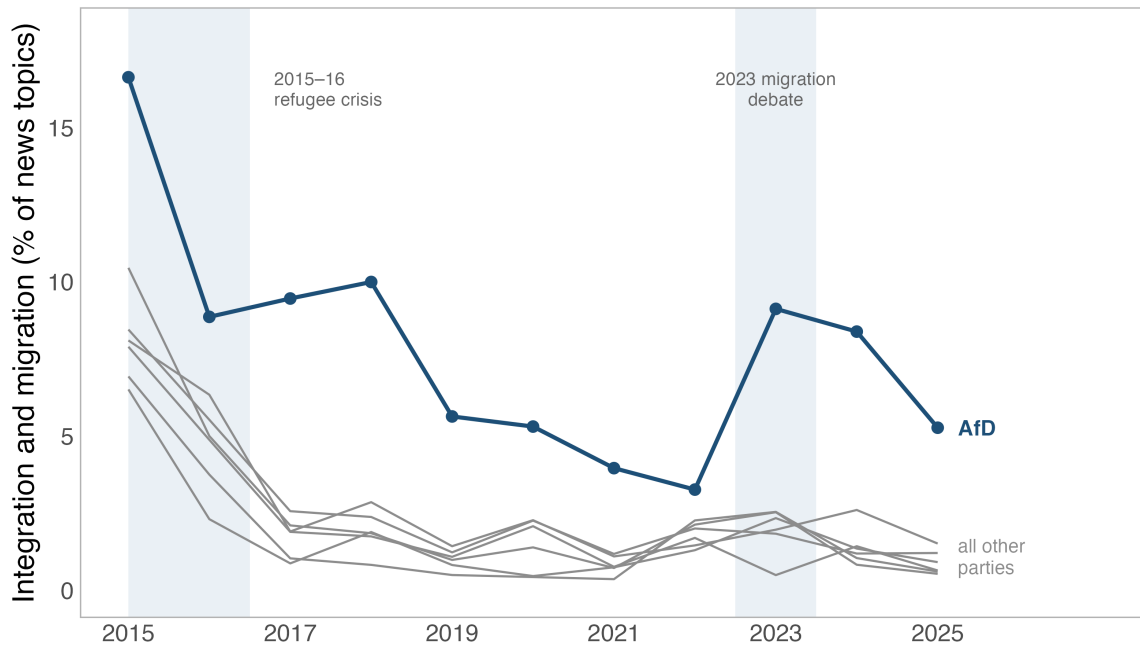


Figure 6: Attention to integration and migration by party, as a share of each party’s classified news topics.

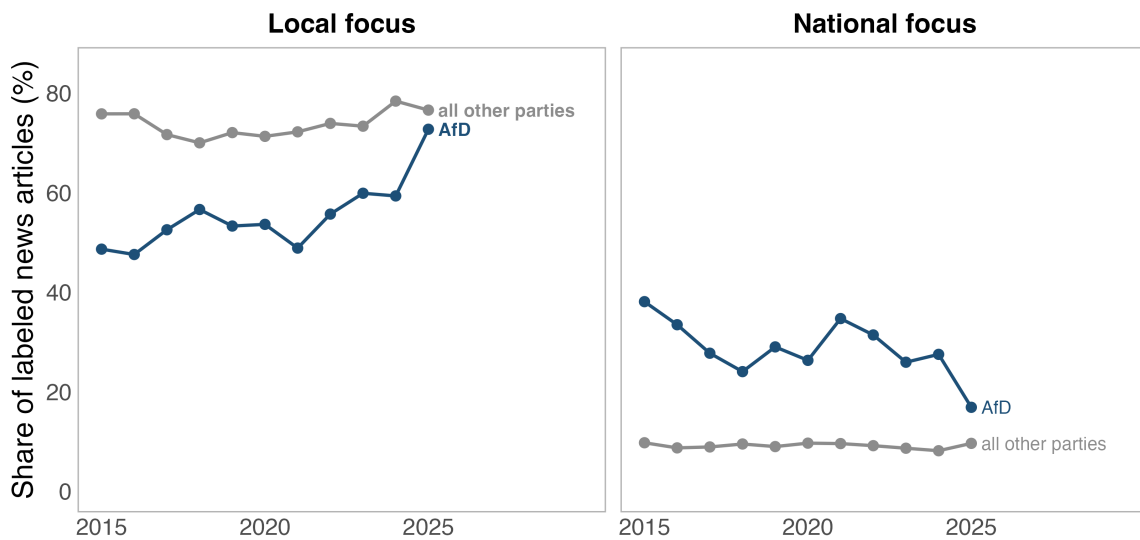


Figure 7: Geographic focus of news articles, AfD versus all other parties.

