Speaker Mining - FAIR Data on Public Broadcasts for Question Answering

Tim Wittenborg[1][0009−0000−9933−8922], Omar Imad Remmo, Claudia Frick[3][0000−0002−5291−4301], Lena John[2][0009−0007−2097−9761], Oliver Karras[2][0000−0001−5336−6899], and Sören Auer[1,2][0000−0002−0698−2864]

[1] L3S Research Center, Leibniz University Hannover, Hannover, Germany
tim.wittenborg@l3s.uni-hannover.de
[2] TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
{lena.john, oliver.karras, soeren.auer}@tib.eu
[3] Institute for Information Science, TH Köln - University of Applied Sciences Cologne, Cologne, Germany
claudia.frick@th-koeln.de

Abstract. Public broadcasts are at the center of civic discourse: Traditional television talk shows, alongside emerging podcast and web video formats, capture and guide the attention of our societies, shaping how citizens encounter politics, science, and societal issues. Yet, systematic or even simple analyses of these formats face similar challenges: guest and content metadata are scarce, fleeting, fragmented, and not standardized. Research conducted and questions answered are based on extensive, laborious, yet isolated data-curation efforts that capture only a fraction of the relevant landscape. This work seeks to address this issue using a scaling-oriented framework for FAIR data curation in public broadcasting. Evaluated on 15 broadcasting programs, the pipeline aggregates ZDF Archive PDFs, fernsehserien.de, and Wikidata into a unified knowledge graph. Of the 31,817 candidate guest mentions from these three sources, 17,729 could be automatically disambiguated, further 5,958 via 64 hours of manual reconciling using OpenRefine. Results are published at speakermining.wikibase.cloud and linked to Wikidata, enabling SPARQL-based question answering based on gender, age, occupation, or institutional affiliation across 8,436 canonical persons with 23,527 appearances in 6,469 aligned episodes. Our iterative experience reveals that correctly disambiguating and deduplicating speaker data from heterogeneous sources demands dedicated effort on sustainable infrastructure. For scalable and reliable question answering on public broadcasts to be accessible to everyone, we recommend fostering the potential of linked open data: Advancing alignment and utilization approaches like this work, particularly towards crowdsourced development and curation, but also more FAIR data interfaces from public broadcast service providers.

Keywords: Science Communication · FAIR Data · Knowledge Graph · Entity Disambiguation · Public Broadcasting

1 Introduction

Public broadcast talk shows and interview formats shape public discourse. Programs such as political talk shows and science podcasts are major drivers of how citizens encounter politics, science, and societal issues [12]. Millions of viewers encounter experts and scientific perspectives through these formats rather than through direct engagement with scientific publications [15]. This underlines the democratic relevance of public broadcasts and science journalism as a specific form of science communication [7].

Systematic analysis of these broadcast formats remains difficult because relevant metadata is fragmented, inconsistent, and often inaccessible. But these formats are especially important because editorial selection determines which experts, topics, and perspectives become publicly visible, making them prone to biases and stereotypical representations of science. Journalistic gatekeeping can privilege certain demographics, institutions, and viewpoints while limiting the visibility of others [14,15]. To investigate these effects, which might stem from any of the identity types evident in the academic wheel of privilege [8], studying science communication across journalistic formats is necessary. Broadcast metadata is rarely FAIR (Findable, Accessible, Interoperable, Reusable). Researchers, therefore, spend substantial effort manually collecting and reconciling guest information before analysis can begin. Structured FAIR broadcast data would also enable question answering (QA) about guest variability, age distribution, or topic visibility at scale [18,19].

This work presents a framework for FAIR curation of public broadcast guest data and accounts for where true scalability remains unsolved. Using 15 broadcast programs as a case study, we aggregate and reconcile heterogeneous metadata from ZDF Archive PDFs, fernsehserien.de, and Wikidata into a shared Wikibase infrastructure. Our contributions include: (1) An open source pipeline[4] for event-sourced candidate generation, alignment across sources, and staged entity disambiguation and deduplication. (2) A curated FAIR dataset[5] covering 6,469 episodes across 12 broadcasting programs (only 10 of which have a total of 4,861 episodes with guest data), disambiguated using OpenRefine. (3) Quantitative analyses across the 4,861 episodes for 15 selected guest properties, particularly gender, age, occupation, position, and party affiliation.

2 Background and Related Work

Science communication is increasingly studied as a socio-technical system shaped by diverse actors, media formats, and institutional constraints. Prior work highlights that different communication channels, including journalism, social media, visual communication, participatory formats, and humour, come

[4] https://github.com/borgnetzwerk/speaker-mining
[5] https://speakermining.wikibase.cloud/

with distinct affordances and challenges for communicating scientific knowledge [9,12,16]. Within this landscape, journalism occupies a particularly influential role because it mediates between scientific communities and broader publics, thereby shaping which topics, experts, and perspectives receive visibility [4,17]. Related work documents persistent inequalities: women are underrepresented as experts in journalism, and media structures can reproduce existing systemic biases, including gender and ethnic disparities in quotation practices and media visibility [1,6,7,11]. However, many such studies rely on purpose-built datasets that remain difficult to reproduce, extend, or reuse across formats and time periods. Public broadcast talk shows occupy a distinctive position within this landscape: they operate under editorial curation, reach millions of viewers per episode, and create a recurring, timestamped record of which voices receive public visibility over years and decades. Yet systematic longitudinal analysis of these records has been hampered by the absence of structured, reusable metadata.

FAIR data practices, particularly Knowledge Graphs (KGs), are promising approaches for addressing these challenges. Public broadcaster archives, web aggregators such as fernsehserien.de, and open knowledge bases such as Wikidata offer complementary but heterogeneous metadata sources. Our prior work on the emerging Science Communication Knowledge Infrastructure (SciCom KI) argued that existing curation infrastructures remain fragmented and insufficiently scalable for audiovisual formats [19]. The rufus [3] portal addresses similar challenges, supported by the ZDF archive providing data on almost 500,000 episodes.

