A Systematic Review of Crowdsourcing Approaches and Their Applicability to Curation

MASTER THESIS

Mohammad Naim

Submitted on 9 March 2026

Friedrich-Alexander-Universität Erlangen-Nürnberg
Faculty of Engineering, Department Computer Science
Professorship for Open Source Software

Supervisor:
Martin Wagner
Prof. Dr. Dirk Riehle, M.B.A.

Friedrich-Alexander-Universität
Faculty of Engineering

Declaration of Originality

I confirm that the submitted thesis is original work and was written by me. Appropriate credit has been given where reference has been made to the work of others. The thesis was not examined before, nor has it been published. The submitted electronic version of the thesis matches the printed version.

Artificial Intelligence (AI) tools ChatGPT (OpenAI) and Gemini (Google) were used for language editing, source summarization, and code development assistance. All AI-generated content was reviewed, verified, and edited by the author. The ideas, analysis, conclusions, and final interpretations presented in this thesis remain entirely my own work.

Erlangen, 9 March 2026

License

Erlangen, 9 March 2026

Abstract

The increasing complexity of software ecosystems requires reliable, continuously maintained Software Composition Analysis (SCA) data. Although automated scanners provide essential scalability, they lack the contextual judgment required to reliably resolve ambiguous licensing and vulnerability scenarios. This raises the question of whether crowdsourcing can support data curation within SCA environments, particularly in the context of the SCA Tool.

To address this question, this thesis conducts a Systematic Literature Review (SLR) following the evidence-based software engineering guidelines proposed by Kitchenham et al. (2004) and Kitchenham, Charters et al. (2007). The review synthesizes existing research on crowdsourcing approaches, application domains, incentive structures, quality assurance mechanisms, and licensing frameworks. The analysis shows that crowdsourcing encompasses structurally distinct approaches that differ in task complexity, aggregation mechanisms, and contribution diversity.

Based on this synthesis, the study evaluates the suitability of these approaches for supporting data curation tasks within the SCA Tool. The findings suggest that selectively aligning crowdsourcing approaches with task characteristics and associated risk levels can improve scalability while preserving data integrity. However, its effectiveness depends on strong governance, appropriate incentives, and proportional quality controls. Overall, the study provides an evidence-based foundation for assessing whether crowdsourcing can complement automated workflows in SCA-based data curation.

Contents

List of Figures

List of Tables

Acronyms

SCA Software Composition Analysis

SBOM Software Bill of Materials

SLR Systematic Literature Review

GPS Global Positioning System

SDT Self-Determination Theory

GWAP Games With A Purpose

NDA Non-Disclosure Agreement

UX User Experience

RQ Research Question

IP Intellectual Property

1 Introduction

Crowdsourcing has emerged as a widely adopted paradigm for organizing distributed human effort in the digital era. By enabling open participation in problem-solving, content creation, and data processing, it has reshaped organizational approaches to innovation and large-scale information management. Across domains including software engineering, healthcare, machine learning, and education, crowdsourcing is increasingly used to complement or extend traditional organizational workflows (Bhatti et al., 2020; Bhuyan & Singh, 2023; Wang et al., 2020).

This chapter introduces the conceptual foundations of crowdsourcing, outlines its historical development, and presents the research problem addressed in this thesis.

1.1 Foundations and Conceptual Development of Crowdsourcing

Crowdsourcing refers to the delegation of tasks, traditionally performed within organizations, to a large and undefined group of individuals through an open call (Bhuyan & Singh, 2023; Ghezzi et al., 2018; Neto & Santos, 2018). Rather than relying solely on internal employees or contracted experts, organizations leverage the collective intelligence and distributed capabilities of the "crowd". This method enables scalable collaboration and provides access to distributed expertise and perspectives, supporting open innovation as well as potential gains in cost and time efficiency beyond traditional organizational structures (Alenezi & Faisal, 2020).

The term "crowdsourcing" was formally introduced by Jeff Howe and Mark Robinson in 2006 in the article "The Rise of Crowdsourcing" (Howe et al., 2006). Their definition emphasized the power of the open-call structure and the accessibility of a vast, diverse network of potential contributors. Crowdsourcing has subsequently become a central model for distributed problem-solving and innovation management across academic research and industrial applications.

Fundamentally, crowdsourcing operates on the principle that the aggregated contributions of many individuals frequently surpass the outcomes generated by a single expert. However, its effectiveness depends on more than just participation volume; it requires deliberate task design, aggregation logic, incentive structures, and governance mechanisms (Blohm et al., 2018; Cappa et al., 2019). Modern platforms increasingly integrate these elements, transforming crowdsourcing from isolated contest-based initiatives into structured digital ecosystems (Ghezzi et al., 2018).

1.2 Evolution of Crowdsourcing: Historical Roots and Modern Expansion

Although the term "crowdsourcing" is relatively recent, the underlying concept of mobilizing collective human effort through open participation has existed for centuries (Allon & Babich, 2020; Bhatti et al., 2020). Early innovation contests and collaborative public projects reveal the long-standing value of engaging the "crowd" to solve complex problems.

Several historical cases illustrate this progression:

- **The Longitude Prize (1714)**: The British government offered a substantial reward to solve the challenge of determining a ship's longitude at sea. John Harrison's marine chronometer eventually won the prize, providing an early and famous example of open innovation (Sobel, 2007).
- **The Food Preservation Contest (1810)**: In France, this competition led to Nicolas Appert's invention of food canning, demonstrating how public contests can generate practical industrial breakthroughs (Stol et al., 2017).
- **The Oxford English Dictionary (1879)**: This massive linguistic undertaking relied on thousands of voluntary contributions from the public to compile word usages and meanings (Bhuyan & Singh, 2023).
- **Galton's "Wisdom of the Crowd" (1907)**: Francis Galton's experiment demonstrated that the average of a crowd's estimates could approximate expert-level accuracy, providing empirical support for the statistical validity of aggregated intelligence (Galton, 1907).
- **The Sydney Opera House (1955)**: An international design competition exemplified the effectiveness of open calls in fostering creative problem-solving (Stol et al., 2017).

These initiatives share a common principle: distributing a problem to a broad population in order to collect solutions or relevant information. However, the systematic, large-scale expansion of crowdsourcing only became possible

with the advent of the Internet and Web 2.0. While the Internet provided global connectivity, Web 2.0 technologies facilitated the transition from passive consumption to active content co-creation (Allon & Babich, 2020; Bhuyan & Singh, 2023).

Further advancements in mobile technologies, Global Positioning System (GPS) systems, and smartphones have enabled real-time, location-based collaboration. These digital infrastructures have significantly reduced transaction costs and enhanced speed, accessibility, and scalability. These developments allow organizations to coordinate distributed contributors efficiently for complex or time-sensitive tasks (Allon & Babich, 2020).

Today, modern crowdsourcing platforms leverage distributed human computation to perform activities such as image labeling, translation, scientific data annotation, and software testing. In many contexts, crowdsourcing complements automated systems by providing human judgment where machine-based approaches remain limited. As a result, crowdsourcing has evolved into a structured framework for distributed problem-solving across multiple industries, including complex environments like software data curation (Bhatti et al., 2020; Bhuyan & Singh, 2023).

1.3 Problem Statement and Research Objective

The SCA Tool, developed at the Professorship for Open-Source Software at FAU Erlangen, supports SCA related tasks including Software Bill of Materials (SBOM) management, license compliance, vulnerability tracking, and governance of open-source components. These functions rely on accurate and continuously maintained data. As software ecosystems evolve, maintaining this data represents a persistent curation challenge.

While automated scanners perform large portions of this work, certain tasks require human judgment, contextual interpretation, and validation of ambiguous or conflicting results. This raises a central design question: whether crowdsourcing can serve as an effective mechanism for enhancing data curation within the SCA Tool.

Crowdsourcing approaches vary significantly in task structure, aggregation logic, incentive design, and quality control (Bhatti et al., 2020; Blohm et al., 2018). However, the evidence required to assess their suitability for SCA-based curation is dispersed across domain-specific studies. To date, no consolidated synthesis provides a structured evaluation of these approaches within the SCA context.

To address this gap, this thesis conducts a SLR following established evidence-based software engineering guidelines by Kitchenham et al. (2004) and Kitchenham,

Charters et al. (2007). This research aims to provide an evidence-based foundation for deciding whether crowdsourcing is a viable data curation strategy for the SCA Tool by:

- Identifying and classifying established crowdsourcing approaches and their operational characteristics;
- Analyzing incentive models, governance structures, and quality assurance mechanisms;
- Synthesizing reported strengths, limitations, and trade-offs; and
- Assessing the suitability of these findings for specific SCA-related curation tasks.

By integrating this dispersed knowledge into a unified analytical framework, this thesis establishes a rigorous basis for determining whether crowdsourcing should be adopted within the SCA Tool's ecosystem.

2 Related Work

This chapter reviews prior research on crowdsourcing to establish the theoretical foundation of the study. It begins by examining existing SLRs, analyzing their scope and limitations. It subsequently evaluates real-world initiatives that apply crowdsourcing to data curation, highlighting practical design challenges that affect adoption and sustainability.

2.1 Prior SLRs on Crowdsourcing

Crowdsourcing has developed into a mature research domain, and several SLRs have attempted to organize its expanding body of knowledge. However, most reviews primarily focus on general conceptual models or traditional software engineering tasks rather than SCA-specific data curation contexts.

Early large-scale reviews focused on conceptualizing and classifying crowdsourcing systems. For example, Hosseini et al. (2015) organized the literature around four structural pillars: the crowd, the crowdsourcer, the task, and the platform. Their objective was to provide a structured taxonomy that clarifies how crowdsourcing systems are designed and how responsibilities are distributed. While this taxonomy provides an important conceptual foundation for understanding crowdsourcing systems, it does not evaluate domain-specific applicability or operational performance in specialized contexts.

Within software engineering, several SLRs have explored how crowdsourcing supports software development processes. For instance, Sarı et al. (2019) structured their review around ten research questions addressing business models, technological platforms, development methodologies, effort estimation, incentive mechanisms, crowd formation strategies, and task decomposition. Their analysis emphasizes economic and organizational aspects of competitive programming environments. However, the focus remains on production-oriented tasks such as coding, testing, and requirements engineering. Similarly, Mao et al. (2017) mapped crowdsourcing applications across software lifecycle phases, identifying trends and coordination challenges in distributed development. These studies

primarily assess productivity and cost efficiency rather than data-centric validation processes.

Several additional reviews provide complementary perspectives. Ambreen and Ikram (2016) examined empirical trends in crowdsourcing research. Thuan et al. (2016) analyzed organizational decision factors influencing crowdsourcing adoption, highlighting task complexity and risk considerations. Morschheuser and Hamari (2019) investigated gamification mechanisms and participant motivation. Stol et al. (2017) discussed onboarding, coordination, and quality assurance challenges in software engineering contexts.

Despite these contributions to understanding crowdsourcing models, engagement strategies, and general applications, an important limitation remains: none systematically assess the applicability of crowdsourcing approaches to structured data curation in SCA environments. Curation tasks differ fundamentally from coding or testing activities, as they require metadata validation, interpretation of complex licensing conditions, consistency checks, and sustained quality assurance. This limitation motivates further investigation.

2.2 Crowdsourcing for Data Curation in Practice

Although academic literature rarely addresses crowdsourcing in SCA contexts, several practical initiatives have attempted to apply community-based approaches to software data curation. Two representative initiatives are ClearlyDefined and OSSelot.

2.2.1 ClearlyDefined

ClearlyDefined is an open initiative that aims to improve the clarity and accessibility of open-source licensing and metadata. The platform allows community members to review, correct, and enrich package information through collaborative workflows. Contributors can propose updates, discuss ambiguities, and refine information iteratively (ClearlyDefined Project, 2026).

The platform emphasizes openness and transparent collaboration. Its strength lies in collective refinement and distributed participation. However, participation depends largely on voluntary engagement, and incentive and governance mechanisms are not systematically structured. As a result, sustained participation and consistent quality control remain ongoing challenges.

2.2.2 OSSelot

OSSelot is another initiative focused on building a trusted database of curated software package information. Unlike open community models, OSSelot adopts a controlled participation approach. A limited group of experienced contributors performs continuous reviews under strict quality assurance processes (OSSelot Project, 2026).

This model prioritizes reliability and trustworthiness. However, restricting participation limits scalability and reduces access to diverse expertise. While quality is enhanced, throughput and community engagement remain constrained.