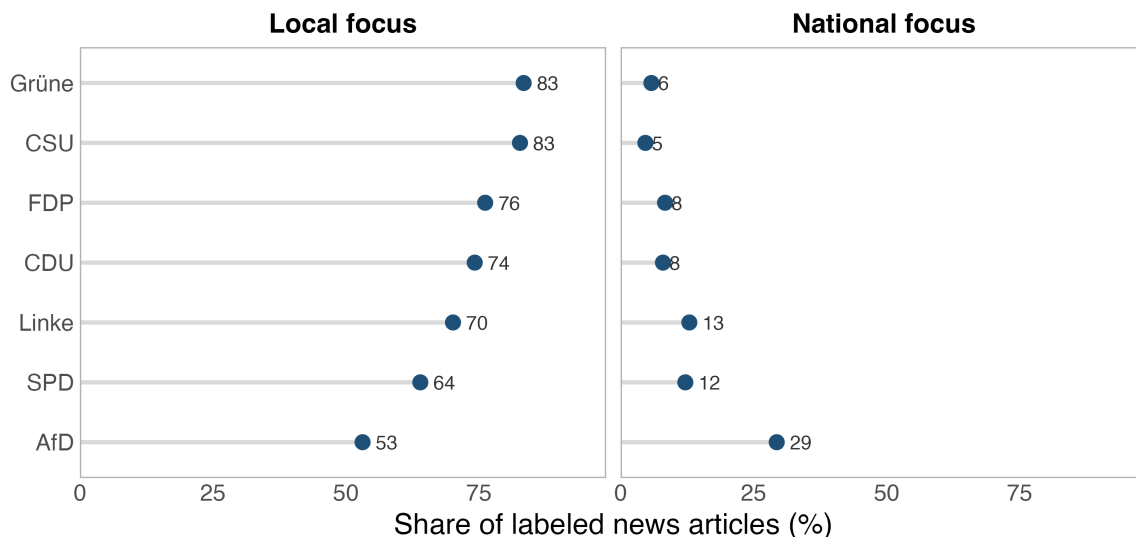


Figure 8: Geographic focus of news articles by party, weighted by article count. Chapter-years in the analysis scope with at least five labeled articles.

between two classifiers on an axis neither was given is stronger evidence that the labels track something real than a precision figure computed against the same pages the model read.

Coverage by variable family

A variable family is a group of measures fed by the same kind of source page: leadership boards come from board pages, contact information from contact and imprint pages, issue positions from issue pages, and so on. Coverage differs sharply across these families, because a chapter’s archived site may contain one kind of page and not another. Table 1 therefore reports two denominators: all 25,509 chapter-years in the analysis scope (defined in the next subsection), and only those chapter-years for which the archive holds a page of the type that family needs. Measured zeros count as observed in both columns.

Table 1: Variable-family coverage. Observed values include measured positives and measured zeros. The source-conditional denominator contains chapter-years with a successfully archived page eligible for that family.

Variable family	Observed in analysis scope	Share of analysis scope	Observed when required source is available	Conditional share
Leadership board	13,064 / 25,509	51.2%	13,064 / 13,075	99.9%
Contact information	14,064 / 25,509	55.1%	14,064 / 14,064	100.0%
Issue positions	6,865 / 25,509	26.9%	6,865 / 6,886	99.7%
News	15,481 / 25,509	60.7%	15,481 / 16,699	92.7%
Social media	18,961 / 25,509	74.3%	18,961 / 18,961	100.0%

Conditional on an eligible source page, extraction coverage is complete or nearly complete in every family; the one residual of any size is news, where some archived news pages yield no parseable

