

A Review Methodologies in HCI

Review methodologies in HCI face persistent challenges when systematic transparency must coexist with interpretive expertise and when evidence circulates across fragmented publication ecosystems [49, 141]. Systematic reviews establish protocols for organizing evidence within bounded domains, and PRISMA frameworks support reproducibility [123]. Both approaches reach limits in interdisciplinary settings where evidence types vary and epistemological frameworks conflict [6, 113].

Multivocal literature reviews (MLRs) offer one way to meet these demands and balance rigor with practical utility, as recent work in responsible AI demonstrates [99]. The approach originated in educational research as a methodological framework to impose systematic rigor on reviews of diverse documents [118], initiating a discussion that clarified that the standard of rigor must be situational and secondary to utility for practitioners [127]. Software engineering later adapted MLRs to capture both state-of-the-art research and state-of-practice knowledge [54, 55]. MLRs integrate peer-reviewed academic literature with grey literature such as organizational reports, policy documents, community statements, technical documentation, practitioner outputs, and multimedia materials. The widely cited Luxembourg definition characterizes grey literature as material produced by government, academia, industry, or community groups that is not controlled by commercial publishers [57]. Diversity of source types and timeliness are core advantages, since emerging practices often circulate outside formal publication channels and appear earlier than peer-reviewed work [122].

Credibility varies across grey-literature types, and assessments often depend on provenance, expertise, and recognized authority [77]. Multivocal approaches are particularly important for scholarship involving Indigenous and underserved communities. Lewis et al. [92] show that multivocality preserves heterogeneous viewpoints. The authors combine essays, protocols, and artistic works rather than imposing a single scholarly mode, a stance aligning with decolonizing methodologies that emphasize community-generated knowledge and Indigenous epistemic authority [10, 159]. Additionally, influential contributions in AI ethics and data governance often appear in organizational reports [26, 72], investigative journalism [65], public initiatives [1], and widely cited preprints [15]. MLR methods accommodate a diversity of distributed knowledge production and support synthesis across venues not fully captured by academic indexing.

B Search Strategy & Corpus Composition

Database searches were conducted iteratively between August 2024 and January 2025, complementing network referrals and citation snowballing. The final structured ACM Digital Library search was executed on January 31, 2025, using the advanced search interface with abstract and full-text indexing via personal subscription. Table 3 reports the four primary ACM query sets and their outcomes.

Aggregate results from the January 2025 ACM searches are summarized in Table 5. Across 1,914 hits, 1,201 items were screened, yielding 153 that met criteria and 48 unique sources after full-text review and duplicate removal. Similar comprehensive search strategies were applied to IEEE Xplore, ScienceDirect, Taylor & Francis Online, Wiley Online Library, Google Scholar, and Springer Link, following the same phased approach for queries.

Table 3. ACM Digital Library search queries and results.

Query	Exact search string	Results screened	Potentially relevant
Q1	((("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	51 abstracts + 300 full-texts	45 (7 abstracts, 38 full-texts)
Q2	((("extractive" OR "exploitative" OR "data colonialism") AND ("data practices" OR "dataset construction") AND ("communities" OR "workers" OR "labor"))	3 abstracts + 182 full-texts	13 (1 abstract, 12 full-texts)
Q3	((("crowdsourcing" OR "platform labor") AND ("bias" OR "fairness" OR "ethics") AND ("marginalized" OR "vulnerable populations" OR "community harm"))	491 full-texts (200 screened)	32
Q4	((("participatory design" OR "community-led" OR "co-design") AND ("ai development" OR "dataset creation") AND ("sovereignty" OR "community engagement" OR "ethical data"))	122 full-texts	12

Table 4. Targeted ACM venue-specific searches.

Venue	Exact search string	Venue filter	Results summary
CHI Conference Proceedings	((("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	CHI Conference on Human Factors in Computing Systems (all years)	296 hits; screened: first 200; potentially relevant: 15
FAccT Proceedings	((("data collection" OR "data production" OR "data curation" OR "dataset development") AND ("artificial intelligence" OR "machine learning" OR "AI") AND ("marginalized" OR "underrepresented" OR "underserved" OR "community" OR "indigenous"))	ACM Conference on Fairness, Accountability, and Transparency	143 hits; screened: all; potentially relevant: 37

Table 5. ACM search results summary.

Category	Count
Total primary searches (query sets)	6
Total venue-specific searches	2
Total hits across all searches	1,914
Total items screened (varied by search size)	1,201
Items meeting inclusion criteria after screening	153
Items retained after full-text review	89
Final unique sources for corpus (after duplicate removal)	48

Table 6. Discovery method distribution (N=350 sources).

Method	Sources	Percent
Database searches	174	50%
Existing networks/organizations	73	21%
Citation snowballing	51	15%
Iterative keyword search	31	9%
Hand-searching journals	21	6%
Total	350	100%

C Corpus Creation Details

Datasheet Fields.