2.3 Adoption Challenges and Research Gap

Comparing ClearlyDefined and OSSelot reveals a core trade-off in crowdsourced data curation. ClearlyDefined follows an open participation model that encourages broad community involvement through voluntary contributions. However, the lack of structured incentives and governance mechanisms raises concerns about long-term sustainability and the reliability of contributions. In contrast, OSSelot relies on a limited number of experienced contributors to ensure high-quality outputs. While this controlled approach improves reliability, it limits scalability by restricting participation. Together, these initiatives illustrate a persistent trade-off between open and controlled environments, as well as between scalability and reliability. Neither initiative demonstrates a crowdsourcing model that effectively combines broad participation with structured governance and proportional quality assurance.

On the other hand, despite the expanding body of literature on crowdsourcing systems and their applications in software engineering, an important gap remains. Existing SLRs primarily focus on general conceptual frameworks, participation models, or software development activities such as coding and testing. However, none of these reviews systematically examine the applicability of crowdsourcing approaches to structured data curation in SCA environments.

Data curation in SCA systems involves tasks such as metadata validation, license interpretation, vulnerability assessment, and the continuous maintenance of software component datasets. These activities differ from traditional crowdsourced software development tasks because they require sustained quality management, contextual interpretation, and governance mechanisms. As a result, existing research does not provide a consolidated framework for evaluating how different crowdsourcing approaches align with the operational requirements of SCA-based curation workflows. This theoretical void often translates into significant adoption challenges during real-world implementation.

This thesis addresses this gap by systematically synthesizing crowdsourcing approaches and evaluating their applicability to SCA-related data curation, with particular reference to the SCA Tool. By integrating insights from existing literature with practical system design considerations, the study aims to provide an evidence-based foundation for assessing whether crowdsourcing can complement automated workflows in SCA-based data curation.

3 Research Methodology

The objective of this thesis is to evaluate the suitability of crowdsourcing as a mechanism for improving data curation within the SCA Tool. As research on crowdsourcing is dispersed across multiple domains and application contexts, a structured synthesis of the existing literature is required to establish a reliable analytical foundation.

To address this need, this study adopts a Systematic Literature Review (SLR) following the guidelines proposed by Kitchenham et al. (2004) and Kitchenham, Charters et al. (2007). An SLR systematically identifies, evaluates, and synthesizes relevant studies in order to provide a comprehensive and unbiased overview of the current state of research related to predefined research questions.

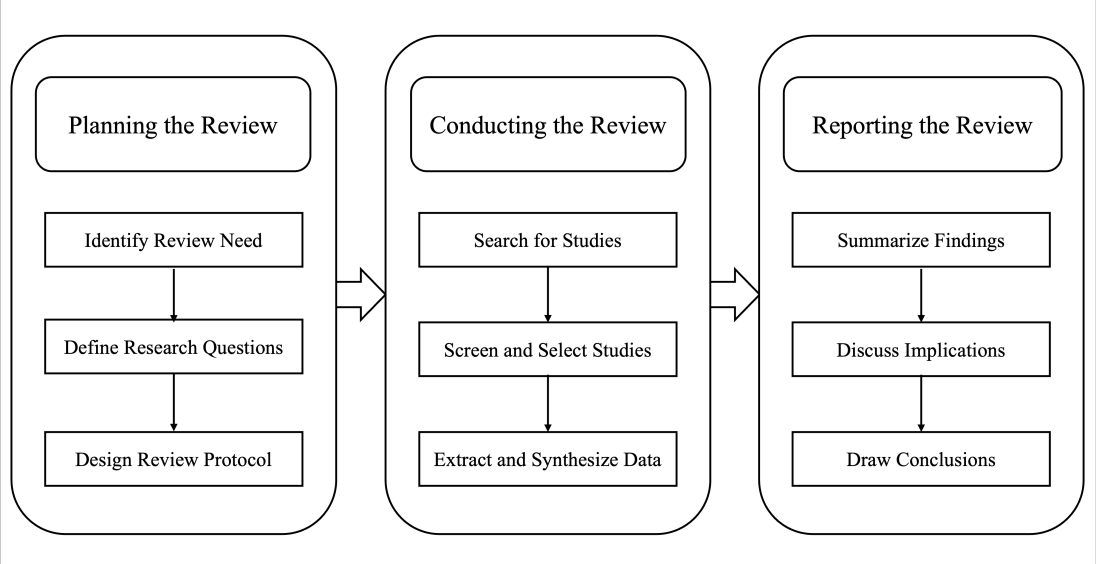

Figure 3.1: Systematic literature review process based on the guidelines of Kitchenham et al. (2004)

According to Kitchenham's framework, the SLR process comprises three main phases: planning, conducting, and reporting the review. Each phase consists of structured activities that guide the identification, selection, and analysis of relevant studies. Figure 3.1 illustrates the overall review process followed in this study.

The remainder of this chapter describes the review protocol, including the formulation of research questions, the search strategy, the study selection process, and the procedures for data extraction and synthesis.

3.1 Review Protocol and Planning

The planning phase establishes the methodological foundation for the SLR. In accordance with the guidelines by Kitchenham et al. (2004) and Kitchenham, Charters et al. (2007), a detailed review protocol was established prior to conducting the search. Defining the protocol in advance reduces the risk of researcher bias and ensures transparency, consistency, and reproducibility throughout the review process.

The protocol comprises the research questions, search strategy, selection criteria, and quality assessment procedures guiding the review. Each component is described in the following subsections.

3.1.1 Research Questions

The Research Questions (RQs) define the scope, structure, and analytical direction of this SLR. Since the objective of this thesis is to evaluate whether crowdsourcing is appropriate for enhancing data curation within the SCA Tool, the RQs are organized into two analytical phases: knowledge synthesis and suitability assessment.

Phase I: Knowledge Synthesis

The first phase of the study focuses on synthesizing existing knowledge from the reviewed literature. RQ1–RQ4 establish the analytical foundation required for evaluation by systematically identifying and structuring prior research on crowdsourcing approaches, real-world implementations, incentive and quality mechanisms, and the licensing frameworks governing crowdsourced outputs.

RQ1: What are the established crowdsourcing approaches, and how are they classified?

Goal: To develop a coherent conceptual framework that allows for the comparative analysis of different crowdsourcing models.

RQ2: In which domains have crowdsourcing methods been implemented, and how does their application vary across different operational contexts?

Goal: To analyze real-world implementations of crowdsourcing across diverse sectors to understand their scope and potential applicability to SCA environments.

RQ3: Which incentive and quality control strategies are most effective in ensuring participant reliability and the integrity of contributions in crowdsourcing systems?

Goal: To identify the motivational and validation frameworks required to sustain a high-integrity contributor community for data curation.

RQ4: How is crowdsourced data licensed, and what licensing categories are commonly used?

Goal: To investigate licensing practices and Intellectual Property (IP) considerations associated with crowdsourced data and outputs.

Phase II: Suitability Assessment

Building on the structured synthesis from Phase I, RQ5 and RQ6 evaluate the applicability of crowdsourcing approaches in the specific context of SCA-based data curation. These questions focus on suitability, feasibility, and associated trade-offs in relation to the operational requirements of the SCA Tool.

RQ5: Which crowdsourcing approaches are suitable for supporting data curation within the SCA Tool?

Goal: To assess the alignment between crowdsourcing models and the operational requirements of SCA-related curation workflows.

RQ6: What are the primary benefits, limitations, and ethical considerations associated with applying crowdsourcing to data curation within the SCA Tool?

Goal: To evaluate the strategic trade-offs and ethical implications of integrating crowdsourced workflows into a SCA environment.

ID	Research Question
Phase I: Knowledge Synthesis	
RQ1	What are the established crowdsourcing approaches, and how are they classified?
RQ2	In which domains have crowdsourcing methods been implemented, and how does their application vary across different operational contexts?
RQ3	Which incentive and quality control strategies are most effective in ensuring participant reliability and the integrity of contributions in crowdsourcing systems?
RQ4	How is crowdsourced data licensed, and what licensing categories are commonly used?
Phase II: Suitability Assessment	
RQ5	Which crowdsourcing approaches are suitable for supporting data curation within the SCA Tool?
RQ6	What are the primary benefits, limitations, and ethical considerations associated with applying crowdsourcing to data curation within the SCA Tool?

Table 3.1: Research Questions

In line with this two-phase structure, RQ1-RQ4 are addressed in Chapter 4, where the synthesized evidence from the literature are presented in a structured and descriptive manner. On the other hand, RQ5 and RQ6 are analyzed in Chapter 5, where the synthesized findings are applied to evaluate the suitability of crowdsourcing for data curation within the SCA Tool. Table 3.1 summarizes the RQs.

3.1.2 Search Strategy

To address the RQs defined in the previous section, a systematic search strategy was implemented to identify relevant studies on crowdsourcing approaches and their applicability to data curation and metadata correction. The aim is to ensure comprehensive coverage of both conceptual frameworks and real-world implementations.

Google Scholar was used as the primary database due to its broad interdisciplinary coverage. Two search strings were developed using Boolean operators (`AND`, `OR`) and the truncation operator (`crowdsourc*`) to capture variations such as *crowdsourcing*, *crowdsourced*, and *crowdsource*. The search was limited to publications dated between January 2018 and August 2025.

String 1: Focus on Curation and Metadata Correction
Designed to retrieve studies that explicitly connect crowdsourcing with data curation activities.

```
crowdsourc* AND ("curation" OR "data curation" OR "metadata correction")
```

String 2: Focus on Approaches, Domains and Use Cases
Designed to identify established crowdsourcing approaches, taxonomies, and applications across different domains.

```
crowdsourc* AND ("approaches" OR "domain" OR "application" OR "use case" OR "context")
```

3.1.3 Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected literature, a set of predefined inclusion and exclusion criteria was applied during the study screening process.

Inclusion Criteria

1. Peer-reviewed journal articles, conference papers, or book chapters.
2. Studies addressing crowdsourcing approaches, applications, or data curation.
3. Studies published between January 2018 and August 2025.

Exclusion Criteria

1. Grey literature and non–peer-reviewed sources, unless considered highly relevant and supported by credible references.
2. Studies that fall outside the scope defined by the RQs or the specified time-frame.
3. Publications written in languages other than English.

3.1.4 Quality Assessment

Following Kitchenham's guidelines, a quality assessment was conducted to evaluate the relevance and reliability of the included studies. Each study was examined for the clarity and completeness of its methodological description, particularly regarding crowdsourcing approaches, application contexts, reported benefits and limitations, and licensing practices related to data curation. Studies that lacked sufficient methodological detail or showed limited relevance to the RQs were excluded during the full-text screening stage. This process ensured that the final synthesis was based on credible and well-documented evidence.

3.2 Study Identification and Selection

Following the establishment of the review protocol, the next stage involved the systematic identification and selection of relevant studies. This phase implemented the predefined search strategy and selection criteria.

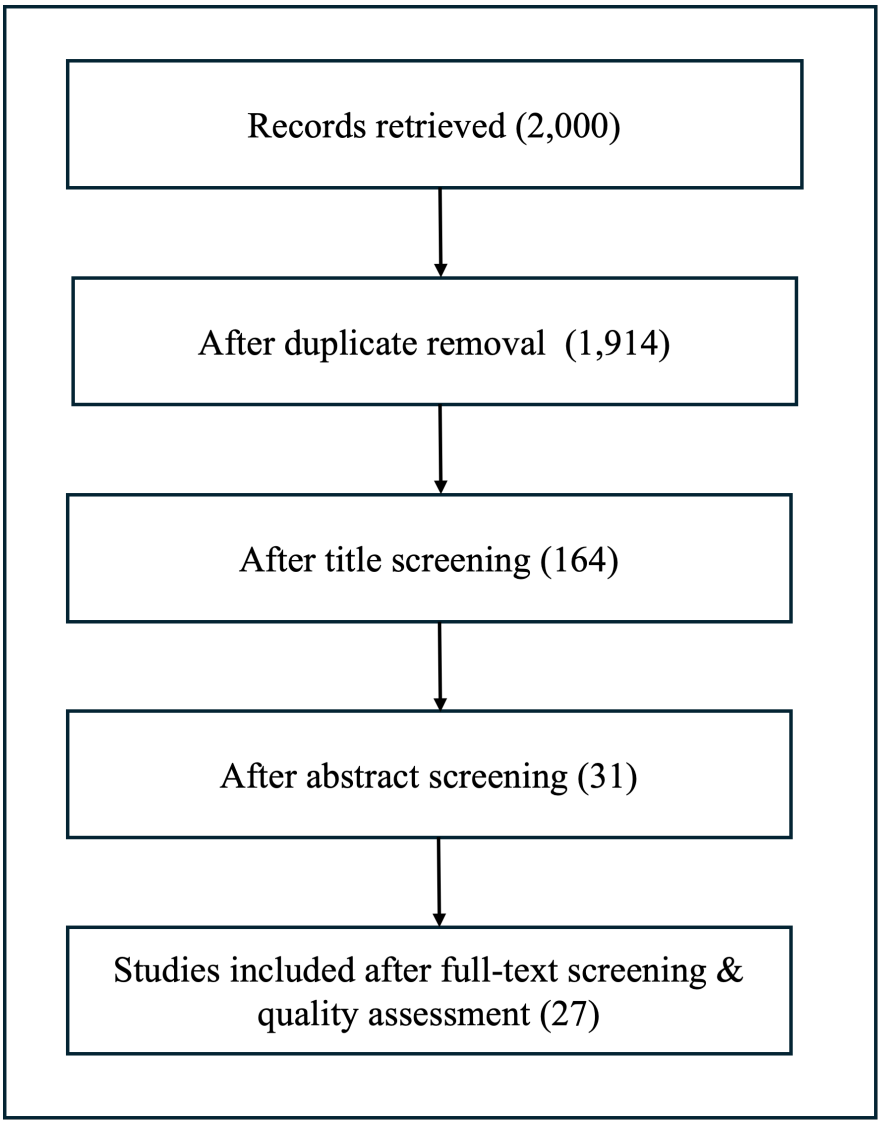

Figure 3.2: Study identification and screening process for the SLR, illustrating the flow of record identification, exclusion, and final inclusion.