Related analyses demonstrate the societal relevance and analytical potential of structured broadcast metadata. Der Spiegel [5] repeatedly investigated recurring guests and representational imbalances in German talk shows. Independent projects such as LanzMining [2] explored automated aggregation of talk show guests using large language models and simple regex-based classification. In collaboration with LanzMining, a community-driven approach coordinated by Stefan "stk" Kaufmann [10] demonstrated the use of Wikidata and linked open data for analyzing guest appearances and party representation in political talk shows. Table 1 contrasts our framework with the prior analyses whose datasets permit cross-validation. Only one other approach provides Wikidata-linked, SPARQL-queryable output. To our knowledge, no other contained an explicit entity deduplication protocol. Our framework extends coverage to 15 programs (12 with episode data, 10 of those with guest data) and links 62.3% of unique persons to Wikidata. Our work builds on this prior work towards a tool-supported, rule-based disambiguation workflow for FAIR broadcast data.

Table 1: Comparison of related talk show guest analysis approaches.

	Period	**Shows**	**Episodes**	**Guests**	**Wikidata**	**Output**
Spiegel [5]	2015–25	6	2854	3,991	—	CSV
LanzMining [2]	2022–25	5	515	813	—	CSV
stk [10]	2024	1	138	262	100%	Wikidata
Our work	1972–26	15	6,469	8,287	62.3%	Wikibase
↳ with guest data	2006–26	10	4,861	8,287	62.3%	Wikibase

3 Approach

We pursued a digital library approach to FAIR broadcast data in two iterations: (1) an explorative single-source run to identify roadblocks, and (2) a full redesign with multiple sources and a contract-coupled, event-sourced pipeline.

3.1 First Iteration: Explorative Discovery

Detailed in a bachelor's thesis [13], four PDFs from the ZDF archive were processed in a four-stage pipeline (PDF extraction, Wikibase setup, OpenRefine entity linking, SPARQL analysis). It extracted 2,035 of 2,036 episodes, created 3,966 person items, and linked 2,984 (75.3%) to Wikidata through eight hours of manual review. The 36.0% female appearance share (31.3% of unique guests) matched Spiegel's independent figure of 35%, validating the approach. The iteration equally revealed systematic roadblocks that motivated the redesign:

1. **Brittle extraction.** The regex parser was tightly coupled to the ZDF archive typographic conventions; minor formatting variations caused 125 episodes (6.1%) to yield no guests; multi-part surnames (e.g., *von Schönburg*) were systematically malformed or not even registered if they contained non-capital ASCII letters (e.g., ß, Ö).
2. **Missing pre-import disambiguation.** Person identity resolution was deferred to OpenRefine; near-duplicate name typos remained as separate items, and distinct persons sharing a name were automatically merged by default.
3. **Manual entity linking does not scale.** With 40.4% of persons requiring manual review, the effort is impractical for larger corpora.
4. **Absent temporal semantics.** Wikidata party memberships were queried without temporal filtering, causing historically inaccurate attributions: Bündnis Sahra Wagenknecht (founded in 2024) appeared in the 2012 timeline.
5. **Single-source design.** The pipeline was built around the ZDF archive structure and did not integrate episode data from fernsehserien.de or Wikidata into the Knowledge Graph.

3.2 Second Iteration: Informed Redesign

Taking the lessons learned from the first iteration into account, the approach was reworked from the ground up to the shape illustrated in Figure 1.

It is composed of primarily three coupled phases with downstream potential:

- **Phase 1** extracts and structures mentions from the ZDF archive export: Episodes, their publications, guests, topics, and related organizations.
- **Phase 2** generates relevant candidates for these instances in different sources and explores their connections to related items, potentially identifying additional relevant instances.
- **Phase 3** synthesizes this information by disambiguating matches describing the same event across sources (person X, guest of episode 1) and deduplicating within the resulting instance list (person X, guest of episode 1, equals person Y, guest of episode 3).

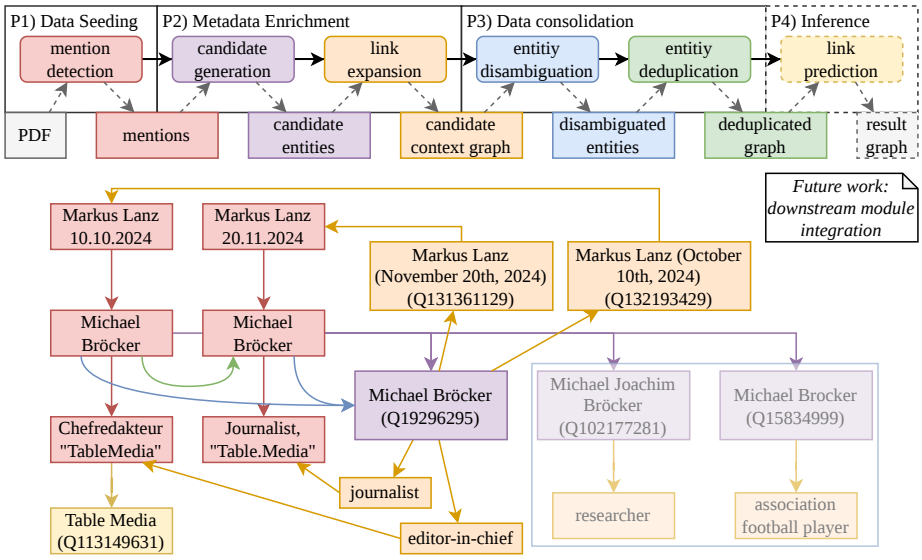


Fig. 1: Final approach after redesign.