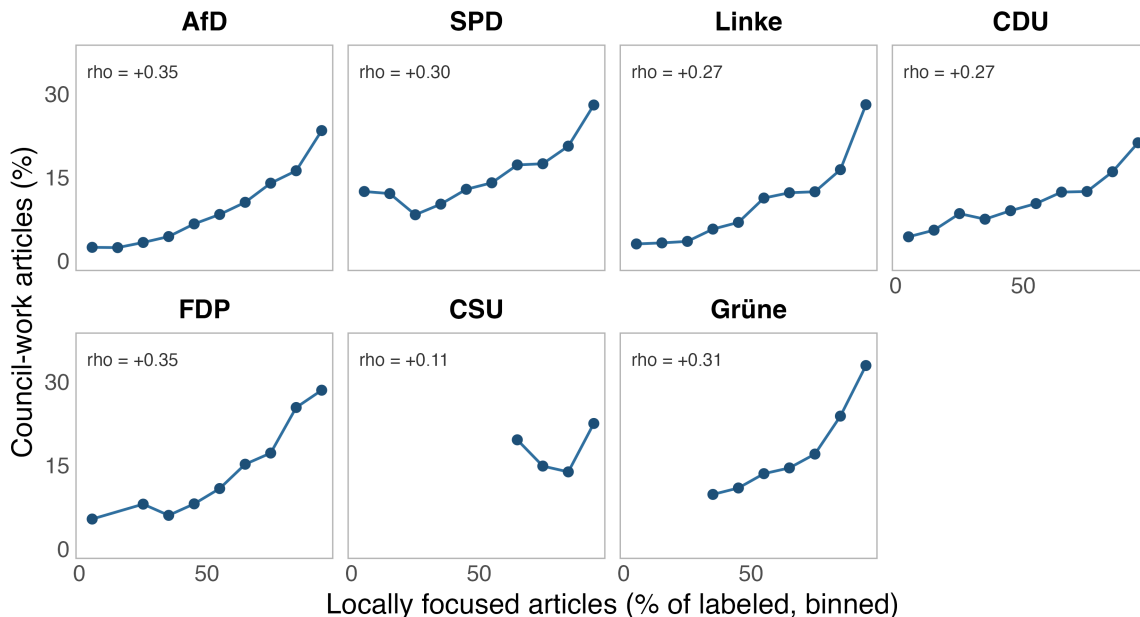


Figure 9: Council-work share against locally focused share, binned into deciles of the horizontal axis. Binned rather than plotted raw because most chapter-years sit on a boundary and small article counts make the underlying shares discrete.

article. Coverage looks low against the full analysis scope because the archived source page is absent, not because extraction failed on it. Users should therefore choose the denominator that matches their research question and retain the family-specific status columns in all analyses. A cause-level decomposition of residual missingness, including a statistical bound on recoverable content the release does not capture, accompanies the deposit (`coverage_and_quality/missingness_audit_addendum.md`).

Coverage and selection into observability

Of 28,985 chapter-years, 25,509 are in the analysis scope. A chapter-year belongs to that scope when its chapter falls inside the 2,394-cell party-by-county sampling frame and survives the roster quality screening, which removes chapters identified as the wrong organizational unit or as junk. The release denominator is researcher-usable coverage: 17,989 chapter-years (70.5 percent of scope), of which 15,118 meet the full structured-coverage threshold (`coverage_basis = full`) and 2,871 fall short of that threshold but still yield an observed feature value, either from pages below the threshold or from pages that extraction reached through a fallback rule (`coverage_basis = feature_level`). For most of the remainder, no further collection effort would help: 6,335 chapter-years (24.8 percent) have no same-year archived snapshot, meaning the Internet Archive holds no capture of the site for that year.

Coverage is strongly non-random, and Figure 10 shows where. Full-threshold coverage rises from about 35 percent in 2015 to 73 percent in 2025, and party levels range from 74.7 percent (SPD) to 37.2 percent (AfD). The CSU panel is the most volatile: many CSU chapters sit on shared state-party infrastructure whose archival capture varies year to year. A descriptive linear-probability model in `missingness_diagnostics.csv`, which regresses a chapter-year's zero-one usable-coverage indicator on chapter and year characteristics, shows usable coverage rising with municipality population and

over time. Missing chapter-years must not be treated as zeros, and analyses should condition on `observability_status` or `coverage_basis`.