The search queries were executed using the Publish or Perish software, which served as the primary tool for accessing and retrieving publications from the Google Scholar database. To maintain consistency with the review protocol, the search was restricted to the specified time-frame and executed using the predefined search strings without modification.

All retrieved records were imported into Zotero for bibliographic management. Zotero was used to organize references, remove duplicates, and support the subsequent screening process. Bibliographic metadata, including titles, authors, publication years, and abstracts, were systematically recorded and reviewed.

The screening process was conducted in sequential stages. First, titles were reviewed to assess their relevance to the research focus on crowdsourcing approaches and data curation. Studies that were clearly unrelated were excluded at this stage. Second, abstracts of the remaining studies were reviewed to evaluate their scope, methodological relevance, and alignment with the inclusion criteria. Only studies meeting these criteria were retained for full-text assessment.

During the full-text screening, studies were further evaluated in accordance with the predefined quality assessment criteria. Publications lacking sufficient methodological clarity or demonstrating limited relevance to the RQs were excluded.

The final set of eligible studies consists of 27 publications, which form the analytical basis of this review. The complete study selection process, detailing the number of records identified, screened, and excluded at each methodological phase, is visualized in Figure 3.2. Furthermore, the complete list of included studies is provided in Appendix A.

3.3 Data Extraction and Synthesis

Following the identification of eligible studies, data extraction and synthesis were conducted to analyze the final dataset. This phase aimed to organize relevant information systematically and derive cross-study insights aligned with the RQs.

The full texts of the included studies were imported into QDAcity, a qualitative data analysis tool supporting structured coding and traceability. Each study was examined in detail, and relevant text segments were coded according to categories derived from the RQs. These categories included crowdsourcing approaches, application domains, incentive and quality mechanisms, licensing models, and reported benefits and challenges.

The coding process was iterative. Initial codes were assigned based on predefined analytical categories, with iterative refinements applied as emergent patterns were identified. Previously coded studies were revisited where necessary to ensure consistency across the dataset. QDAcity's linking functionality enabled each coded segment to remain connected to its original source context, thereby maintaining transparency and auditability.

Subsequently, the coded data were synthesized using a descriptive and thematic approach. Codes were grouped into broader themes to identify recurring patterns, similarities, and distinctions across studies. This process enabled the classification of crowdsourcing approaches, comparison of application contexts, and identification of common incentive structures, licensing practices, and operational challenges.

The synthesis focused on conceptual integration and cross-study interpretation. The results of the thematic structure form the analytical foundation for the findings presented in Chapter 4 and the applicability assessment conducted in Chapter 5.

4 Research Results

This chapter presents the structured synthesis of the 27 studies included in the SLR. The findings are organized in alignment with RQ1–RQ4, which constitute the knowledge synthesis phase of this study. The primary focus of these RQs is on crowdsourcing approaches, application domains, incentive and quality mechanisms, and licensing practices associated with crowdsourced data. Additionally, this chapter synthesizes the benefits, challenges, and ethical concerns of crowdsourcing reported in the literature.

The following sections synthesize the reviewed evidence into a coherent framework that provides the foundation for the application-oriented analysis presented in Chapter 5.

4.1 Classification of Crowdsourcing Approaches

The analysis of the included studies shows that crowdsourcing approaches can be classified along several complementary dimensions. The literature commonly distinguishes these approaches based on contribution diversity, aggregation mechanisms, task complexity, and the type of value generated by the crowd. The following subsections outline these classifications.

4.1.1 Classification by Contribution Diversity and Result Aggregation

Blohm et al. (2018) provide a classification of crowdsourcing approaches based on two key dimensions:

I. **Contribution Diversity:** referring to whether participants provide homogeneous contributions where inputs are similar and comparable, or heterogeneous contributions where individuals submit diverse ideas, solutions, or perspectives;

II. **Aggregation Mechanism:** referring to whether value is created by selecting the best individual contribution from the crowd or by integrating

multiple contributions into a combined result.

Combining these dimensions yields four distinct crowdsourcing approaches: microtasking, information pooling, broadcast search, and open collaboration. Each approach is briefly introduced in the following subsections to highlight its core characteristics and typical use cases.

Microtasking

Microtasking involves decomposing tasks into small, repetitive units, with each participant performing the same type of action. Contributions are homogeneous and highly standardized, and value is generated through selective aggregation, where individual submissions are evaluated independently. This approach is well suited to repetitive and structured tasks that require human intelligence but can be executed in parallel at large scale.

Typical applications include image labelling, data classification, and basic annotation tasks. Platforms such as Amazon Mechanical Turk, Galaxy Zoo, and Facebook Translations adopt microtasking to accelerate large-scale data processing (Blohm et al., 2018).

Information Pooling

Information pooling also relies on homogeneous contributions; however, individual inputs have limited standalone value. Meaningful insights emerge only after contributions are integrated using aggregation techniques such as averaging, voting, or statistical modelling.

Examples include prediction markets such as the Hollywood Stock Exchange and sensing-based platforms such as AT&T's Mark the Spot and Google Maps traffic reporting, where collective inputs are combined to generate real-time or predictive information (Blohm et al., 2018).

Broadcast Search

Broadcast search is characterised by heterogeneous contributions, where participants submit diverse ideas, solutions, or models. Value is created through selective aggregation, as only the most promising or high-quality submissions are retained. This approach is particularly effective for innovation contests and complex problem-solving tasks that require creativity or specialised expertise.

Examples include the Netflix Prize, GE Ecomagination Challenge, InnoCentive, and Applause, where a small number of outstanding solutions are selected from a diverse pool of submissions (Blohm et al., 2018).

Open Collaboration

Open collaboration involves heterogeneous contributions that become valuable only through integration and iterative refinement. Individual contributions are often partial; however, continuous interaction and co-creation facilitate the development of coherent and high-quality outcomes.

Common examples include Wikipedia, OpenIDEO, LEGO Ideas, and open-source software communities, where contributors collectively develop content, designs, or codebases over time (Blohm et al., 2018).

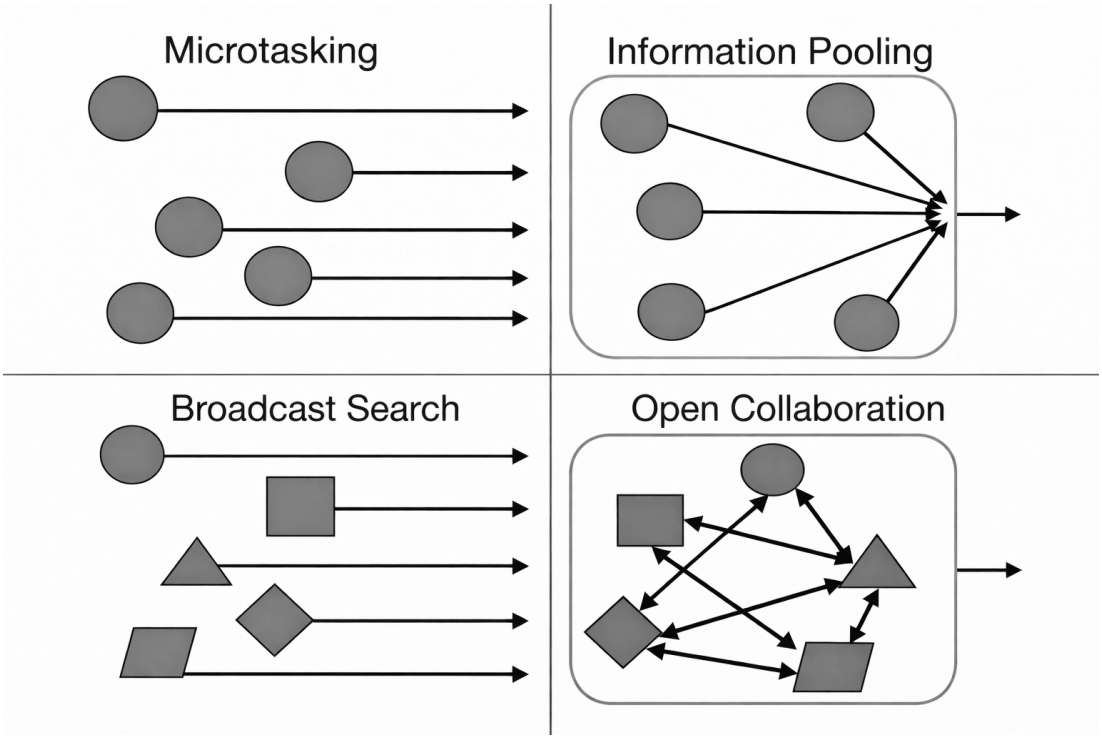

Figure 4.1: Classification of crowdsourcing approaches by contribution diversity and result aggregation (Blohm et al., 2018)

An overview of how different crowdsourcing approaches emerge from the interaction between contribution diversity and aggregation logic, as proposed by Blohm et al. (2018), is shown in Figure 4.1.

4.1.2 Classification by Task Complexity

In addition to contribution and aggregation based distinctions, crowdsourcing approaches can also be classified according to task complexity. Bhatti et al. (2020) propose a task-based classification that distinguishes crowdsourcing approaches according to the required level of cognitive effort, expertise, and task decomposability. Under this framework, tasks are categorized into micro-tasks, complex tasks, macro-tasks, and creative tasks. Each category is briefly discussed in the following subsections to clarify its defining characteristics and typical applications.

Micro-tasks

Conceptually parallel to the microtasking category of Blohm et al. (2018), micro-tasks are simple, well-defined, and highly decomposable units of work that can be completed independently using general skills. They support large-scale parallel execution and typically rely on aggregation to ensure accuracy.

Common examples include image labeling, transcription, and rating tasks, often implemented on platforms such as Amazon Mechanical Turk, FigureEight, microTask, and MicroWorkers (Bhatti et al., 2020).

Complex Tasks

Complex tasks require higher cognitive effort and domain-specific knowledge than micro-tasks but remain decomposable into smaller sub-tasks. These tasks are commonly executed through structured workflows and require greater coordination and compensation.

Examples include article writing, large-scale proofreading, data generation, and natural language processing tasks. Platforms such as PlateMate, Turkomatic, TurKit, and DataSift support this approach through task decomposition and recomposition mechanisms (Bhatti et al., 2020).

Macro-tasks

Macro-tasks are non-decomposable complex tasks that cannot be effectively divided without losing essential context. They require expert-level knowledge, sustained cognitive effort, and often collaborative or iterative participation. Macro-tasks are commonly associated with expert crowdsourcing and are generally less scalable than micro-tasks or complex tasks due to their higher complexity and limited decomposability.

Typical examples include drafting technical documents, defining research methodologies, and solving complex R&D problems. Platforms such as Upwork,

CrowdSpring, and OpenIDEO are frequently used to support macro-task crowdsourcing (Bhatti et al., 2020).

Creative Tasks

Creative tasks focus on ideation, originality, and innovation rather than predefined solutions. These tasks are open-ended and depend heavily on individual creativity and unique perspectives. Contributions are typically evaluated through selective mechanisms rather than aggregation, where only the most promising or high-quality submissions are chosen.

Common applications include design competitions, software innovation challenges, and product or logo design. Representative platforms include TopCoder, InnoCentive, 99designs, and Threadless, where contributors compete or collaborate to generate novel solutions (Bhatti et al., 2020).