– **Phase 4** is scoped as future work: it would inspect the remaining instance set for unresolved links and properties, attempting to predict additional linkages beyond the disambiguation and deduplication of Phase 3.

4 Implementation

The second iteration included not only ZDF Archive PDFs, but also Wikidata and fernsehserien.de as additional sources for the 15 configured programs. The pipeline is implemented as a sequence of Jupyter notebooks, each orchestrating a set of dedicated Python modules. Intermediate output projections are written to CSV and JSON files under a phase-bound directory contract, so any notebook can be rerun independently once its upstream inputs are available.

4.1 Phase 1: Text Extraction and Mention Detection

PDF-to-text conversion uses `pdfplumber`; each episode receives a SHA1-based stable identifier derived from title and first broadcast date. Guest extraction follows a four-strategy rule hierarchy, from a structured info-section parser to a surname fallback for ALL-CAPS tokens, in descending confidence order. A pattern-based institution extraction module was also initially developed; however, due to a high false-positive rate, this module was deferred until after semantic validation via Wikidata was concluded for additional context.

4.2 Phase 2: Candidate Generation

Wikidata candidate generation performs a relevancy-seeded BFS from the configured broadcasting programs: episodes inherit relevancy from their series, guests

from relevant episodes. Nodes are classified `basic_fetched` (label, description, alias, and class hierarchy), `all_outlinks_fetched`, or `full_fetched` (in- and outlinks) based on relevance rules. A precision-tuned threshold prevents runoff into unrelated entities; a permissive threshold causes relevance to spread to thousands of unrelated entities, while a strict threshold drops valid guests. Our implemented design leans toward precision: only seed broadcasts are `full_fetched`, shows listing them as `part of series` are `all_outlinks_fetched`, as well as their guests and a selection of guest property objects (e.g. *occupation*, *position*).

Wikidata graph expansion can take hours, and interruptions risk discarding current or corrupting prior progress. Logging infrastructure, issue tracing, checkpoints, iterative improvement, and append-only caching were needed. Initial runs required over 25,000 queries over several hours, creating an equal number of JSON artifacts to avoid data loss during a write interruption. While event-sourcing was not the initial design choice, it naturally emerged as a favorable solution and met all of the identified requirements. Every query result, entity fetching, triple, relevancy assignment, and rate-limit observation is written as an immutable event. The pipeline is now built around an append-only JSONL event log. Event handlers can replay this log at any time to regenerate all projection artifacts. Subsequent runs replay the cache in seconds.

The fernsehserien.de pipeline reused these event-sourcing and cache-first patterns from Wikidata. Once the Wikidata pipeline structure was robust and iteratively improved, adding this second source required minimal additional effort.

4.3 Phase 3: Entity Disambiguation

Phase 3 comprises three sequential steps: Step 3.1.1: Alignment, Step 3.1.2: Reconciliation, and Step 3.2: Deduplication.

Step 3.1.1: Alignment normalizes and aligns entities across all sources within seven classes: programs, seasons, episodes, persons, topics, roles, organizations. Primary alignment targets are episodes. Normalization applies a schema-mapping table that spans all properties across all sources, as well as accent removal, lowercasing, and whitespace collapsing before comparison. Normalized label matches and publication time matches, along with belonging to the same configured broadcasting program, were the highest-quality alignment signals. High-confidence near-matches were technically also possible, and the remainder were flagged `UNRESOLVED` with typed reason codes (*no_candidate*, *low_confidence*, *contradiction*, *insufficient_context*). Quality gates at the end of Step 3.1.1 validate row counts and schema completeness before the handoff to manual disambiguation.

Step 3.1.2: Reconciliation imports these rows to OpenRefine for manual reconciliation. Two curators independently reviewed separate alphabetical slices (A–L: 16,667 rows; M–Z: 15,043 rows).

Step 3.2 Deduplication clustered canonical persons by shared Wikidata QID or normalized name similarity as a lower-confidence fallback. Other entity classes could be sufficiently disambiguated via data-driven alignment (episodes and broadcasts) or were deferred (institutions and topics).

5 Results

Results! Why, man, I have gotten a lot of results! I know several thousand things that won't work.

Thomas Edison (Q8743)

The main result of our two-iteration approach is that additional iterations are needed. Resolving the known issues of the first iteration exposed a cascade of new ones, yet this growth proved frontloaded rather than exponential: the deeper we walked down any issue path, the more issues we could close and systems we could refine (such as our cached Wikidata interface), catching most future cases with minimal overhead. As such, we now provide insights into our use-case results and highlight the framework that aggregated them in the discussion section.

Of the 2,036 episodes extracted from the 2008 to 2024 ZDF corpus, only thirteen episodes produced no guest rows, down from 125 in the first iteration. Many of the remaining episodes without guests are correctly identified as moderator-only special episodes. When extracting the 6,459 episodes from fernsehserien.de, 1,639 were without guests, 2,434 were without a moderator. Where those could be matched to the ZDF Archive, these should have had guests, hinting at gaps in the database. The Wikidata BFS also identified 638 series instances, with 382,400 Wikidata triples spanning 8,221 unique classes and 2,324 properties.

In Step 3.1.1: Alignment, 2,673 of 6,849 episodes (39%) were successfully aligned across sources. Reduced to only episodes with guest data, we arrived at 4,861 episodes. The automated person alignment matched 17.2% (5,487) of the 31,817 person appearance rows as high-confidence matches. This context was presented to the OpenRefine curators, who conducted Step 3.1.2 Reconciliation: First, OpenRefine automatically reconciled 55.91% of the 31,817 person rows. Over 64 hours, these matches and reconciliation were confirmed and another 18.79% were manually mapped. For another 12.17%, no suitable candidate could be found, which mostly came down to encoding errors, single-word names, or organizations mislabeled as persons. 13.13% remain unresolved within the available time window. We arrive at 26,659 QID-matched entries of 8,436 canonical persons, of which 5,257 (62.3%) could be linked to Wikidata. Reduced to only guests of our target episodes, we arrive at 23,527 appearances of 8,287 unique guests.

