

Auditable Crowdsourced Digital Dictionary for Low-Resource Toulour with Gamification and Functional Evaluation

¹Kenneth Yosua Palilingan*, ²Salaki Reynaldo Joshua, ³Salvius Paulus Lengkong
^{1,2,3}Department of Informatics Engineering, Faculty of Engineering, Sam Ratulangi University
^{1,2,3}Kampus UNSRAT Bahu, Manado 95115, North Sulawesi, Indonesia
*e-mail: kennethpalilingan@unsrat.ac.id

(**received:** 10 September 2026, **revised:** 23 September 2026, **accepted:** 24 September 2026)

Abstract

Toulour, also documented as Tondano/Toundano, is a low-resource Minahasan language whose digital lexical resources require accessible retrieval and accountable community maintenance. This study designs and evaluates an auditable crowdsourced and gamified Toulour–Indonesian digital dictionary in which participation incentives are separated from lexical authority. Using Design Science Research Methodology (DSRM), the artifact integrates bidirectional lexical access, contributor proposals, validator-controlled publication, persistent before/after provenance, experience points, quizzes, badges, feedback, and role-based administration. Functional evaluation used 50 predefined black-box cases structured by ISO/IEC 25010:2023 Functional Suitability and informed by ISO/IEC/IEEE 29119-4:2021; 49 cases passed (98.0%), with one oversized-query boundary failure retained. Supplementary evaluation included a pilot user study, automated application-surface security scanning, and staged local load testing. Of 40 questionnaire submissions, 34 eligible adults who had tried the application were analyzed; mean System Usability Scale (SUS) was 77.79, while descriptive perceived usefulness, perceived ease of use, and behavioral intention means were 4.10, 4.07, and 3.90 on a 1–5 scale. Final context-scoped OWASP ZAP verification reported 0 High, 3 Medium, 3 Low, and 4 Informational alert types. k6 testing at 1, 10, 25, and 50 virtual users observed no HTTP request failures, although latency rose substantially at higher concurrency and throughput gains flattened between 25 and 50 virtual users. The results support an operational integration of governed crowdsourcing, reconstructable provenance, participation-linked incentives, and complementary evaluation while retaining explicit pilot, local-deployment, and residual-security limitations.

Keywords: controlled crowdsourcing, digital dictionary, gamification, lexical provenance, low-resource language

1 Introduction

Local languages require digital resources that reflect their social functions, community identities, and practical conditions rather than merely reproducing infrastructure designed for globally dominant languages [1]. Toulour is a Minahasan language spoken in and around the Tondano area of North Sulawesi, Indonesia. The principal lexicographic source used in this study is Kamus Tondano–Indonesia, published in 1985 [2], while the implemented system uses Toulour as its community-facing designation. In this article, Toulour and Tondano/Toundano refer to the same ethnolinguistic language entity, with the latter retained when discussing the historical dictionary and linguistic literature. Contemporary description identifies Tondano/Toundano as an endangered Minahasan language [3]. These conditions make lexical access, maintenance, and provenance important information-system design concerns.

Digitizing a printed dictionary is necessary, but digitization alone does not create a sustainable lexical information system. Legacy entries must be converted into structured records, normalized, checked, and presented in forms that support discovery and reuse [7]. Once the resource becomes online and editable, additional questions arise: who may propose a word or correction, who determines whether a proposal becomes public lexical content, how a correction changes an approved entry, and whether earlier states remain inspectable. Open contribution without explicit governance

can weaken lexical reliability, whereas a closed database can become detached from speaker knowledge. The design problem is therefore to enable community participation while preserving accountable linguistic review and traceable change.

This study addresses three research questions. RQ1 asks how a low-resource bilingual dictionary can combine community crowdsourcing, validator-controlled acceptance, persistent provenance, and gamification without allowing incentives to determine lexical acceptance. RQ2 asks to what extent the artifact satisfies its specified functional requirements under ISO/IEC 25010:2023 Functional Suitability using predefined black-box scenarios. RQ3 asks what supplementary descriptive evidence can be obtained on user-perceived usability and acceptance, application-surface security, and local scalability. The study therefore contributes an operational Toulour–Indonesian artifact and an auditable governance design linking proposals, validator decisions, public entries, revision history, and participation rewards. Its central design contribution is the separation of participation incentives from lexical authority while retaining reconstructable evidence of lexical change.