Task Type	**Complexity Level**	**Decomposability**	**Typical Participation**
Micro-Tasks	Low	Decomposable	Individual (Collective)
Complex Tasks	Medium	Decomposable	Individual (Aggregative)
Macro-Tasks	High	Non-decomposable	Collaborative (Iterative)
Creative Tasks	High	Often Non-decomposable	Competitive (Selective)

Table 4.1: Classification of Crowdsourcing Approaches by Task Complexity

An overview of the classification of crowdsourcing approaches based on task complexity proposed by Bhatti et al. (2020), highlighting differences in complexity level, task decomposability, and typical participation modes, is presented in Table 4.1.

4.1.3 Other Specific Types of Crowdsourcing

Beyond classifications based on contribution diversity, aggregation logic, and task complexity, the literature also identifies crowdsourcing approaches based on the type of value generated by the crowd. Bhuyan and Singh (2023) categorize these approaches according to the nature of crowd contributions across activities such as funding, software evaluation, creative work, collaborative knowledge creation, and problem-solving. Two representative examples are briefly discussed below to illustrate their key characteristics and practical relevance.

Crowdfunding

Crowdfunding is a specialised form of crowdsourcing in which the primary contribution from participants is financial rather than intellectual or physical. It involves collecting small monetary contributions from a large number of individuals, typically via online platforms, to support projects, products, or initiatives. It enables entrepreneurs, non-profit organisations, and creators to raise monetary contributions by presenting their projects to a wide audience.

Although crowdfunding differs from labour or expertise based crowdsourcing, its open participation structure and reliance on digital platforms place it within the broader crowdsourcing domain, while its focus on financing distinguishes it as a separate category (Bhuyan & Singh, 2023).

Crowd Testing

Crowd testing applies crowdsourcing principles to software quality assurance by distributing testing tasks to a diverse pool of external contributors working in real-world conditions. By using a wide range of devices, operating systems, and usage contexts, this approach helps uncover defects and usability issues that are often missed in controlled, in-house testing environments.

Typical activities include exploratory testing, usability checks, bug identification, and feedback on User Experience (UX). The primary strength of crowd testing lies in its scalability and diversity, as parallel testing across diverse users increases coverage and the likelihood of uncovering edge cases or unexpected behaviors (Bhuyan & Singh, 2023).

These complementary dimensions, including contribution diversity, aggregation logic, task complexity, and contribution type, provide the structural foundation necessary to evaluate which approaches align with the specific data curation requirements of the SCA Tool in the next chapter.

4.2 Real-world Applications of Crowdsourcing

This section examines how crowdsourcing approaches are applied in real-world systems to illustrate their practical characteristics and outcomes. The examples illustrate how different crowdsourcing approaches organize participation, task design, and aggregation mechanisms in practice. They also provide concrete insights into how these models operate at scale and support different forms of collaborative work.

4.2.1 Wikipedia

Wikipedia is one of the most widely cited and well-established examples of large-scale open collaboration. It shows how distributed volunteers can collectively create, maintain, and refine a shared knowledge resource through continuous and incremental contributions. The platform relies on heterogeneous contributions, open participation, and iterative peer review. Content quality emerges from the aggregation and revision of many small edits rather than from centralized control (Blohm et al., 2018).

Participation in Wikipedia is voluntary and non-monetary, and is primarily driven by intrinsic motivations such as knowledge sharing, community recognition, and contributing to a public good. Beyond encyclopedia content, Wikipedia also illustrates broader principles of peer production, including self-organization, transparency, and community-based governance. Prior studies have shown that this collaborative model can achieve accuracy levels comparable to expert-edited reference works such as the Encyclopedia Britannica, highlighting the viability of large-scale, non-monetary collaboration for producing high-quality informational resources (Giles, 2005; Malone et al., 2010; Uhlmann et al., 2019).

4.2.2 Google Maps

Google Maps illustrates the information pooling approach, in which a large number of users contribute simple and mostly homogeneous inputs. These inputs are aggregated to create and maintain a shared information resource. Users can update place details, upload photos or videos, correct inaccuracies, and provide ratings or comments. These contributions complement automated data collection and support the continuous improvement of geographic information across a wide range of locations (Goodchild, 2007).

Beyond explicit user contributions, the platform also utilizes passively collected data such as GPS signals from mobile devices to determine traffic conditions, congestion patterns, and average travel speeds. Active reports of incidents, road closures, or delays further enhance situational awareness. The effectiveness of this information pooling model depends on sustained user participation and the rapid

aggregation of distributed observations. This allows the system to provide reliable real-time insights and accurate routing recommendations globally (Haklay, 2016).

4.2.3 Zooniverse

Zooniverse, which hosts well-known projects such as Galaxy Zoo, is a prominent example of microtasking within the broader domain of citizen science[1]. It demonstrates how large-scale volunteer participation can be organized around small, well-defined tasks such as image classification, pattern recognition, and short text transcription. By decomposing complex scientific problems into independent microtasks, the platform allows large datasets to be processed efficiently in parallel by geographically distributed contributors (Lintott et al., 2008).

Zooniverse relies on structured task workflows and aggregation mechanisms to ensure reliability. It also combines lightweight gamification elements to sustain participant engagement. Instead of relying on a single expert, data quality emerges from many independent contributions, where repeated inputs and consensus among contributors help validate the results. Empirical studies show that the aggregated outputs produced by citizen scientists can achieve accuracy levels comparable to expert analysis. These findings demonstrate that microtasking-based crowdsourcing can provide a robust and scalable approach for curating, validating, and processing large scientific datasets for downstream analysis. (Simpson et al., 2014; Uhlmann et al., 2019).

4.2.4 Comparison of Selected Crowdsourcing Systems

Wikipedia, Google Maps, and Zooniverse illustrate how different platforms adopt different crowdsourcing approaches according to their primary objectives. As discussed in the previous sections, crowdsourcing approaches can be distinguished by key dimensions such as contribution diversity, aggregation mechanisms, and task complexity. The design of each platform reflects how these dimensions are combined to support its specific goals. Examining these systems helps to clarify why different crowdsourcing models are suitable for different types of tasks and organizational objectives. A summary of the key characteristics of these platforms across these dimensions is presented in Table 4.2.

Wikipedia primarily aims to support collaborative knowledge creation and maintenance. To achieve this objective, it adopts an open collaboration model that allows contributors to freely create, edit, and refine content. The tasks involved are often complex and creative, with medium to high levels of task complexity. In contrast, Google Maps focuses on maintaining accurate and up-to-date geographic

[1] Citizen science is a form of crowdsourcing applied to scientific research, where the public actively participates in tasks such as data collection and analysis to support scientific goals (Vohland et al., 2021).

information. It therefore relies on an information pooling approach, where users contribute relatively simple and homogeneous inputs. These tasks typically have low complexity and can be aggregated to improve the overall accuracy of the platform. Zooniverse represents a microtasking model used in citizen science projects. Although the underlying scientific problems are complex, they are decomposed into small micro-tasks that volunteers can complete independently, which keeps the task complexity low at the contributor level.

Dimension	**Wikipedia**	**Google Maps**	**Zooniverse**
Crowdsourcing Approach	Open Collaboration	Information pooling	Microtasking
Diversity of Contributions	Heterogeneous	Homogeneous	Homogeneous
Aggregation of Contributions	Integrative	Integrative	Selective
Task Complexity	Medium to High	Low	Low
Primary Goal	Collaborative knowledge creation and maintenance	Up-to-date geospatial and place information	Large-scale scientific data annotation

Table 4.2: Comparison of Wikipedia, Google Maps, and Zooniverse Across Key Crowdsourcing Dimensions

This comparison shows that the choice of a crowdsourcing approach depends on the primary objective of the platform and the characteristics of the tasks involved. The required level of contribution diversity, the aggregation mechanism used to combine results, and the complexity of tasks all influence which model is most appropriate. These dimensions also provide a useful framework for evaluating which crowdsourcing approach is most suitable for data curation in the SCA Tool. However, selecting an appropriate crowdsourcing model alone does not guarantee success. The effectiveness of these systems also depends on the incentive and motivational mechanisms that encourage participation, as well as the quality assurance mechanisms used to ensure reliable crowd contributions. These aspects are presented in the following section.

4.3 Incentive, Motivation & Quality Mechanisms in Crowdsourcing

Crowdsourcing systems depend on effective incentive and motivation mechanisms to encourage sustained participation and ensure high-quality contributions. These incentives are commonly classified as extrinsic and intrinsic, both of which aim to increase contributor engagement and performance. However, increased participation alone does not guarantee reliable results. Therefore, crowdsourcing platforms often combine incentive mechanisms with quality and performance mechanisms that evaluate contributor reliability and submission accuracy (Bhatti et al., 2020; Cappa et al., 2019).

This section first discusses extrinsic and intrinsic incentives and then examines the mechanisms used to maintain the quality of crowd contributions.

4.3.1 Extrinsic Incentives

Extrinsic incentives refer to external rewards offered in exchange for task completion, such as monetary payments, prizes, or other measurable benefits. These incentives are commonly used in commercial and task-oriented crowdsourcing platforms and are typically implemented through direct monetary compensation or competitive prize-based contests, which are briefly discussed in the following subsections.

Monetary Compensation

Direct payment is the most common form of extrinsic incentive, especially in microtask marketplaces such as Amazon Mechanical Turk (AMT). Contributors are typically compensated through fixed payments, where a predefined amount is paid upon successful task completion and acceptance. For monetary rewards to be effective, compensation must be perceived as fair, aligning with principles from fairness expectation theory. Payment levels are usually determined based on estimated task duration and complexity, often targeting an implied hourly rate comparable to minimum wage standards in microtasking contexts (Bhatti et al., 2020; Cappa et al., 2019).

However, existing research also highlights the crowd-out effect, where excessively high financial rewards may reduce overall participation. Large payments can discourage non-expert contributors, who may perceive such tasks as intended only for experts and specialists, resulting in fewer submissions overall. From the perspective of Self-Determination Theory (SDT), high monetary incentives may be perceived as controlling, thereby diminishing autonomy and reducing intrinsic engagement (Cappa et al., 2019).

Prizes and Contests

Prizes and contests are commonly used for complex, creative, or problem-solving tasks, particularly in broadcast search approaches such as idea generation or design competitions. These systems typically offer substantial rewards for winning submissions. To reduce the limitations of winner-takes-all dynamics and encourage broader participation, platforms may also provide additional incentives such as runner-up prizes, milestone-based rewards, or dynamic payments based on auction-like mechanisms. These strategies help to manage participant expectations and sustain engagement in competitive environments (Blohm et al., 2018; Cappa et al., 2019; Ghezzi et al., 2018).

4.3.2 Intrinsic Incentives

Intrinsic incentives arise from the satisfaction that participants experience from the activity itself. They are particularly important for sustaining long-term participation and are central to open collaboration models, where contributions are often voluntary and non-monetary (Blohm et al., 2018). These incentives are typically driven by non-monetary factors such as reputation and recognition, gamification and enjoyment, social interaction and collaboration, prosocial or altruistic motivation, and opportunities for learning and skill development, which are discussed in the following subsections.

Reputation and Recognition

Reputation and recognition are strong intrinsic motivators in crowdsourcing systems. Many platforms implement reputation systems that visualize contributor activity, experience, and perceived quality through points, badges, levels, or rankings. Such mechanisms allow contributors to signal expertise and social status within the community, satisfying psychological needs for recognition and achievement. In competitive software engineering contexts, reputation can also serve as a long-term incentive by providing access to more complex or higher-paid tasks (Blohm et al., 2018; Sarı et al., 2019).

Gamification and Enjoyment

Gamification motivates participation in repetitive or cognitively demanding tasks by introducing elements of entertainment and enjoyment. It introduces game design elements such as challenges, narratives, points, badges, and leaderboards into non-gaming contexts to enhance engagement and performance. By making participation more enjoyable, these mechanisms help to satisfy contributors' needs for competence and achievement. A well-known example is Games With A Purpose (GWAP), where users solve complex problems as a by-product of

gameplay, such as protein structure prediction in FoldIt (Morschheuser & Hamari, 2019).

Social Connection and Collaboration

Social connection and collaboration further contribute to intrinsic motivation. Many contributors are driven by the desire to belong to a community and collaborate with peers. Consequently, crowdsourcing platforms often provide socialization mechanisms such as forums, messaging systems, and discussion boards. These features allow contributors to exchange knowledge, seek support, resolve disputes, and recreate social interactions similar to those found in traditional workplace environments (Blohm et al., 2018).

Prosocial and Altruistic Motivation

Prosocial and altruistic motivations refer to the desire to help others or contribute to a meaningful social cause. In crowdsourcing, contributors may participate because they want to support goals such as environmental protection, disaster response, or medical research. Platforms that clearly communicate the social impact of contributions often attract more participants and higher levels of engagement. Research shows that highlighting the real-world impact of tasks can increase contributor motivation and participation (Blohm et al., 2018; Vaughan, 2018).