We classify this data along four data quality tiers: Tier 1: 2,394 (28.4%) persons have a Wikidata QID and are annotated in at least two sources; Tier 2: 2,863 (33.9%) have a Wikidata QID and only one source; Tier 3: 628 (7.4%) are disambiguated via two non-Wikidata sources but have no QID; Tier 4: 2,551 (30.2%) are single-source only, with no QID.

Alphabetically sorting and stacking same-name entries alongside their appearance timelines proved highly effective for manual review. Data quality issues illustrate the value of structured data: We found institutions and editorial staff tagged as guests, prefixes (e.g., "Prof.") or relationship descriptors captured as names (e.g., "her son Bob"), or compound descriptors (e.g., "family Miller"). Due to alphabetical sorting, most of these could quickly be batch-resolved.

5.1 Analysis

An episode-guest occurrence matrix was built and used to navigate the analysis. Every guest's Wikidata entity was fully resolved with all outgoing links, enabling by-property analysis across 15 specified properties. Of the 8,436 deduplicated persons, 39 are moderators, 22 staff members, and 88 additional persons incidentally captured (such as episode topics or relatives of guests), yielding $8{,}436-39-22-88 = 8{,}287$ guests. The 15 programs were selected in four groups: The first group comprises the German political talk shows most extensively studied by prior work [5,2,10], enabling cross-validation against independently derived figures. The shows are *Markus Lanz*, *Maischberger*, *Hart aber fair*, *Maybrit Illner*, *Caren Miosga*, and *Phoenix Runde*. The second group of *Presseclub* and *Der internationale Frühschoppen* follows the same format and audience profile but was absent from those analyses, making it a natural extension. The third group of *Precht*, *Lanz und Precht*, and *unbouble* shares structural properties with political talk shows while differing in scope or medium, airing primarily online, and are testing the pipeline generalizability across format variation. *StarTalk*, *scobel*, *couchwissen*, and *wtf_talk* are included to extend this corpus towards science communication, broadening the scope from the political-talk-show focus of prior analyses [2,5] toward the cross-format vision of WissKomm Wiki [18]. Five programs ultimately lacked usable guest data: *unbouble*, *wtf_talk*, and *Lanz und Precht* had no extractable data at all; *Phoenix Runde* and *couchwissen* had episode data but no guest entries. Table 2 shows the per-show breakdown for the 12 with at least some episode data. Analyses of Wikidata-sourced properties (gender, age, party, occupation) apply only to the 5,257 Tier 1 and Tier 2 persons with Wikidata links; the remaining 37.7% of canonical persons carry no property data and are excluded from all property-based analyses below.

Gender and Age. Of 5,016 guests with gender annotations, 3,127 (62.34%) were male and 1,880 (37.48%) female; male appearances accounted for 65.76% of appearances with known gender. Nine individuals (0.18%) carry non-binary annotations (non-binary: 4, trans woman: 2, trans man: 1, agender: 1, transmasculine: 1). Of the 4,861 episodes with guest data, we could not identify a female guest in 16.97%, compared to 3.64% for male guests. Figure 2 shows that the gender gap has been narrowing since 2006 but the consistent decline appears to have ceased around 2020, partly because *Markus Lanz* has again featured more male guests. A Spearman rank correlation ($\rho = -0.72$, $p < 0.001$) confirms the long-term decline; a monthly split search across all 4,819 dated episodes identifies April 2020 as the earliest month after which the declining trend is no longer statistically significant (post-period $\rho = -0.17$, $p = 0.14$). While scalable topic analysis was deferred, early manual analysis showed that one episode with the topic *Was Frauen wollen* (what women want) contained no female guests.

A methodological ambiguity warrants attention: the identified gender gap may partly reflect Wikidata's systematic underrepresentation of women[6] rather

[6] https://meta.wikimedia.org/wiki/Women_in_Wikipedia/Home

Table 2: Per-show episode and guest statistics across the 12 analyzed programs.

Broadcasting Program	Episodes		Guests		Episodes without ...	
	Total	with Guest Data	Appearances	Unique	Male Guest	Female Guest
Total	**6,469**	**4,861 (75 %)**	**23,527**	**8,287**	**138 (2 %)**	**787 (12 %)**
Markus Lanz	2,248	2,205 (98 %)	11,469	5,234	36 (2 %)	331 (15 %)
Maischberger	952	661 (69 %)	3,542	1,581	31 (3 %)	104 (11 %)
Hart aber fair	835	518 (62 %)	2,768	1,519	6 (1 %)	75 (9 %)
Presseclub	806	402 (50 %)	1,598	573	8 (1 %)	39 (5 %)
Maybrit Illner	587	450 (77 %)	2,413	871	2 (0 %)	36 (6 %)
scobel	450	231 (51 %)	663	581	30 (7 %)	74 (16 %)
Frühschoppen	326	218 (67 %)	925	177	4 (1 %)	40 (12 %)
Precht	85	68 (80 %)	68	62	19 (22 %)	42 (49 %)
Caren Miosga	79	74 (94 %)	231	137	2 (3 %)	14 (18 %)
StarTalk	60	34 (57 %)	70	11	0 (0 %)	32 (53 %)
couchwissen	37	0 (0 %)	0	0	—	—
Phoenix Runde	4	0 (0 %)	0	0	—	—

than editorial selection alone, since guests without Wikidata entries contribute no property data. Yet, even if none of the 3,271 guests without gender annotations were male, the overall male appearance share would not fall under 50%. The true unique-guest male rate lies within $[37.7\%, 77.2\%]$, and the appearance-level rate within $[50.8\%, 73.6\%]$. The observed rates of 62.3% (unique guests) and 65.8% (appearances) apply only to the Wikidata-annotated subset. Inferring gender from ZDF and fernsehserien.de descriptions was explored but rejected: The German position and occupation labels often encode the feminine form but have no equivalent signal for masculine, introducing a directional skew rather than correcting one. First-name-to-gender mapping was similarly ruled out.