2 Literature Review

The local scarcity is part of a broader resource imbalance in Indonesian language technology. NusaX provides parallel data and bilingual lexicons for ten low-resource Indonesian languages [4], NusaWrites develops human-annotated corpora for underrepresented languages [5], and NusaCrowd consolidates dispersed Indonesian NLP resources through collaborative collection and quality assessment [6]. Although these initiatives target broader language-resource construction rather than dictionary governance, they reinforce the need for deliberate curation, documented quality control, and sustainable contribution channels for underrepresented languages.

Prior work addresses important parts of this problem. The Jejueo Talking Dictionary demonstrates collaborative online lexicography for revitalization [8], while a Papuan regional dictionary reports development and validation of a regional lexical information system [9]. Community consultation shaped the Akuzipik–English dictionary [10], and the Enggano dictionary adopts a design-science and participatory orientation [11]. Living Dictionaries illustrates co-creative dictionary building [12], and LingoGap combines crowdsourcing with cross-validation, agreement measures, expert verification, and task logs [13]. GUIDE automates semantic-domain dictionary construction [14]; Kuene uses role-based, multi-round approval for Hawaiian neologisms [15]; and the LiFE Suite emphasizes community-centered collection with linguistic, cultural, and metadata context [16]. Collectively, these studies establish digitization, participation, quality control, and provenance as important concerns, but the cited publications do not uniformly document one lifecycle linking proposals, validator decisions, publication, persistent entry history, and contribution-linked rewards.

Table 1 Comparative emphasis of related low-resource lexical and community-resource studies

Study	Primary emphasis reported in the cited publication
Jejueo [8]	Collaborative online lexicography for language revitalization
Papuan regional dictionary [9]	Regional digital-dictionary development and validation
Akuzipik, Enggano, Living Dictionaries [10]–[12]	Community consultation and participatory/co-creative dictionary building
LingoGap [13]	Crowdsourcing with cross-validation, agreement, expert verification, and task logs
GUIDE [14]	Automated semantic-domain dictionary construction for low-resource languages
Kuene [15]	Role-based, multi-round approval for Hawaiian neologisms
LiFE Suite [16]	Community collection preserving linguistic, cultural, and metadata context
Present study	Controlled proposals, validator publication, before/after provenance, contribution-linked incentives separated from lexical authority, and traceable functional plus supplementary evaluation

Table 1 summarizes the emphasis explicitly reported by the cited studies. It is used to compare documented design emphases rather than to infer that a feature not described in a publication is absent from every version of the corresponding platform.

The related work can be viewed across three information-system layers. Community contribution is central to collaborative and crowd-oriented projects [8], [10], [12], [13], [16]. Lexical authority is explicit in LingoGap's cross-validation and expert verification [13] and Kuene's role-based approval [15]. Traceability appears through task logs and preserved linguistic, cultural, or metadata context [13], [16]. The present study focuses on coupling these layers within one dictionary workflow, particularly when participation incentives must remain separate from lexical authority.

Gamification adds a related design concern. Points, badges, and leaderboards are common mechanisms in gamified crowdsourcing [17], and language-resource studies have used gamification for dialogue collection, idiom corpora, linguistic association data, and word-sense disambiguation [18]–[21]. In the present artifact, points reward selected verified activities, while accept/reject authority remains restricted to validators and administrators. The supplementary user study measures perceived usability and acceptance but does not isolate causal effects of gamification on motivation, learning, retention, adoption, or revitalization.

Artifact-oriented information-systems research also requires evaluation that separates implemented functionality from user or operational claims. DSRM structures artifact construction and evaluation [22], FEDS emphasizes explicit evaluation goals and strategies [23], ISO/IEC 25010:2023 defines Functional Suitability [24], and ISO/IEC/IEEE 29119-4:2021 supports systematic test design [25]. These sources support requirement-traceable functional evaluation but do not establish user acceptance, security, or scalability. Accordingly, the study treats SUS/TAM, OWASP ZAP, and k6 as supplementary evidence streams rather than as extensions of the Functional Suitability score.