Learning and Skill Development

Learning and skill development represent another important intrinsic driver. Many contributors participate in crowdsourcing initiatives to acquire new knowledge, improve creativity, or develop professional skills. Platforms that provide tutorials, training materials, or opportunities for peer learning tend to boost sustained engagement and higher-quality contributions over time, supporting both individual development and overall crowd performance (Acar, 2019; Blohm et al., 2018).

4.3.3 Quality and Performance Mechanisms

To ensure that increased participation does not reduce output quality, crowdsourcing systems often combine incentive mechanisms with quality and performance control strategies. These mechanisms help to assess contributor reliability, evaluate the quality of submissions, and reduce bias or malicious behavior.

One common approach is the use of performance-based schemes, where rewards are linked to the quality of the submitted work. For example, platforms may use pay-per-bug models in software testing or offer bonus payments for meeting

predefined quality thresholds. These schemes encourage contributors to focus on accuracy and effort rather than simply completing tasks (Blohm et al., 2018; Vaughan, 2018).

Platforms also assess worker reliability through qualification tests that contributors must pass before accessing tasks. In addition, many systems monitor acceptance rates and past performance to identify reliable contributors. These measures help filter low-quality submissions and assign tasks to more dependable workers (Blohm et al., 2018).

Another widely used method is the inclusion of gold standard or ground truth questions with known correct answers. These questions are embedded within task sets and allow platforms to evaluate contributor accuracy and honesty before accepting or combining submissions (Neto & Santos, 2018).

For tasks that involve subjective judgments, platforms often use evaluation and aggregation mechanisms. One common method is majority voting, where the same task is assigned to several contributors and the consensus result is accepted. This approach helps reduce individual bias and improves the reliability of the results (Neto & Santos, 2018).

Finally, peer review mechanisms are often used in collaborative or educational crowdsourcing settings. In these systems, contributors review and evaluate each other's work. For example, CrowdSorcerer enables participants to assess peer submissions, creating a self-correcting environment that encourages accountability, learning, and sustained participation (Pirttinen, 2021).

The literature shows that incentive and quality assurance mechanisms cannot be designed in isolation; their effectiveness depends on task characteristics and platform governance. While sustainable crowdsourcing requires integrating these operational controls, their success is also legally bounded by data ownership and licensing. This separate structural challenge is explored in the following section.

4.4 Licensing of Crowdsourced Data

The reviewed literature shows that licensing practices in crowdsourcing systems vary widely and are often not explicitly defined. Instead of being treated as a core design aspect, licensing is frequently embedded within platform terms of service or project-specific agreements, which often favor the organization or seeker rather than individual contributors or solvers. The open-call nature of crowdsourcing, which invites participation from a broad and diverse group of contributors, further complicates traditional legal concepts such as authorship, ownership, and remuneration. As a result, determining who owns a crowdsourced idea or dataset, and under what conditions it can be reused, shared, or rewarded, becomes a complex legal challenge (Ghezzi et al., 2018; Standing & Standing, 2018).

Based on the reviewed sources, the licensing and management of data generated through crowdsourcing can be broadly classified into three main categories: ownership transfer and seeker-centric licensing, open licensing models based on Creative Commons and open-source frameworks, and contractual or confidentiality-based licensing. These categories are briefly discussed in the following subsections.

4.4.1 Ownership Transfer and Seeker-Centric Licensing

In many commercial and competitive crowdsourcing environments, the dominant licensing model is ownership transfer, where contributors assign full or partial IP rights to the seeker, typically the individual, organization, or platform that defines the task and requests the work. Once a submission is accepted, the organizing entity gains exclusive legal control over the resulting data, code, or invention. This seeker-centric approach is widely used in commercial microtasking and competitive platforms, including those supporting data labeling, survey-based research, and software engineering tasks (Standing & Standing, 2018).

Commercial platforms such as Topcoder standardize these ownership transfer mechanisms to enable the seamless integration of crowd-generated outputs into proprietary systems. However, the literature consistently highlights that this model creates an unequal power dynamic, as the legal interests of seekers are prioritized over those of individual contributors. Moreover, legal ambiguity may arise in cases of rejected submissions, where platform terms may still permit seekers to retain access to unselected ideas, raising concerns regarding uncompensated use and potential copyright ownership conflicts (Sheehan, 2018; Standing & Standing, 2018).

4.4.2 Open Licensing Models

Crowdsourcing initiatives aimed at scientific research, public knowledge creation, or community-driven collaboration often adopt open licensing models rather than proprietary ownership frameworks. In such contexts, crowd-generated outputs are typically released under Creative Commons or open-source licenses, which allow reuse, redistribution, and modification while usually requiring attribution to the original contributors.

In medical and scientific domains, open access is often emphasized to promote transparency, reproducibility, and collective learning. Similarly, open-source software projects use collaborative licensing frameworks that support continuous improvement by enabling contributors to iteratively build upon existing work. These open licensing models prioritize shared value creation and long-term accessibility over exclusive organizational control (Standing & Standing, 2018; Tucker et al., 2019).

4.4.3 Contractual and Confidentiality-Based Licensing

In certain crowdsourcing contexts including crowdtesting, enterprise software evaluation, and security-sensitive tasks, licensing arrangements are supplemented or preceded by contractual and confidentiality-based mechanisms. Contributors may be required to sign Non-Disclosure Agreements (NDAs) before accessing tasks or associated datasets, especially when organizations expose proprietary software, internal systems, or sensitive information for testing or debugging purposes.

These agreements primarily function to protect the organization's pre-existing IP and to prevent unauthorized disclosure of confidential information. While such contractual safeguards are necessary in high-risk environments, The reviewed studies indicate that they further reinforce asymmetries between organizations and contributors and may limit transparency regarding the subsequent use of crowd-generated outputs (Bhuyan & Singh, 2023; Kohler, 2018).

The reviewed literature shows that licensing should not be treated as a minor legal issue but as an important part of crowdsourcing system design. The choice between ownership transfer, open licensing frameworks, or contractual and confidentiality-based arrangements affects contributor rights, data reuse conditions, and the overall governance of the platform. Integrating licensing considerations into system design is therefore important for ensuring transparency, legal clarity, and long-term sustainability in crowdsourcing environments. These licensing choices also influence the benefits, challenges, and ethical implications of crowdsourcing, which are discussed in the following section.

4.5 Benefits, Challenges, and Ethical Concerns of Crowdsourcing

Crowdsourcing has been widely adopted across domains due to its potential to overcome limitations of traditional, centralized work models. The reviewed literature consistently highlights a set of recurring benefits, operational challenges, and ethical concerns that characterize crowdsourcing initiatives regardless of application context. This section synthesizes these aspects based on the findings reported in the included studies.

4.5.1 Major Benefits of Crowdsourcing

One of the most frequently cited benefits of crowdsourcing is scalability and speed. By distributing tasks among a large number of contributors, crowdsourcing enables parallel execution, allowing large volumes of work to be completed within short time-frames. This property makes crowdsourcing particularly suitable for tasks that require rapid turnaround or continuous updates (Bhatti et al., 2020; Chen et al., 2020).

Cost efficiency is another commonly reported advantage. Crowdsourcing allows organizations to access labor or expertise on demand without maintaining large, permanent teams. Tasks can be decomposed and outsourced flexibly, reducing fixed operational costs and enabling more efficient resource allocation (Alenezi & Faisal, 2020; Kohler, 2018).

The literature also emphasizes crowdsourcing's ability to provide access to human intelligence for tasks that are difficult to automate. Many activities require contextual understanding, subjective judgment, creativity, or domain knowledge that automated systems cannot reliably replicate. Crowdsourcing enables organizations to leverage these human capabilities at scale (Bhatti et al., 2020).

Finally, diversity of contributors is often cited as a strength. Aggregating inputs from individuals with different backgrounds and perspectives can improve robustness and reduce individual bias, particularly when appropriate validation and aggregation mechanisms are applied (Cappa et al., 2019; Wang et al., 2020).

4.5.2 Operational Challenges

Despite these benefits, the reviewed studies also report several challenges associated with crowdsourcing. Data quality is the most common concern. Contributors vary in expertise, motivation, and effort, which can lead to noisy, incomplete, or low-quality submissions. Therefore, effective quality assurance mechanisms are essential to ensure reliable outcomes (Bhatti et al., 2020; Neto & Santos, 2018).

Another major challenge lies in task design and decomposition. Tasks must be clearly specified and structured in a way that contributors can understand and execute consistently. Poorly designed tasks or ambiguous instructions often lead to inconsistent results and increased error rates (Bhatti et al., 2020).

Aggregation of contributions also presents difficulties, particularly for subjective or complex tasks. Simple aggregation techniques such as majority voting may be insufficient when contributors disagree or when expertise varies significantly. More advanced aggregation methods are often required to derive meaningful results from multiple inputs (Bhatti et al., 2020).

Lastly, recruiting and retaining contributors remains an ongoing challenge. Sustained participation is difficult to achieve, especially for tasks requiring specialized skills. Prior research highlights the importance of carefully designed incentive structures and participation models to sustain contributor engagement over time, as discussed in the previous section (Bhatti et al., 2020; Kong et al., 2019).

Dimension	**Key Factors**
Benefits	Scalability and speed
	Cost efficiency
	Access to human intelligence for complex tasks
	Diversity of contributions
Challenges	Data quality management (noise or spam)
	Complex task decomposition
	Aggregation of conflicting results
	Worker retention
Ethical Concerns	Low wages and potential exploitation
	Power asymmetry
	Data privacy risks
	IP issues

Table 4.3: Benefits, Challenges, and Ethical Concerns of Crowdsourcing in Data Curation

4.5.3 Ethical Concerns

Beyond operational issues, crowdsourcing raises several ethical concerns that are widely discussed in the literature. Fair compensation is a central issue, as many platforms offer payments that fall below minimum wage standards. This has led to concerns about labor exploitation and the long-term sustainability of crowdsourcing ecosystems (Bhatti et al., 2020; Bhuyan & Singh, 2023; Standing & Standing, 2018).

Power asymmetry between task requesters and contributors is another recurring concern. Requesters usually control task acceptance, evaluation, and payment, while contributors often have limited ways to contest decisions or resolve issues. This imbalance can lead to unfair treatment (Wang et al., 2020).

Privacy and data protection also pose ethical risks. Crowdsourcing tasks may expose contributors to sensitive data, while platforms often collect detailed personal and behavioral information about participants. These practices raise concerns about data misuse, surveillance, and insufficient transparency (Bhatti et al., 2020; Hosseini et al., 2019).

Finally, IP ownership is frequently unclear in crowdsourcing arrangements. Contributors are often required to transfer rights to their outputs without negotiation or explicit attribution, leading to ambiguity regarding ownership, reuse, and recognition of contributed work (Bhatti et al., 2020; Standing & Standing, 2018).

Taken together, crowdsourcing is a powerful yet complex paradigm. While it offers significant advantages in terms of scalability, flexibility, and access to diverse human capabilities, these benefits are accompanied by important operational and ethical challenges. Addressing these challenges is essential for the responsible and sustainable use of crowdsourcing across different application domains. Table 4.3 summarizes the key benefits, operational challenges, and ethical concerns identified in the reviewed literature.

The synthesized findings presented in this chapter provide a structured overview of crowdsourcing approaches, application domains, incentive mechanisms, licensing practices for crowdsourced data, and the associated benefits and challenges. Together, these findings establish the analytical foundation for evaluating the applicability of crowdsourcing to structured data curation tasks. Building on this synthesis, the following chapter examines how the identified approaches align with the operational requirements of the SCA Tool and discusses their implications for supporting data curation in SCA environments.

5 Discussion

This chapter interprets the findings in relation to RQ5 and RQ6, which constitute the suitability assessment phase of this study. Building on Chapter 4, it evaluates the applicability of established crowdsourcing approaches to data curation within the SCA Tool, considering their potential benefits as well as the associated operational and ethical challenges. In addition, the chapter discusses strategies for addressing participation, quality, and ethical challenges when applying crowdsourcing to structured data curation.

5.1 Reflection on Research Questions RQ1-RQ4

This section reflects on the findings presented in Chapter 4 in relation to RQ1–RQ4. It analyzes the results of the knowledge synthesis phase and discusses their implications for understanding the design and operation of crowdsourcing systems.

RQ1: What are the established crowdsourcing approaches, and how are they classified?

The findings presented in Section 4.1 show that crowdsourcing is not a single, uniform paradigm but a set of distinct approaches that differ in task complexity, contribution diversity, aggregation mechanisms, and the type of value generated by crowd contributions. The classification based on contribution diversity and aggregation mechanisms distinguishes microtasking, information pooling, broadcast search, and open collaboration as different models of value creation. Complementing this perspective, the task complexity classification differentiates micro-tasks, complex tasks, macro-tasks, and creative tasks according to their level of cognitive demand and degree of decomposability. In addition to these structural classifications, the literature also identifies specific forms of crowdsourcing that reflect particular application contexts, such as crowdfunding and crowd testing, which apply crowdsourcing principles to financial support and software quality assurance activities.