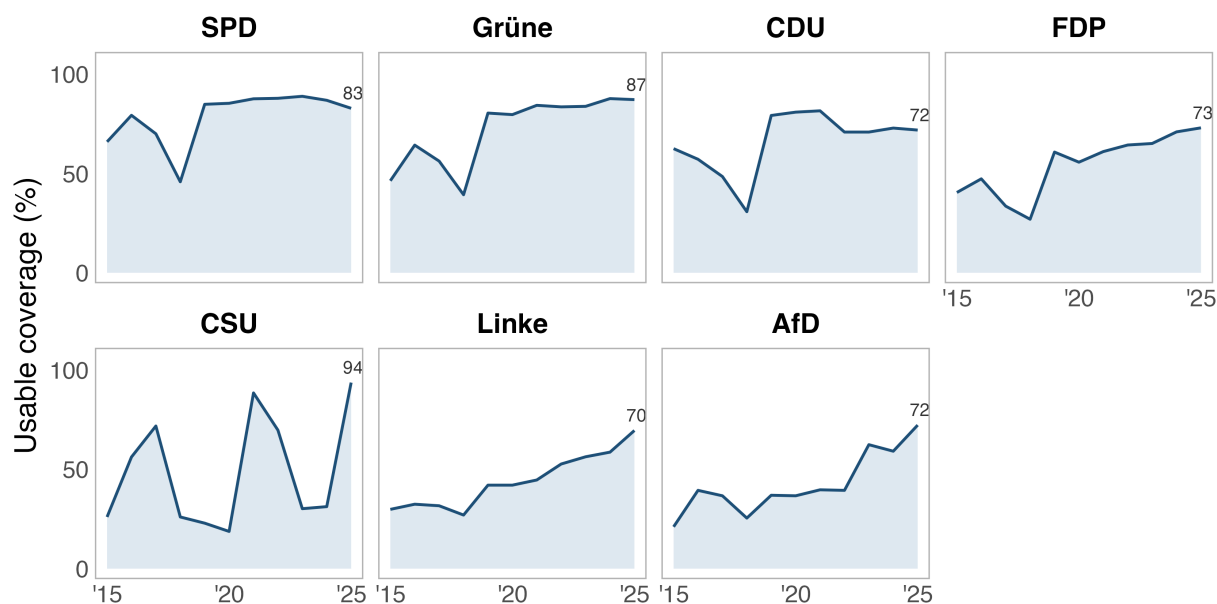


Figure 10: Usable structured coverage by party and year. Each panel shows the share of analysis-scope chapter-years meeting the full coverage threshold.

Usage Notes

Three rules are important when using the release. First, read `ags` and `kreis_id` as strings. Both are zero-padded, and a loader that guesses their type will read them as integers and silently drop the leading zero, which affects roughly 80 percent of chapters. Second, filter to `analysis_scope_included = 1`, and then either to `coverage_basis = 'full'` for a conservative sample or to any non-blank `coverage_basis` for the full researcher-usable denominator, using the two samples as a robustness split. Third, treat a feature value as informative only when its status column reports an observed positive or an observed zero.

The measurement caveats that follow from the construction are listed in `README.md` under known limitations: what the extracted boards pool together and why `officials_source` conditions their interpretation, why issue fields measure salience rather than position, what a link or social-media flag does and does not imply, and why the news count tracks archival density as well as posting volume.

Annual within-chapter changes require particular caution. A single annual snapshot may fall before or after a local election or a site redesign, so year-to-year change is noisy, and part of that noise is systematic: the count measures move with how much of a site the archive captured that year. Adding a control for `n_pages_total` to a chapter fixed-effects regression cuts the annual trend in `issue_block_count` by about three quarters and reverses the sign of the trend in `official_count`. Because page counts partly reflect real activity, that control can absorb signal along with noise, so report estimates both ways and read the difference as a bound. The codebook reports per-variable reliability diagnostics. Finally, I match each chapter to a single seat municipality even though a

Kreisverband organizes an entire county, so `ags` and the `population` column describe that one town rather than the area the chapter covers; merge county-level covariates on `kreis_id`.

The public release contains chapter-year aggregates and coverage metadata only. `DATA_ACCESS.md` states what I withhold and why, and `LICENSE-CC-BY-4.0.md` carries the license terms.

Code Availability

The full extraction and panel-construction pipeline is available from the author on request. `release/figures_report.R` reproduces the release figures from the public CSVs alone, and a researcher with the code and the raw corpus can regenerate the panel and all reports end to end.

Citation

Hilbig, Hanno (2026). *CLARA: Communication of Local German Party Associations, Recorded Annually (2015–2025)*, Version 1.0.0. Harvard Dataverse. <https://doi.org/10.7910/DVN/2KJSCO>

References

- Gibson, R. and Ward, S. (2000). A proposed methodology for studying the function and effectiveness of party and candidate web sites. *Social Science Computer Review* 18(3), 301–319.
- Krause, W., Cohen, D. and Abou-Chadi, T. (2023). Does accommodation work? Mainstream party strategies and the success of radical right parties. *Political Science Research and Methods* 11(1), 172–179.
- Schroeder, V., Ziblatt, D. and Bochert, F. (2025). Is the Brandmauer breaking from below? Party cooperation with the AfD in German local councils. WZB Discussion Paper.
- Comparative Agendas Project. Master codebook (21 major topics).
- Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science* 4:e156.