A distributional difference emerges in the age data (Figure 3): male guests have a median age of 54 versus 49 for female guests, and the pattern of older men and younger women is visible across most programs. A Mann-Whitney U test confirms this ($p = 2.8 \times 10^{-187}$, rank-biserial $r = 0.27$): male guests were older in 63% of our recorded gender-tagged appearance pairs, a small-to-medium effect. Science communication formats (*scobel*, *StarTalk*) skew toward older expert guests; Political formats (*Hart aber fair*, *Maischberger*) show broader age distributions, reflecting the different expertise profiles each format draws on.

Occupation, Party, and Property Coverage. Only 55 appearances were by guests classified as scientist (Q901) directly, but 725 by economists (Q188094, a

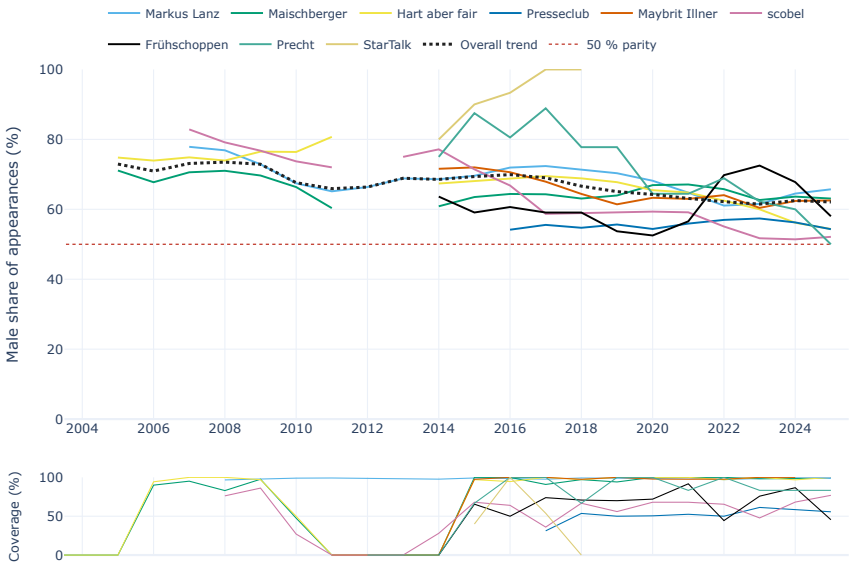
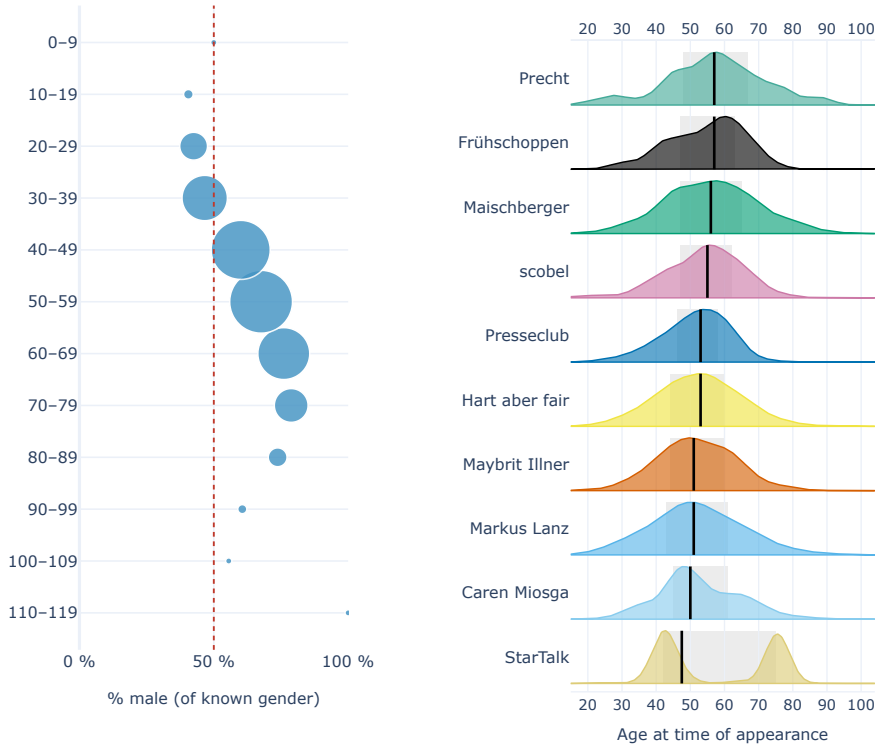


Fig. 2: Gender share of guest appearances over time. The 2011 gap reflects missing source coverage on guest data for those years. The long-term trend toward gender parity appears to have ceased around 2020.

subclass), illustrating why subclass-aware queries are essential. Yet, Wikidata's occupation taxonomy contains loops (scientist $\xrightarrow{P279}$ researcher $\xrightarrow{P279}$ academic professional $\xrightarrow{P279}$ academic $\xrightarrow{P279}$ scientist) and requires pipeline-level resolution. Person-to-organization relationships proved sparser than expected: *Die Welt* appears only 110 times as employer despite Robin Alexander appearing 131 times and having been *Die Welt*'s deputy editor-in-chief from 2019 to 2025, but with an empty Wikidata employer field. Wikidata nonetheless proved a very beneficial resource for visualization: *short labels (P1813)* reduced "Bündnis Sahra Wagenknecht – Vernunft und Gerechtigkeit" to "BSW", and party colors (*P462*/*P6364*/*P465*) enabled consistent political-spectrum ordering across visualizations. Both properties would have required extensive manual curation to replicate and maintain consistency. Additionally, labels of each visualization can be translated to any Wikidata-covered language almost automatically. Additional statistics and visualization permutations are available in the repository[7].