Taken together, these dimensions provide a broader analytical framework that clarifies how crowdsourcing systems are structured and why certain approaches

are more suitable for particular types of problems. The synthesis indicates that no single classification fully captures the diversity of crowdsourcing models. Instead, a multidimensional perspective is required to understand the structural characteristics and operational implications of different crowdsourcing approaches.

RQ2: In which domains have crowdsourcing methods been implemented, and how does their application vary across different operational contexts?

The real-world systems examined in Section 4.2 demonstrate that crowdsourcing has been applied across a range of domains, including knowledge production, geospatial information systems, and scientific research. Platforms such as Wikipedia, Google Maps, and Zooniverse illustrate how different crowdsourcing approaches are adopted according to the primary objectives of the platform. Wikipedia relies on open collaboration to support collaborative knowledge creation and long-term content maintenance. Google Maps uses an information pooling model to aggregate large volumes of simple user-generated updates and observations. Zooniverse applies a microtasking approach that decomposes complex scientific problems into small tasks that can be processed at scale by distributed contributors.

The comparison shows that the effectiveness of crowdsourcing depends largely on the alignment between platform objectives, task characteristics, and the chosen crowdsourcing approach. Factors such as task complexity, contribution diversity, and aggregation mechanisms influence how participation is organized and how contributions are integrated to produce reliable outcomes.

RQ3: Which incentive and quality control strategies are most effective in ensuring participant reliability and the integrity of contributions in crowdsourcing systems?

The analysis in Section 4.3 shows that participation in crowdsourcing systems is shaped by a combination of incentive structures, motivational factors, and explicit quality control mechanisms. The literature distinguishes between extrinsic incentives, such as monetary compensation and competitive prizes, and intrinsic incentives, including reputation and recognition, gamification and enjoyment, social interaction and collaboration, prosocial motivation, and opportunities for learning and skill development. While financial rewards can stimulate short-term participation, long-term engagement is often associated with intrinsic motivations and community-oriented governance structures.

At the same time, incentive mechanisms alone are insufficient to guarantee reliable outcomes. Effective crowdsourcing systems combine motivational incentives with quality assurance mechanisms such as qualification tests, performance monitoring, gold-standard validation tasks, aggregation methods (e.g., majority voting), and peer review processes. The balance between motivational design and quality

control depends on task characteristics and platform objectives. This highlights that sustainable crowdsourcing requires a careful integration of participation incentives and quality management mechanisms.

RQ4: How is crowdsourced data licensed, and what licensing categories are commonly used?

The findings in Section 4.4 show that licensing practices in crowdsourcing systems are often not explicitly defined and are frequently embedded within platform terms of service or project-specific agreements rather than standardized legal frameworks. Three dominant licensing models emerge from the review: ownership transfer and seeker-centric licensing, open licensing frameworks such as Creative Commons and open-source models, and contractual or confidentiality-based arrangements including NDAs.

These licensing categories reflect different strategic orientations, ranging from proprietary control to open knowledge sharing. Licensing choices influence contributor rights, conditions for data reuse, transparency, and the distribution of control between platform operators and participants. The synthesis therefore suggests that licensing should be considered a structural component of crowdsourcing system design, shaping governance structures, long-term sustainability, and ethical considerations.

Overall, the reflections on RQ1–RQ4 show that crowdsourcing systems are shaped by multiple interacting design dimensions. The classification of crowdsourcing approaches (RQ1), their application across different operational contexts (RQ2), the incentive and quality mechanisms that sustain participation and reliability (RQ3), and the licensing frameworks governing crowdsourced outputs (RQ4) collectively define how such systems operate in practice. These dimensions are closely interrelated and influence the effectiveness, governance, and sustainability of crowdsourcing platforms. Understanding these relationships provides the conceptual basis for evaluating the suitability of crowdsourcing approaches for structured data curation in the SCA Tool.

5.2 Applicability of Crowdsourcing Approaches to Data Curation in the SCA Tool (RQ5)

Building on the reflections presented in the previous section on RQ1–RQ4, this section evaluates how established crowdsourcing approaches can support data curation in the SCA Tool in relation to the operational requirements of SCA-based curation. Within the SCA Tool, curation activities include validating, correcting, and continuously updating data related to software components, licenses, and vulnerabilities. These tasks vary in complexity, required expertise, and associated risk.

For the purpose of evaluating crowdsourcing applicability, these activities can be categorized according to their operational risk levels, ranging from low-risk routine verification tasks to high-risk vulnerability assessments that may influence security-related decision making. Consequently, no single crowdsourcing approach is sufficient to support the entire curation workflow. Instead, effective curation requires a structured combination of approaches aligned with the characteristics and risk levels of different tasks.

The following subsections examine how established crowdsourcing models can be mapped to specific curation activities within the SCA Tool.

5.2.1 Microtasking for Routine Verification

Microtasking is well suited for repetitive and high-volume curation activities that require limited contextual interpretation. In the SCA Tool, such tasks may include verifying automated scanner outputs, correcting inaccurate license references, or confirming metadata fields associated with software components. These activities are typically well-defined, standardized, and can be performed independently.

Microtasking relies on decomposing large workloads into small and homogeneous task units that can be executed in parallel by multiple contributors. Quality can be maintained through selective aggregation mechanisms, such as majority voting or weighted consensus, which combine multiple independent submissions to produce reliable outcomes. This approach enables scalable human verification and introduces structured human oversight into automated SCA workflows.

However, microtasking is primarily suitable for low-risk tasks with low complexity and explicitly defined outcomes. It is less appropriate for curation activities that require deep technical reasoning, interpretation of ambiguous information, or broader contextual understanding.

5.2.2 Macrotasking for Vulnerability Assessment

More complex curation activities, particularly those related to vulnerability assessment, require a different approach. Determining whether a reported vulnerability affects a specific software component often depends on configuration details, usage context, dependency chains, and technical assumptions. Such decisions require contextual interpretation and cannot easily be decomposed into simple validation steps.

In these cases, macrotasking enables the involvement of experienced contributors who can evaluate vulnerability reports within their broader technical context. Macro-tasks are characterized by high complexity and limited decomposability, and therefore rely on expert knowledge and holistic assessment rather than parallel task execution. This makes macrotasking suitable for high-risk curation activities that require careful reasoning and domain expertise.

Given the potential downstream impact of incorrect vulnerability assessments, appropriate governance mechanisms are essential. Decisions should remain traceable, and clear escalation procedures should be defined for high-risk cases. Without structured accountability, distributed expert assessment may introduce inconsistency rather than reliability. In the context of the SCA Tool, macrotasking therefore requires controlled participation, clear responsibility boundaries, and documented decision authority.

5.2.3 Information Pooling for Resolving Uncertain or Conflicting Findings

Not all curation decisions produce clear or binary outcomes. In cases where scanner results conflict or the relevance and severity of an issue are uncertain, the SCA Tool can adopt an information pooling approach.

In this model, contributors independently evaluate the same finding, and the system aggregates their inputs to derive a collective assessment or confidence level. By combining multiple homogeneous contributions, information pooling produces more reliable results. This reduces reliance on individual judgments and helps mitigate bias.

However, aggregation strategies must account for differences in contributor expertise. Simple majority voting may be insufficient when technical knowledge varies among participants. Weighted aggregation or reputation-based scoring mechanisms may therefore be required to maintain decision quality. In the SCA Tool context, information pooling is most appropriate for medium-risk curation tasks characterized by uncertainty but not requiring full expert analysis.

5.2.4 Open Collaboration for Shared Knowledge Curation

Certain outputs, such as license knowledge bases, component relationship mappings, and shared curation guidelines, represent long-term knowledge resources rather than isolated validation tasks. Maintaining these resources requires continuous updates and iterative refinement over time.

Open collaboration is well suited for this type of activity, as it allows heterogeneous contributions to be gradually integrated into a shared resource. Contributors collaboratively add, review, and refine information. While individual inputs often represent incremental improvements, quality emerges over time through iterative integration and community review.

For open collaboration to function effectively in the SCA Tool context, moderation and governance mechanisms are necessary. Version control, clear contribution guidelines, and structured review procedures help to maintain consistency, ensure traceability, and prevent the degradation of shared knowledge resources. Because these resources evolve gradually and are subject to community review and moderation, the associated operational risk remains moderate compared to direct vulnerability assessment tasks.

5.2.5 Hybrid Human-Machine Curation System

To maintain scalability while preserving quality, the SCA Tool can adopt a hybrid human–machine architecture that integrates the crowdsourcing approaches discussed above. In this model, automated components process well-defined and high-volume tasks, while human contributors are engaged selectively for ambiguous, complex, or high-risk cases. This layered architecture allows the system to handle tasks across multiple risk levels, combining automated processing for low-risk tasks with human oversight for cases involving medium to high risk.

Most scanner outputs can be handled automatically, with only uncertain findings escalated to human review. Routine verification tasks can be addressed through microtasking, conflicting findings through information pooling, and complex vulnerability assessments through macrotasking. In addition, open collaboration can support the continuous maintenance of shared curation knowledge.

Feedback from human interventions can be fed into system improvement processes, gradually increasing automated accuracy over time. The literature consistently indicates that crowdsourcing is most effective when it complements automation rather than replacing it. A hybrid human–machine architecture therefore provides a practical foundation for scalable and reliable SCA-based data curation.

Curation Activity	Risk Level	Suitable Approach
Routine verification of automated scanner outputs	Low	**Microtasking**
Analysis of complex vulnerability reports	High	**Macrotasking**
Resolving uncertain or conflicting findings	Medium	**Information pooling**
Collaborative maintenance of shared knowledge resources	Medium	**Open collaboration**
Scalable curation combining automation and human review	Multi-level	**Hybrid human-machine system**

Table 5.1: Risk-Tiered Crowdsourcing Approaches for Curation Activities in the SCA Tool

Taken together, the analysis indicates that no single crowdsourcing model is sufficient to support the full range of curation activities required in the SCA Tool. Instead, effective data curation requires a structured combination of approaches in which microtasking, information pooling, macrotasking, and open collaboration are applied according to task complexity, uncertainty, and operational risk levels. This leads to a risk-tiered crowdsourcing architecture, where different crowdsourcing approaches are systematically assigned to tasks based on their associated risk and expertise requirements. As summarized in Table 5.1, low-risk routine verification tasks are well suited for microtasking, medium-risk and uncertain cases benefit from information pooling and collaborative knowledge maintenance, while high-risk vulnerability assessments require expert-driven macrotasking. When integrated within a hybrid human–machine architecture, these approaches enable scalable yet reliable curation workflows across different risk levels. However, implementing such a model also introduces operational and ethical considerations, which are discussed in the following section.

5.3 Benefits, Challenges, and Ethical Concerns of Crowdsourcing in Data Curation (RQ6)

Building on the general findings presented in Section 4.5, this section interprets the benefits, operational challenges, and ethical concerns of applying crowdsourcing to data curation in the SCA Tool. While crowdsourcing offers clear advantages for scalable and distributed curation, its application also introduces risks that must be carefully managed. The following discussion shows how these factors influence the practical adoption of crowdsourcing within SCA-based curation workflows.

5.3.1 Benefits in SCA Context

From a practical perspective, crowdsourcing directly addresses key limitations of traditional expert-driven curation. Data within SCA tooling platforms require continuous validation and updating, particularly with regard to software components, licenses, and vulnerabilities. Maintaining this data exclusively through internal expert teams may lead to bottlenecks, limited scalability, and delayed updates.

Crowdsourcing enables parallel processing of curation tasks, especially for routine verification activities. As discussed in the previous section, microtasking supports scalable human validation of automated scanner outputs, while information pooling can assist in resolving uncertain findings. This distributed model increases responsiveness and reduces dependency on a small group of maintainers.

In addition, crowdsourcing provides access to distributed human expertise. Certain curation tasks, such as vulnerability assessment or license interpretation, require contextual understanding that automated tools cannot fully provide. Macrotasking allows qualified contributors to apply domain knowledge in situations where automation reaches its limits.

Another important benefit lies in the diversity of perspectives. Aggregating inputs from multiple contributors can improve robustness and reduce individual bias, particularly in medium-risk tasks characterized by uncertainty. When supported by structured quality mechanisms, this diversity can strengthen overall data reliability.

5.3.2 Operational Challenges

Despite these advantages, applying crowdsourcing to SCA-based curation introduces significant operational challenges. The most critical concern relates to data quality where incorrect curation decisions may affect compliance analyses, vulnerability management workflows, and legal assessments. In contrast to some crowdsourcing applications where errors remain isolated, inaccuracies in SCA data

may have downstream regulatory or security implications. Quality assurance is therefore essential, particularly for tasks with medium or high risk.

