

### 1353 **A Review Methodologies in HCI**

1354 Review methodologies in HCI face persistent challenges when systematic transparency must coexist with interpretive  
1355 expertise and when evidence circulates across fragmented publication ecosystems [49, 141]. Systematic reviews establish  
1356 protocols for organizing evidence within bounded domains, and PRISMA frameworks support reproducibility [123].  
1357 Both approaches reach limits in interdisciplinary settings where evidence types vary and epistemological frameworks  
1358 conflict [6, 113].

1361 Multivocal literature reviews (MLRs) offer one way to meet these demands and balance rigor with practical utility, as  
1362 recent work in responsible AI demonstrates [99]. The approach originated in educational research as a methodological  
1363 framework to impose systematic rigor on reviews of diverse documents [118], initiating a discussion that clarified  
1364 that the standard of rigor must be situational and secondary to utility for practitioners [127]. Software engineering  
1365 later adapted MLRs to capture both state-of-the-art research and state-of-practice knowledge [54, 55]. MLRs integrate  
1366 peer-reviewed academic literature with grey literature such as organizational reports, policy documents, community  
1367 statements, technical documentation, practitioner outputs, and multimedia materials. The widely cited Luxembourg  
1368 definition characterizes grey literature as material produced by government, academia, industry, or community groups  
1369 that is not controlled by commercial publishers [57]. Diversity of source types and timeliness are core advantages, since  
1370 emerging practices often circulate outside formal publication channels and appear earlier than peer-reviewed work  
1371 [122].

1372 Credibility varies across grey-literature types, and assessments often depend on provenance, expertise, and rec-  
1373 ognized authority [77]. Multivocal approaches are particularly important for scholarship involving Indigenous and  
1374 underserved communities. Lewis et al. [92] show that multivocality preserves heterogeneous viewpoints. The authors  
1375 combine essays, protocols, and artistic works rather than imposing a single scholarly mode, a stance aligning with  
1376 decolonizing methodologies that emphasize community-generated knowledge and Indigenous epistemic authority  
1377 [10, 159]. Additionally, influential contributions in AI ethics and data governance often appear in organizational reports  
1378 [26, 72], investigative journalism [65], public initiatives [1], and widely cited preprints [15]. MLR methods accommodate  
1379 a diversity of distributed knowledge production and support synthesis across venues not fully captured by academic  
1380 indexing.

### 1381 **B Search Strategy & Corpus Composition**

1382 Database searches were conducted iteratively between August 2024 and January 2025, complementing network referrals  
1383 and citation snowballing. The final structured ACM Digital Library search was executed on January 31, 2025, using  
1384 the advanced search interface with abstract and full-text indexing via personal subscription. Table 3 reports the four  
1385 primary ACM query sets and their outcomes.

1386 Aggregate results from the January 2025 ACM searches are summarized in Table 5. Across 1,914 hits, 1,201 items  
1387 were screened, yielding 153 that met criteria and 48 unique sources after full-text review and duplicate removal. Similar  
1388 comprehensive search strategies were applied to IEEE Xplore, ScienceDirect, Taylor & Francis Online, Wiley Online  
1389 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%
<b>Total</b>	<b>350</b>	<b>100%</b>

**C Corpus Creation Details**

**Datasheet Fields.**

Table 7. Description of datasheet fields

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

Column	Content
1561	
1562	
1563	<b>Pipeline Stage</b> Specific process within a pipeline stage (controlled list):
1564	• Problem Understanding and Formulation
1565	• ML System Design and Development
1566	• Deployment and Impact
1567	• Cross-pipeline
1568	
1569	
1570	
1571	<b>Pipeline Sub-stage</b> Specific process within a pipeline stage (controlled list):
1572	• Institutional Prioritization and Funding
1573	• Product Conception and Design
1574	• Data Selection, Collection and Annotation
1575	• Model Architecture Selection and Design
1576	• Model Training and Evaluation
1577	• Product Testing
1578	• Product Launch
1579	• Cross-pipeline
1580	
1581	
1582	
1583	<b>Historical Era</b> Era of data production practice (controlled list):
1584	
1585	• Era 1: Curated datasets (pre-2009); no sources in corpus
1586	• Era 2: Crowdsourced benchmarks (2009–2017)
1587	• Era 3: Web-scraped/foundation models (2017–present)
1588	• Multi-era: Spans multiple eras or provides historical analysis
1589	
1590	
1591	<b>Primary Pattern(s) / Pathway(s)</b> 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).
1592	
1593	
1594	
1595	
1596	
1597	<b>Triangle Coverage</b> Engagement with the three scoping domains that defined corpus eligibility:
1598	
1599	• A – AI contexts
1600	• D – Data production practices
1601	• C – Community impacts
1602	
1603	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.
1604	
1605	
1606	
1607	
1608	An accompanying Rationale column explains the basis for inclusion and the specific ways each source engages A/D/C beyond passing mention.
1609	
1610	
1611	
1612	

Column	Content
<b>How Source Was Found</b>	White literature (journal papers, conference proceedings, books) or Grey literature (reports, policy documents, theses, community outputs, blogs).
<b>Keywords</b>	3–5 terms for coding/search, ordered Geography → Data/technical → Community/impact.
<b>Geographic Region of Focus</b>	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).
<b>Author Affiliations</b>	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
<b>Geographic Area of Author(s)</b>	Full institution name and country of the lead author(s)
<b>Institution</b>	Region of lead author's institution (controlled list, same regions as above).
<b>Authorship and Positionality Context</b>	Complete authorship profile, including all institutions, geographic distribution, equal contribution notes, and any relevant statements on positionality or disciplinary traditions.
<b>Summary</b>	≤ 120-word synopsis. Structure: Topic → Method → Findings → Link to AI data production + community impacts.

**1665 Corpus Summary.**

**1666**  
**1667** Table 8. Corpus composition summary (N=350 sources)

<b>1669 Category</b>	<b>Sub-category</b>	<b>Count (%)</b>
<b>1670 Orientation</b>	Extractive Practices	141 (41%)
	High-Agency Principles	116 (33%)
	High-Agency Practices	93 (27%)
<b>1674 Source Type</b>	White literature	258 (74%)
	Grey literature	92 (26%)
<b>1676 Geographic Focus</b>	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%)
	MENA	3 (1%)
<b>1686 Author Affiliation (lead only)</b>	Academic	182 (52%)
	Mixed	96 (27%)
	Industry	37 (11%)
	NGO/Non-profit	20 (6%)
	Journalist/Other	11 (3%)
<b>1692 Pipeline Stage</b>	Government	4 (1%)
	ML System Design & Development	183 (52%)
	Problem Understanding & Formulation	90 (25%)
	Cross-pipeline	55 (16%)
	Deployment & Impact	22 (6%)
<b>1698 Historical Era</b>	Era 3 (2017–present)	241 (69%)
	Multi-era	95 (27%)
	Era 2 (2009–2017)	14 (4%)
	Era 1 (pre-2009)	0 (0%)

**1703** Received 12 February 2026; revised 5 June 2009; accepted 5 June 2009