Table 3 shows Wikidata property coverage across the 10 programs for all 8,287 analyzed guests. Date of birth (P569) is well covered for prominent guests but sparse for lesser-known individuals. Occupation (P106) is moderately covered, though the hierarchy issues documented above complicate aggregation. Coverage patterns generally align with show type: political shows have better

[7] https://github.com/borgnetzwerk/speaker-mining



(a) Male share by guest age.

(b) Age distribution per program.

Fig. 3: Age and gender distributions across the analyzed broadcasting programs.

party affiliation and position held coverage. Interestingly, though, there seems to be no clear indicator for the small sample of scientific programs. Coverage is computed over all guests per show; the 3,179 Tier 3 and Tier 4 guests (37.7%) carry no Wikidata data and contribute zero coverage for all properties, pulling per-show rates below those of the Wikidata-linked subset alone. The Wikidata graph also supports page-rank and co-occurrence analyses; structurally central entities (ZDF, Markus Lanz) rank highest, confirming the graph's integrity.

5.2 Threats to Validity

Construct validity. Guest appearance counting potentially conflates frequency with importance. Moreover, our data captures *presence*, not *participation*: a guest listing carries no information about speaking time, interruption patterns, or discourse content. Measuring actual participation would require complementary transcript-based speaker analysis; our knowledge graph provides the struc-

Table 3: Wikidata property coverage for guests (n = 8,287), total and per show.

	Total	Markus Lanz	Maischberger	Hart aber fair	Presseclub	Maybrit Illner	scobel	Frühschoppen	Precht	Caren Miosga	StarTalk
Wikidata entry	62 %	61 %	72 %	72 %	77 %	77 %	71 %	72 %	90 %	93 %	91 %
sex or gender	60 %	60 %	70 %	70 %	75 %	76 %	69 %	70 %	90 %	92 %	91 %
occupation	59 %	59 %	68 %	68 %	74 %	75 %	69 %	68 %	90 %	91 %	91 %
date of birth	58 %	58 %	69 %	69 %	70 %	74 %	66 %	67 %	90 %	91 %	82 %
place of birth	52 %	53 %	64 %	65 %	57 %	69 %	53 %	57 %	90 %	86 %	82 %
country of citizenship	50 %	51 %	63 %	63 %	53 %	67 %	51 %	55 %	87 %	85 %	82 %
Commons category	36 %	40 %	51 %	52 %	34 %	56 %	28 %	33 %	82 %	79 %	73 %
employer	18 %	15 %	26 %	22 %	29 %	32 %	44 %	34 %	53 %	43 %	27 %
award received	20 %	21 %	31 %	26 %	20 %	28 %	20 %	18 %	63 %	36 %	46 %
social media followers	10 %	12 %	17 %	17 %	16 %	27 %	6 %	15 %	32 %	37 %	64 %
position held	11 %	10 %	22 %	25 %	8 %	32 %	5 %	8 %	24 %	42 %	9 %
member of political party	11 %	10 %	23 %	27 %	3 %	33 %	3 %	2 %	21 %	39 %	9 %
academic degree	6 %	5 %	9 %	10 %	5 %	14 %	14 %	5 %	29 %	18 %	9 %
religion or worldview	4 %	4 %	10 %	9 %	1 %	12 %	2 %	2 %	11 %	18 %	27 %
number of viewers/listeners	1 %	1 %	1 %	1 %	0 %	1 %	0 %	1 %	2 %	2 %	9 %
political ideology	0 %	0 %	1 %	0 %	0 %	0 %	0 %	0 %	3 %	1 %	0 %

tured backbone such analyses can build upon and be integrated into. Wikidata properties are crowd-maintained and variable in completeness; 37.7% of deduplicated persons are without Wikidata links and carry no property data. All Wikidata-dependent findings (gender, age, occupation, party) are conditional on this linked subset and should not be interpreted as representative of the full guest population. This selection bias is structural, not incidental: women, private citizens, and recent or regional public figures are systematically underrepresented in Wikidata, meaning that representational patterns observed in the linked subset might not reflect the full guest population. Any interpretation of gender, age, or occupation distributions must account for this compounding coverage effect. At the same time, every analysis and inspection of Wikidata claims identifies issues and increases their overall quality. Wikipedia has shown that crowdsourcing this knowledge leads to higher-quality knowledge representation overall. Our experience is consistent with this. Overall, observed disparities cannot be cleanly attributed to editorial guest selection yet, and representative investigations of non-Wikidata-annotated guests are recommended as future work to assess the degree of this effect.

Internal validity. Episodes without guests in the ZDF archive and fernsehserien.de remain unaccounted for. Relevance propagation in Wikidata cannot reach guests unless their episode correctly references both the seed and the guest.

100 % deduplication rate is likely neither achievable nor desirable in practice: Limiting factors include ambiguous or absent context, inaccurate source data, and most importantly GDPR considerations. While our analysis excludes properties of individuals not captured in Wikidata, it also excludes the uncertain noise of duplicate guests and malformed textual data interpretation. The semi-automated reconciliation was laborious, but also intentionally precision-tuned to ensure the resulting mapping is of high quality. The two curators reviewed disjoint alphabetical slices; quantifying inter-annotator agreement via overlapping slices is recommended for future work.