Another challenge concerns task design and aggregation mechanisms. Curation tasks must be defined clearly while preserving essential technical context. Over-simplification may remove critical information, whereas excessive complexity may discourage participation. Achieving the right level of task granularity is therefore necessary to balance scalability and reliability. Similarly, aggregation mechanisms must be carefully selected. For low-risk tasks, majority voting may be sufficient, but in cases involving technical interpretation or expert disagreement, more advanced approaches such as weighted scoring or reputation-based filtering are required. Without structured aggregation, contribution volume alone does not guarantee decision quality.

Finally, sustaining participation over time can be difficult. High-complexity tasks demand significant expertise and cognitive effort. Without appropriate incentives and recognition structures, contributor retention may decline, particularly for macrotasking activities.

5.3.3 Ethical Considerations

Ethical considerations are closely linked to operational design choices. Decisions about task allocation, incentive structures, quality control, and access management directly influence fairness, transparency, and accountability within the system. Fair compensation is a central issue when contributors are asked to perform cognitively demanding tasks. If compensation levels are not proportionate to task complexity and responsibility, participation may become unsustainable or ethically problematic.

Power asymmetry between platform operators and contributors must also be addressed. In SCA tooling environments, contributors may have limited visibility into how their decisions are used or evaluated. Clear feedback mechanisms and transparent governance structures are therefore necessary to maintain trust. Privacy and data protection are also particularly relevant in SCA-based curation, as some tasks may involve proprietary software data or sensitive vulnerability information. Access control mechanisms and clearly specified data boundaries are required to prevent unintended disclosure.

IP and licensing transparency represent another important ethical dimension. Contributors should understand how their curation inputs are stored, reused, or redistributed. Licensing models must align with the governance goals of the SCA Tool while respecting contributor rights.

Together, applying crowdsourcing to SCA-based data curation involves clear trade-offs. Scalability and access to distributed expertise increase responsiveness, but they also require stronger quality assurance and governance as task complexity and risk levels rise. As discussed in Section 5.2, different crowdsourcing approaches align with different risk levels. Successful integration therefore depends not only on selecting appropriate approaches, but also on aligning risk management, quality control, incentive structures, and ethical safeguards. Without such alignment, increased participation may reduce consistency rather than improve reliability. The following section therefore examines how these participation, quality, and ethical challenges can be addressed through appropriate system design and governance mechanisms.

5.4 Addressing Participation, Quality & Ethical Challenges in Crowdsourced Curation

The challenges discussed in Section 5.3 indicate that applying crowdsourcing to SCA-based data curation requires deliberate and structured system design. Participation, quality assurance, and ethical safeguards cannot be treated as isolated concerns. Instead, they must be integrated into the operational architecture of the platform.

Effective participation depends on aligning incentive mechanisms with task complexity and responsibility. For routine and low-risk tasks, lightweight incentive mechanisms are likely to be most effective. Gamification-based incentives, including contribution points, progress indicators, badges, leaderboards, or task completion milestones, can encourage engagement while keeping participation voluntary and intrinsically motivated. These approaches support sustained involvement without creating strong dependence on monetary rewards. As discussed earlier, excessive reliance on financial incentives may weaken intrinsic motivation and reduce long-term contributor commitment.

For higher-complexity tasks, such as vulnerability assessment or context-dependent license interpretation, stronger intrinsic motivators are necessary. Reputation systems, visible indicators of expertise, and role-based privileges can help signal accountability and attract qualified contributors. Structured collaboration mechanisms, including discussion spaces and peer interaction, can further support knowledge exchange and shared responsibility.

Quality assurance mechanisms must operate in alignment with risk levels. Low-risk tasks can rely on aggregation methods such as majority voting. Medium-risk tasks require controlled aggregation, potentially incorporating weighted scoring or reputation-based evaluation. High-risk tasks demand well-structured responsibility frameworks, including expert review and documented decision authority. Such risk-aware quality control mechanisms ensure that decision rigor increases in proportion to the potential impact of the task.

Building on the ethical considerations discussed in the previous section, safeguards must be embedded within platform governance and operational design. Platforms should implement transparent contribution guidelines, clear feedback mechanisms, and structured governance processes to reduce power imbalances between platform operators and contributors. Incentive policies should remain fair and proportionate to task complexity and risk level. Additionally, contributors should understand how their efforts are evaluated and recognized. In contexts where sensitive or proprietary data are involved, appropriate access control mechanisms and well-specified information boundaries must be included into the platform governance framework.

In summary, sustainable crowdsourced curation depends on a structured alignment between incentive mechanisms, quality assurance processes, risk management strategies, and ethical governance. As task complexity and potential risk level increase, stronger accountability and clearer decision authority become necessary. Crowdsourcing in SCA-based data curation is neither inherently reliable nor inherently problematic. Its effectiveness is determined by how well system design reflects task characteristics, risk exposure, incentive structures, quality control mechanisms, and governance safeguards. When these elements are coherently integrated, crowdsourcing can serve as a viable and structured complement to automated SCA workflows.

6 Conclusion

The primary objective of this thesis is to examine whether crowdsourcing can support data curation in SCA environments, particularly in the context of the SCA Tool that motivated this study. The SCA Tool analyzes software components, identifies license obligations, and detects known vulnerabilities within software dependencies. Its effectiveness depends on the accuracy and continuous maintenance of component metadata, licensing information, and vulnerability records.

As software ecosystems grow in complexity, SCA tooling platforms increasingly rely on continuously updated datasets to ensure compliance and security. Although automated scanners provide scalability and efficiency, they cannot fully resolve contextual ambiguity, interpret complex licensing conditions, or assess the practical relevance of vulnerabilities in specific operational environments. These limitations highlight the need for structured human involvement in data validation and refinement.

To address this objective, an SLR was conducted in accordance with the guidelines of Kitchenham et al. (2004) and Kitchenham, Charters et al. (2007), synthesizing evidence from 27 selected studies.

The study followed a two-phase analytical structure. First, it synthesized existing knowledge on crowdsourcing approaches, application domains, incentive and quality mechanisms, and licensing models for crowdsourced data, as presented in Chapter 4. Second, it evaluated the suitability of these approaches for supporting data curation within the SCA Tool, assessing their practical implications, benefits, challenges, and governance considerations, as discussed in Chapter 5.

6.1 Summary of Research Findings

The analysis of established crowdsourcing approaches (RQ1) shows that crowdsourcing is not a monolithic paradigm but a set of structurally distinct models that differ in task structure, contribution diversity, aggregation mechanisms, and the type of value generated by crowd contributions. The literature distinguishes microtasking, information pooling, broadcast search, and open collaboration as primary structural approaches based on contribution diversity and aggregation logic. From a task-complexity perspective, crowdsourcing activities can further be categorized into micro, complex, macro, and creative tasks according to their level of cognitive demand and degree of decomposability. In addition, domain-specific forms such as crowd testing and crowdfunding demonstrate how crowdsourcing principles are adapted to different functional contexts and participation models. Together, these complementary classifications provide a multidimensional framework for understanding how distributed participation generates value and how different approaches align with varying levels of expertise, coordination requirements, and task complexity.

The examination of real-world implementations (RQ2) confirms that crowdsourcing approaches have been successfully applied across diverse domains, including collaborative knowledge production, geospatial information systems, and scientific research. The comparative analysis of representative platforms illustrates how different crowdsourcing approaches organize participation, contributions, and aggregation mechanisms according to their primary objectives. Wikipedia exemplifies open collaboration, where contributors provide heterogeneous inputs that are iteratively integrated to create and maintain shared knowledge resources. Google Maps represents information pooling, in which large numbers of users submit relatively homogeneous observations that are aggregated to maintain up-to-date geospatial information. Zooniverse demonstrates microtasking, where complex scientific problems are decomposed into small and well-defined tasks that volunteers can complete independently. These cases highlight how crowdsourcing systems differ in contribution diversity, aggregation logic, and task complexity depending on their operational goals. The findings therefore indicate that the effectiveness of crowdsourcing systems depends on the alignment between platform goals, task characteristics, and the selected crowdsourcing approach.

With regard to participation sustainability and contribution reliability (RQ3), the literature shows that engagement in crowdsourcing systems is shaped by a combination of extrinsic and intrinsic incentives. While extrinsic incentives such as monetary compensation and competitive prizes can stimulate short-term engagement, long-term participation is more closely associated with intrinsic motivators such as reputation, learning opportunities, and prosocial contribution. At the same time, incentive structures alone are insufficient to ensure high-quality

outputs. Effective crowdsourcing systems therefore adopt structured quality assurance mechanisms, including qualification filters, gold-standard validation tasks, aggregation techniques, and peer review processes. These mechanisms help to evaluate contributor reliability and mitigate bias or low-quality submissions, ensuring that reliable outcomes emerge from structured governance rather than participation volume alone.

Finally, licensing practices governing crowdsourced data in crowdsourcing systems (RQ4) lack a uniform standard and are often embedded within platform-specific governance frameworks or terms of service. Licensing models range from seeker-centric ownership transfer, commonly used in commercial crowdsourcing environments, to open licensing frameworks based on Creative Commons or open-source principles, which are typical in scientific, collaborative and open knowledge projects. In addition, some platforms employ contractual or confidentiality-based arrangements, such as NDAs, particularly in contexts involving proprietary data or sensitive tasks. These differences reflect varying governance objectives and have direct implications for contributor rights, data reuse, and system transparency. The findings suggest that licensing should be treated as an integral structural dimension of crowdsourcing system design rather than as a peripheral legal detail.

6.2 Practical Implications for the SCA Tool

The core contribution of this thesis is the assessment of how these synthesized findings support the specific curation needs of the SCA Tool (RQ5 and RQ6). The analysis indicates that crowdsourcing can support SCA workflows when applied selectively and structured according to task characteristics and associated risk levels.

Low-risk routine verification tasks are well aligned with microtasking, where well-defined and repetitive activities can be distributed efficiently. Medium-risk or uncertain cases benefit from structured aggregation through information pooling, which enables independent assessments and controlled aggregation mechanisms. High-risk and technically complex assessments require macrotasking with clearly structured responsibility frameworks and expert oversight to ensure traceability and accountability.

In addition, open collaboration is particularly suitable for maintaining shared and evolving curation resources, such as license knowledge bases, component relationship mappings, and documentation guidelines. Through iterative contributions and transparent revision processes, open collaboration enables long-term knowledge development beyond isolated validation tasks.

Furthermore, hybrid human–machine systems combine automated processing with multiple crowdsourcing models to support scalable curation workflows. Automated components handle well-defined and high-volume tasks, while human contributors are engaged selectively for cases that require contextual interpretation or technical expertise. Within this architecture, different crowdsourcing approaches can be invoked depending on the nature and risk level of the task, enabling flexible coordination between automated processing and human judgment. Feedback from human interventions can then be fed into iterative system improvement processes, which will gradually improve the accuracy of automated outputs over time. This structured interaction between machine efficiency and human judgment enables scalability while preserving contextual oversight.

Effective implementation also requires operational and ethical safeguards embedded within the overall system design. Quality assurance mechanisms must operate according to risk levels, with more rigorous verification applied to complex and high-impact tasks. Such risk-aware quality control ensures that decision rigor increases in proportion to potential impact. Participation design should also align incentive structures with task complexity, combining gamification-based engagement for routine tasks with reputation systems and role-based recognition for expert contributions. Incentive policies must remain fair and proportionate to responsibility, while excessive reliance on financial rewards should be avoided, as it may weaken intrinsic motivation and reduce long-term contributor commitment.

Transparent governance processes, clear feedback mechanisms, and appropriate access controls are essential to maintain trust and accountability within the platform. Contributors should understand how their inputs are evaluated, stored, and reused, while licensing transparency and data protection measures must be integrated into the governance framework of the system.

Overall, the findings indicate that crowdsourcing can serve as a reliable and scalable complement to SCA-based data curation when it is implemented through a structured and risk-tiered framework. As discussed in Chapter 5, effective integration requires aligning different participation models with the complexity and risk level of curation tasks while combining automated processing with selective human oversight. A hybrid human–machine architecture enables scalable workflows, while governance mechanisms such as risk-aware quality control, transparent participation structures, and appropriate incentive designs help maintain reliability and accountability. When these elements are coherently integrated, crowdsourcing can strengthen the responsiveness, accuracy, and long-term sustainability of SCA data curation.