Table 7. Description of datasheet fields

Column	Content
Identifier	In-line APA citation (author surname and year) used as a unique ID for tracking within the corpus.
APA Citation	Full APA reference for the source.
Title	Title of the publication or output.
Analytic Domain	Controlled list, multiple possible: <ul style="list-style-type: none"> • Data Relations: how data is scoped, justified, and negotiated • Data Labor: how curation work is arranged and carried out • Data Representation: how categories are constructed and based on what presences and absences • Data Infrastructure: how technical systems mediate data movement • Data Governance: how authority over data shapes downstream use
Orientation	Controlled list: <ul style="list-style-type: none"> • Extractive: undermines consent, compensation, or benefit • High-Agency Principles: normative frameworks promoting stewardship, sovereignty, accountability • High-Agency Practices: operationalized, community-led, participatory, or sovereignty-based initiatives
General Theme	Controlled list, multiple possible: <ul style="list-style-type: none"> • Community impacts and relations • Critical theory • Data labor • Data practices • Ethics frameworks

Column	Content
Pipeline Stage	Specific process within a pipeline stage (controlled list): <ul style="list-style-type: none"> • Problem Understanding and Formulation • ML System Design and Development • Deployment and Impact • Cross-pipeline
Pipeline Sub-stage	Specific process within a pipeline stage (controlled list): <ul style="list-style-type: none"> • Institutional Prioritization and Funding • Product Conception and Design • Data Selection, Collection and Annotation • Model Architecture Selection and Design • Model Training and Evaluation • Product Testing • Product Launch • Cross-pipeline
Historical Era	Era of data production practice (controlled list): <ul style="list-style-type: none"> • Era 1: Curated datasets (pre-2009); no sources in corpus • Era 2: Crowdsourced benchmarks (2009–2017) • Era 3: Web-scraped/foundation models (2017–present) • Multi-era: Spans multiple eras or provides historical analysis
Primary Pattern(s) / Pathway(s)	The specific extractive or high-agency behavior described in the source. Between one and three tags were assigned per source in order of relevance. For sources that provide conceptual, historical, or framing contributions without mapping directly onto an identified pattern, we assigned Other/NA (conceptual framing).
Triangle Coverage	Engagement with the three scoping domains that defined corpus eligibility: <ul style="list-style-type: none"> • A — AI contexts • D — Data production practices • C — Community impacts <p>Because community impacts (C) establish the outer bounds of the review, included sources substantively address all three domains, though with varying emphases. Codes (A, D, C or combinations ADC, DC, AD) indicate which domains are explicitly developed within the source. An accompanying Rationale column explains the basis for inclusion and the specific ways each source engages A/D/C beyond passing mention.</p>

Column	Content
How Source Was Found	White literature (journal papers, conference proceedings, books) or Grey literature (reports, policy documents, theses, community outputs, blogs).
Keywords	3–5 terms for coding/search, ordered Geography → Data/technical → Community/impact.
Geographic Region of Focus	Region or community under study (controlled list): Africa, APAC, EU/UK, LatAm, MENA, North America, Oceania, Multiple regions, Not regionally specific (globally framed advocacy, transnational collectives, or technical works not tied to one region).
Author Affiliations	High-level institutional grouping of authors. If multiple affiliations, code majority grouping here; record full details in Authorship & Positionality Context. Controlled list: Academic; Government; Industry; NGO/Non-profit; Mixed; Journalist/Other/Not sure
Geographic Area of Author(s)	Full institution name and country of the lead author(s)
Institution	Region of lead author's institution (controlled list, same regions as above).
Authorship and Positionality Context	Complete authorship profile, including all institutions, geographic distribution, equal contribution notes, and any relevant statements on positionality or disciplinary traditions.
Summary	≤ 120-word synopsis. Structure: Topic → Method → Findings → Link to AI data production + community impacts.

Corpus Summary.

Table 8. Corpus composition summary (N=350 sources)

Category	Sub-category	Count (%)
Orientation	Extractive Practices	141 (41%)
	High-Agency Principles	116 (33%)
	High-Agency Practices	93 (27%)
Source Type	White literature	258 (74%)
	Grey literature	92 (26%)
Geographic Focus	Not regionally specific	150 (43%)
	Multiple areas	61 (17%)
	North America	41 (11%)
	Africa	38 (11%)
	Oceania	19 (5%)
	APAC	15 (4%)
	EU/UK	11 (3%)
	LatAm	12 (3%)
Author Affiliation (lead only)	MENA	3 (1%)
	Academic	182 (52%)
	Mixed	96 (27%)
	Industry	37 (11%)
	NGO/Non-profit	20 (6%)
	Journalist/Other	11 (3%)
Pipeline Stage	Government	4 (1%)
	ML System Design & Development	183 (52%)
	Problem Understanding & Formulation	90 (25%)
	Cross-pipeline	55 (16%)
Historical Era	Deployment & Impact	22 (6%)
	Era 3 (2017–present)	241 (69%)
	Multi-era	95 (27%)
	Era 2 (2009–2017)	14 (4%)
	Era 1 (pre-2009)	0 (0%)

Received 12 February 2026; revised 5 June 2009; accepted 5 June 2009