External validity. Most sources are German-language, the pipeline was developed and evaluated primarily on the Markus Lanz corpus. The ZDF Archive PDF structure, the Wikidata entity landscape and especially fernsehserien.de are particular for this dataset on German public television; but applying the pipeline to a different language context, broadcaster, or show format would only require retuning extraction regexes, and seed QIDs. The degree of manual effort required for Step 3.1.2 Reconciliation will vary substantially with corpus size and source quality. Yet we found that this work is heavily frontloaded, and picking up where we left off should be much easier than starting from scratch, as ours had to.

Conclusion validity. All results should be treated as indicative rather than exhaustive. In particular, qualifier-based assessments, such as temporal properties, are implemented, but are currently dropping potentially relevant claims from the past that qualified guests for their appearance (e.g., former presidents). The current pipeline state represents a proof of concept with quantified gaps, not a production system with end-to-end coverage. To our knowledge, no gold-standard guest corpus annotation exists against which precision or recall could be computed; establishing one is a recommendation for future error analysis.

6 Discussion

FAIR data infrastructure is the bottleneck. A substantial share of the overall project time was spent processing raw, unstructured data sources. Every source had systematic gaps: the ZDF Archive had encoding failures and malformed episodes; fernsehserien.de had episodes without guest data. The root cause is structural: as Kaufmann noted [10], *"it would make much more sense if the public broadcasters maintained this information as a knowledge graph on their own platforms and then linked the individuals via Wikidata or GND ID"*. We recommend merging Phase 1 and 2 into a single extraction step that aggregates all sources and aligns them on a shared Wikibase before querying Wikidata. Infrastructure efforts complementary to this are rufus [3] for institutional archive access and WissKomm Wiki [19] for crowdsourced curation. Collaborative ecosystems built around linked open data already enable federated SPARQL queries from our speakermining.wikibase.cloud to the Wikiverse.

At the data level, future work should connect additional sources (e.g., YouTube, Spotify) and untapped properties (e.g., topics, audience ratings, qualifiers, references); preliminary YouTube-based transcript retrieval for *Lanz und Precht* has

been explored as a proof of concept. Transcript-based speaker (share) annotation is particularly promising, yet faces even more complex GDPR and copyright challenges. While expanding the data model accordingly, a mapping to the academic wheel of privilege [8] could serve as a powerful framework for multidimensional representation analysis.

Disambiguation at scale is likely frontloaded. The manual curation bottleneck documented in Results, and the Step 3.1.1 Alignment automation ceiling for persons, are the clearest signals that entity disambiguation cannot yet be called scalable. At the same time, as the system improved and lessons were learned, we could reduce the number of manual matches required. If we scale all three pillars, from data providing from broadcasters to data processing from pipelines such as ours to data curation through crowdsourced infrastructures, we assume the largely front-loaded curation task will eventually become achievable. The Wikiverse has repeatedly shown that large-scale knowledge work can be achieved when the community has sufficient infrastructure.

Despite current gaps, the curated dataset already enables questions previously unanswerable at scale: Which shows have the least gender-balanced guest profiles? Which occupational groups are systematically absent from political discourse? How has guest diversity evolved over 20 years? These are the questions motivating the SciCom KI, and they require a structured, community-driven infrastructure that this work contributes to.

Sustainability focused via SciCom KI. The long-term curation for this data can be facilitated in the knowledge infrastructure for science communication. Particularly the WissKomm Wiki [18] is designed as a sustainable, crowdsourced successor to isolated efforts like our related work: using a shared knowledge base, editorial governance, and community tooling for individual researchers and projects. Speaker Mining provides a proof of concept for WissKomm Wiki, demonstrating which data can be curated at scale, which disambiguation workflows are reusable, and which legal and technical open questions must be resolved before such a platform can operate at full scope. The speakermining.wikibase.cloud instance and the pipeline developed here will be migrated into WissKomm Wiki as a founding dataset, alongside the lessons learned on event-sourced extraction and the extent of Wikidata coverage into the next iteration. These challenges require coordination with data protection expertise as well as ethics and open science standards. Every such contribution reduces the technological and legal uncertainty currently hindering the SciCom KI, sustainably reduces onboarding costs, and advances towards FAIR question answering on public broadcasts at scale.

7 Conclusion

This paper presents a scaling-oriented framework for FAIR curation of public-broadcast guest data, evaluated across 15 configured broadcasting programs. Starting from three heterogeneous sources, we produced a unified, queryable knowledge base of 8,436 canonical persons with 23,527 appearances across 6,469

aligned episodes, published at speakermining.wikibase.cloud. Our contributions are (1) a reusable, open-source pipeline for multi-source broadcast metadata curation; (2) a curated FAIR dataset covering 12 broadcasting programs with episode data (10 with guest data); (3) quantitative analyses of primarily gender, age, party, and occupation distributions across the 10 analyzed programs. Within the Wikidata-annotated subset, a gender gap is visible in our data slice of public broadcasting: male guests account for 65% of appearances with known gender, and a distributional age difference is observed towards older male guests. The long-term trend toward gender parity appears to be slowing since 2021, though the recency of this period limits how firmly this can be characterised. Our main contribution, the pipeline and infrastructure enabling this analysis, is a promising proof of concept, with much potential to grow and contribute towards a crowdsourced SciCom KI. Complementary efforts such as the Wikiverse, rufus and WissKomm Wiki[8] [19,18] face similar technical and legal challenges ahead. To overcome them, we must share not only data, but also open-source software, reconciliation workflows, and legal lessons learned, fostering collaboration of public broadcasters, research institutions, and individuals across fields towards scaling FAIR data on public broadcasts for reliable question answering.

Acknowledgments. This study was funded as part of the WissKomm Wiki project, part of the Lower Saxony Research Data Management Initiative (FDM-NDS), a funding program by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation. We want to acknowledge help from the ZDF archive (Veit Scheller), Armin Müller-von Fischer (arrrrrmin) and Stefan Kaufmann (stk), as well as Der Spiegel for providing data and support, and our students Jamal Eldemashki, Abdul Aahad Qureshi, and Rownak Deka for their extensive work on the wikibase and dataset curation.