6.3 Limitations and Future Work

This study is based on a systematic synthesis of existing literature and does not include empirical validation within a live SCA tooling environment. The applicability assessment therefore remains conceptual. In addition, the review was limited to studies published between January 2018 and August 2025. While this time-frame ensured a focus on recent developments in crowdsourcing research, earlier foundational studies as well as developments emerging after August 2025 were not included. Consequently, some theoretical perspectives and more recent advancements in crowdsourcing platforms, incentive design, and hybrid human–machine systems may not have been fully captured.

Future research should evaluate the proposed models through practical implementation within SCA tooling environments. Empirical studies could measure data quality outcomes, review efficiency, scalability, and contributor engagement under different incentive structures and governance configurations. Such investigations would clarify how risk-tiered crowdsourcing architectures perform in real-world SCA workflows. In addition, machine learning–based methods could be developed to support adaptive task assignment, contributor trust scoring, anomaly detection, and automated quality validation within crowdsourced curation processes. Integrating learning-based models with human feedback loops may strengthen the interaction between automated scanners and human contributors, improving both accuracy and efficiency over time.

In conclusion, this thesis demonstrates that crowdsourcing can serve as a viable complement to automated SCA workflows when implemented through a structured and risk-aware architecture. Rather than replacing automated analysis, crowdsourcing can enhance SCA systems by introducing scalable human judgment for tasks involving ambiguity, contextual interpretation, and evolving software ecosystems. By strategically aligning task characteristics, participant incentives, quality assurance mechanisms, and ethical safeguards, this study provides a concrete foundation for integrating human intelligence into future SCA tooling environments. As software ecosystems continue to expand in complexity, hybrid human–machine approaches will play an increasingly important role in maintaining reliable, continuously updated software metadata and vulnerability information.

Appendices

A List of Studies Included in the SLR

Included Studies

Acar, O. A. (2019). Motivations and solution appropriateness in crowdsourcing challenges for innovation. *Research Policy*, *48*(8), 103716

Alenezi, H. S., & Faisal, M. H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. *Education and Information Technologies*, *25*(4), 2971–2986

Allon, G., & Babich, V. (2020). Crowdsourcing and crowdfunding in the manufacturing and services sectors. *Manufacturing & Service Operations Management*, *22*(1), 102–112

Bhatti, S. S., Gao, X., & Chen, G. (2020). General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey. *Journal of Systems and Software*, *167*, 110611

Bhuyan, B. P., & Singh, M. (2023). Introduction to crowdsourcing. In *Social media and crowdsourcing* (pp. 1–31). Auerbach Publications

Blohm, I., Zogaj, S., Bretschneider, U., & Leimeister, J. M. (2018). How to manage crowdsourcing platforms effectively? *California Management Review*, *60*(2), 122–149

Cappa, F., Rosso, F., & Hayes, D. (2019). Monetary and social rewards for crowdsourcing. *Sustainability*, *11*(10), 2834

Chen, T., Han, L., Demartini, G., Indulska, M., & Sadiq, S. (2020). Building data curation processes with crowd intelligence. *International Conference on Advanced Information Systems Engineering*, 29–42

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170

Ghezzi, A., Gabelloni, D., Martini, A., & Natalicchio, A. (2018). Crowdsourcing: A review and suggestions for future research. *International Journal of management reviews*, *20*(2), 343–363

Hosseini, M., Angelopoulos, C. M., Chai, W. K., & Kundig, S. (2019). Crowdcloud: A crowdsourced system for cloud infrastructure. *Cluster Computing*, *22*(2), 455–470

Kohler, T. (2018). How to scale crowdsourcing platforms. *California Management Review*, *60*(2), 98–121

Kong, X., Liu, X., Jedari, B., Li, M., Wan, L., & Xia, F. (2019). Mobile crowdsourcing in smart cities: Technologies, applications, and future challenges. *IEEE Internet of Things Journal*, *6*(5), 8095–8113

Li, M., Weng, J., Yang, A., Lu, W., Zhang, Y., Hou, L., Liu, J.-N., Xiang, Y., & Deng, R. H. (2018). Crowdbc: A blockchain-based decentralized framework for crowdsourcing. *IEEE transactions on parallel and distributed systems*, *30*(6), 1251–1266

Lykourentzou, I., Khan, V.-J., Papangelis, K., & Markopoulos, P. (2019). Macrotask crowdsourcing: An integrated definition. In *Macrotask crowdsourcing: Engaging the crowds to address complex problems* (pp. 1–13). Springer

Morschheuser, B., & Hamari, J. (2019). The gamification of work: Lessons from crowdsourcing. *Journal of Management Inquiry*, *28*(2), 145–148

Neto, F. R. A., & Santos, C. A. (2018). Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management. *Information Processing & Management*, *54*(4), 490–506

Pirttinen, N. (2021). Crowdsourcing in computer science education. *Proceedings of the 17th ACM Conference on International Computing Education Research*, 421–422

Sheehan, K. B. (2018). Crowdsourcing research: Data collection with amazon's mechanical turk. *Communication monographs*, *85*(1), 140–156

Sheng, V. S., & Zhang, J. (2019). Machine learning with crowdsourcing: A brief summary of the past research and future directions. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 9837–9843

Standing, S., & Standing, C. (2018). The ethical use of crowdsourcing. *Business Ethics: A European Review*, *27*(1), 72–80

Sarı, A., Tosun, A., & Alptekin, G. I. (2019). A systematic literature review on crowdsourcing in software engineering. *Journal of Systems and Software*, *153*, 200–219

Tucker, J. D., Day, S., Tang, W., & Bayus, B. (2019). Crowdsourcing in medical research: Concepts and applications. *PeerJ*, *7*, e6762

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific utopia iii: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733

Vaughan, J. W. (2018). Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, *18*(193), 1–46

Wang, C., Han, L., Stein, G., Day, S., Bien-Gund, C., Mathews, A., Ong, J. J., Zhao, P.-Z., Wei, S.-F., Walker, J., et al. (2020). Crowdsourcing in health and medical research: A systematic review. *Infectious diseases of poverty*, *9*(1), 8

Wazny, K. (2018). Applications of crowdsourcing in health: An overview. *Journal of global health*, *8*(1), 010502

References

Acar, O. A. (2019). Motivations and solution appropriateness in crowdsourcing challenges for innovation. *Research Policy*, *48*(8), 103716.

Alenezi, H. S., & Faisal, M. H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. *Education and Information Technologies*, *25*(4), 2971–2986.

Allon, G., & Babich, V. (2020). Crowdsourcing and crowdfunding in the manufacturing and services sectors. *Manufacturing & Service Operations Management*, *22*(1), 102–112.

Ambreen, T., & Ikram, N. (2016). A state-of-the-art of empirical literature of crowdsourcing in computing. *2016 IEEE 11th International Conference on Global Software Engineering (ICGSE)*, 189–190.

Bhatti, S. S., Gao, X., & Chen, G. (2020). General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey. *Journal of Systems and Software*, *167*, 110611.

Bhuyan, B. P., & Singh, M. (2023). Introduction to crowdsourcing. In *Social media and crowdsourcing* (pp. 1–31). Auerbach Publications.

Blohm, I., Zogaj, S., Bretschneider, U., & Leimeister, J. M. (2018). How to manage crowdsourcing platforms effectively? *California Management Review*, *60*(2), 122–149.

Cappa, F., Rosso, F., & Hayes, D. (2019). Monetary and social rewards for crowdsourcing. *Sustainability*, *11*(10), 2834.

Chen, T., Han, L., Demartini, G., Indulska, M., & Sadiq, S. (2020). Building data curation processes with crowd intelligence. *International Conference on Advanced Information Systems Engineering*, 29–42.

ClearlyDefined Project. (2026). *Clearlydefined documentation: Helping open source projects be more successful through clearly defined licensing data* [Accessed: 2026-02-18]. ClearlyDefined. https://docs.clearlydefined.io/

Eickhoff, C. (2018). Cognitive biases in crowdsourcing. *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

Galton, F. (1907). Vox populi.

Ghezzi, A., Gabelloni, D., Martini, A., & Natalicchio, A. (2018). Crowdsourcing: A review and suggestions for future research. *International Journal of management reviews*, *20*(2), 343–363.

Giles, J. (2005). Special report internet encyclopaedias go head to head. *nature*, *438*(15), 900–901.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Haklay, M. M. (2016). Why is participation inequality important? In *European handbook of crowdsourced geographic information*. Ubiquity Press.

Hosseini, M., Angelopoulos, C. M., Chai, W. K., & Kundig, S. (2019). Crowdcloud: A crowdsourced system for cloud infrastructure. *Cluster Computing*, *22*(2), 455–470.

Hosseini, M., Shahri, A., Phalp, K., Taylor, J., & Ali, R. (2015). Crowdsourcing: A taxonomy and systematic mapping study. *Computer Science Review*, *17*, 43–69.

Howe, J., et al. (2006). The rise of crowdsourcing. *Wired magazine*, *14*(6), 176–183.

Kitchenham, B., et al. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1–26.

Kitchenham, B., Charters, S., et al. (2007). Guidelines for performing systematic literature reviews in software engineering.

Kohler, T. (2018). How to scale crowdsourcing platforms. *California Management Review*, *60*(2), 98–121.

Kong, X., Liu, X., Jedari, B., Li, M., Wan, L., & Xia, F. (2019). Mobile crowdsourcing in smart cities: Technologies, applications, and future challenges. *IEEE Internet of Things Journal*, *6*(5), 8095–8113.

Li, M., Weng, J., Yang, A., Lu, W., Zhang, Y., Hou, L., Liu, J.-N., Xiang, Y., & Deng, R. H. (2018). Crowdbc: A blockchain-based decentralized framework for crowdsourcing. *IEEE transactions on parallel and distributed systems*, *30*(6), 1251–1266.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., et al. (2008). Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179–1189.

Lykourentzou, I., Khan, V.-J., Papangelis, K., & Markopoulos, P. (2019). Macrotask crowdsourcing: An integrated definition. In *Macrotask crowdsourcing: Engaging the crowds to address complex problems* (pp. 1–13). Springer.

Malone, T. W., Laubacher, R., & Dellarocas, C. (2010). The collective intelligence genome. *MIT Sloan management review*.

Mao, K., Capra, L., Harman, M., & Jia, Y. (2017). A survey of the use of crowdsourcing in software engineering. *Journal of Systems and Software*, *126*, 57–84.

Morschheuser, B., & Hamari, J. (2019). The gamification of work: Lessons from crowdsourcing. *Journal of Management Inquiry*, *28*(2), 145–148.

Neto, F. R. A., & Santos, C. A. (2018). Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management. *Information Processing & Management*, *54*(4), 490–506.

OSSelot Project. (2026). *Osselot wiki: Documentation for crowdsourced open source data curation* [Accessed: 2026-02-18]. OSSelot. https://wiki.osselot.org/index.php/Main_Page

Pirttinen, N. (2021). Crowdsourcing in computer science education. *Proceedings of the 17th ACM Conference on International Computing Education Research*, 421–422.

Sarı, A., Tosun, A., & Alptekin, G. I. (2019). A systematic literature review on crowdsourcing in software engineering. *Journal of Systems and Software*, *153*, 200–219.

Sheehan, K. B. (2018). Crowdsourcing research: Data collection with amazon's mechanical turk. *Communication monographs*, *85*(1), 140–156.

Sheng, V. S., & Zhang, J. (2019). Machine learning with crowdsourcing: A brief summary of the past research and future directions. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 9837–9843.

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. *Proceedings of the 23rd international conference on world wide web*, 1049–1054.

Sobel, D. (2007). *Longitude: The true story of a lone genius who solved the greatest scientific problem of his time*. Bloomsbury Publishing USA.

Standing, S., & Standing, C. (2018). The ethical use of crowdsourcing. *Business Ethics: A European Review*, *27*(1), 72–80.

Stol, K.-J., LaToza, T. D., & Bird, C. (2017). Crowdsourcing for software engineering. *IEEE software*, *34*(2), 30–36.

Thuan, N. H., Antunes, P., & Johnstone, D. (2016). Factors influencing the decision to crowdsource: A systematic literature review. *Information Systems Frontiers*, *18*(1), 47–68.

Tucker, J. D., Day, S., Tang, W., & Bayus, B. (2019). Crowdsourcing in medical research: Concepts and applications. *PeerJ*, *7*, e6762.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific utopia iii: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733.

Vaughan, J. W. (2018). Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, *18*(193), 1–46.

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., & Wagenknecht, K. (2021). *The science of citizen science*. Springer Nature.

Wang, C., Han, L., Stein, G., Day, S., Bien-Gund, C., Mathews, A., Ong, J. J., Zhao, P.-Z., Wei, S.-F., Walker, J., et al. (2020). Crowdsourcing in health and medical research: A systematic review. *Infectious diseases of poverty*, *9*(1), 8.

Wazny, K. (2018). Applications of crowdsourcing in health: An overview. *Journal of global health*, *8*(1), 010502.