Data Availability Statement. The knowledge graph is publicly queryable at speakermining.wikibase.cloud/query; user documentation with the data model, example entities, and tested SPARQL queries is available on the instance's main page and in the repository[9]. Wikidata-derived projections, aggregated statistics and a complete RDF dump are being published incrementally in line with legal requirements. The WissKomm Wiki project [18] was in part created to resolve this challenge; the Speaker Mining repository documents the current status.

Use of AI tools declaration. During the preparation of this work, the author(s) used **GitHub Copilot**, **Claude Code**, **DeepL**, **Grammarly**, **LanguageTool** in order to: **translate text**, **grammar and spelling check**, **paraphrase and reword**, **peer review simulation**, according to the CEUR GenAI Usage Taxonomy[10]. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

[8] https://tech-news.wikimedia.de/2026/06/24/knowledge-wants-to-be-found/
[9] https://github.com/borgnetzwerk/speaker-mining
[10] https://ceur-ws.org/GenAI/Taxonomy.html

References

1. Arabi, S., Zheng, X., Hutchins, B.I., Ni, C.: Equity in Science Journalism: Investigating Gender Disparities in News Media Coverage of Scientific Research. Science Communication (2025). https://doi.org/10.1177/10755470251360187
2. Arrrrrmin: Lanzmining: Der journalismus (2025), https://lanz-mining.arrrrrmin.dev
3. Blume, P.F.: Public Broadcasting Archives and Academic Use in Germany. Scopes of the Portal for Broadcast Search 'rufus'. VIEW Journal of European Television History and Culture **13** (2024). https://doi.org/10.18146/view.336
4. Brüggemann, M., Lörcher, I., Walter, S.: Post-normal science communication: Exploring the blurring boundaries of science and journalism. JCOM **19**(3), A02 (2020). https://doi.org/10.22323/2.19030202
5. Buß, C., Dambeck, H., Krug, N., Ohdah, D., Tack, A.: Talkshow-könig karl, berlins balzac und der ewige tv-schattenkanzlerkandidat (2025), https://www.spiegel.de/kultur/tv/sandra-maischberger-markus-lanz-und-caren-miosga-talkshows-in-der-datenanalyse-a-b7e46530-d826-4424-95c5-6c2f220595b1?sara_ref=re-xx-cp-sh
6. CL, C.M., MG, W.: The role of media professionals in perpetuating and disrupting stereotypes of women in Science, Technology, Engineering and Math (STEM) fields. Frontiers in Communication (2022). https://doi.org/10.3389/fcomm.2022.1027502
7. Davidsona, N.R., Greene, C.S.: Analysis of science journalism reveals gender and regional disparities in coverage. eLife **12** (2024). https://doi.org/10.7554/eLife.84855.3
8. Elsherif, M.M., Middleton, S.L., Phan, J.M., Azevedo, F., Iley, B.J., Grose-Hodge, M., Tyler, S.L., Kapp, S.K., Gourdon-Kanhukamwe, A., Grafton-Clarke, D., Yeung, S.K., Shaw, J.J., Hartmann, H., Dokovova, M.: Bridging neurodiversity and open scholarship: How shared values can guide best practices for research integrity, social justice, and principled education (2022). https://doi.org/10.31222/osf.io/k7a9p, https://osf.io/k7a9p
9. Enzingmüller, C.: Visuelle wissenschaftskommunikation: Ein forschungsüberblick, https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-52199
10. Kaufmann, S.: (2025), https://stefan.bloggt.es/2025/06/wer-war-wie-oft-bei-lanz-schoener-leben-mit-linked-data/
11. Peng, H., Teplitskiy, M., Jurgens, D.: Author mentions in science news reveal widespread disparities across name-inferred ethnicities. Quantitative Science Studies **5** (2024). https://doi.org/10.1162/qss_a_00297
12. Phogat, P., Rab, S., Wan, M.: Science communication in the digital age: Trends, gaps, and interdisciplinary opportunities. Information Services and Use **45**(1), 148–163 (2025). https://doi.org/10.1177/18758789251342896
13. Remmo, O.I.: LanzMining aber FAIR: Empirische Fragen zur Medienlandschaft mittels FAIRer Talkshow-Daten beantworten (2026). https://doi.org/10.15488/20751
14. Riedl, A.A., Rohrbach, T., Krakovsky, C.: "I Can't Just Pull a Woman Out of a Hat": A Mixed-Methods Study on Journalistic Drivers of Women's Representation in Political News. Journalism & Mass Communication Quarterly **101** (2024). https://doi.org/10.1177/10776990211073454
15. Rosen, C., Guenther, L., Froehlich, K.: The Question of Newsworthiness: A Cross-Comparison Among Science Journalists' Selection Criteria in Argentina, France, and Germany. Science Communication **38** (2016). https://doi.org/10.1177/1075547016645585

16. Ruf, O.: Humor und wissenschaftskommunikation: Ein forschungsüberblick, https://transferunit.de/thema/humor-in-der-wissenschaftskommunikation/
17. Schwind, M.: Wissenschaftskommunikation – konzepte und begriffe: Ein forschungsüberblick (2024), https://transferunit.de/thema/wissenschaftskommunikation-konzepte-und-begriffe/
18. Wittenborg, T., Stöhr, M., Rossenova, L., Auer, S.: WissKomm Wiki: Digitale Wissens-Infrastruktur für audio-visuelle Forschungsdaten. Zenodo (2026). https://doi.org/10.5281/zenodo.18714623
19. Wittenborg, T., Stehr, N., Karras, O., Auer, S.: SciCom Wiki: A Digital Library to Support the Science Communication Knowledge Infrastructure for Videos and Podcasts. arXiv Preprint (2025). https://doi.org/10.48550/arXiv.2511.09248