The research gap is therefore one of integration and traceability, not a claim that digital dictionaries, crowdsourcing, validation, gamification, or individual governance mechanisms are new. The contribution lies in coupling controlled non-public proposals, validator-controlled publication, persistent before/after lexical provenance, and participation rewards that do not grant lexical authority, then evaluating the resulting artifact through traceable functional testing and separate user, security, and load assessments. This is a documented integration claim rather than a claim of global firstness. Bidirectional retrieval is treated here as a supporting subsystem; detailed retrieval-algorithm evaluation is reserved for a companion study.

3 Research Method

3.1 Design Science Research Methodology

The study adopted Design Science Research Methodology (DSRM) because its principal output is an implemented information-system artifact [22]. FEDS further emphasizes explicit evaluation goals and strategies for design-science artifacts [23]. Problem identification established the need for searchable Toulour lexical access with governed community maintenance; solution objectives specified approved-content access, controlled contribution, provenance, incentives, and role-based governance. Design and development produced the data model, permissions, workflows, interfaces, and PHP/MariaDB system. Demonstration exercised representative new-word and correction lifecycles, evaluation examined the frozen artifact, and communication is represented by this article and its supporting evidence.

For RQ1 and RQ2, the primary unit of analysis remained the operational artifact: whether authorized functions were available, invalid or unauthorized actions were rejected, and successful actions produced the intended state. RQ3 added complementary evidence through a short user pilot, automated application-surface scanning, and staged HTTP load testing. These procedures are interpreted descriptively and do not imply causal motivation effects, security certification, or population-level adoption.

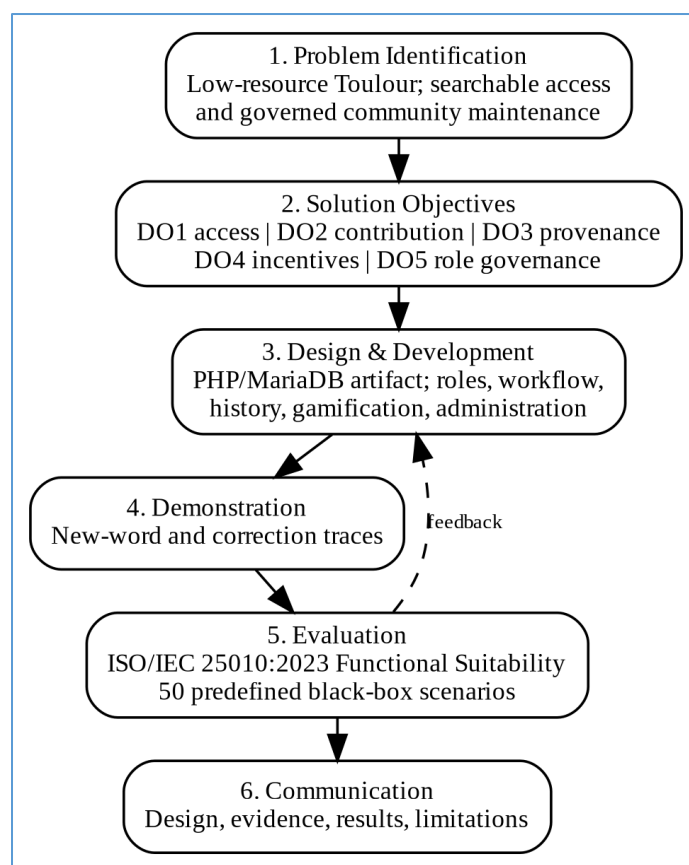


Figure 1 DSRM activities applied to the toulour digital dictionary study

Figure 1 summarizes the primary artifact-oriented DSRM cycle. The supplementary user, security, and load evaluations described in Sections 3.7–3.9 provide complementary evidence for RQ3 and are not represented as additional DSRM phases.

3.2 Design Objectives and Artifact Scope

Five design objectives were fixed before functional interpretation. Table 2 links each objective to its principal artifact mechanism and functional evidence. The supplementary evaluations in Sections 3.7–3.9 are cross-cutting checks rather than one-to-one measures of individual design objectives.

Table 2 Design objectives and evaluation evidence

Objective	Artifact mechanism	Primary evaluation evidence
DO1	Approved bidirectional lexical access and browsing	S01–S08
DO2	Controlled new-word and correction proposals	C01–C09
DO3	Validator decisions and persistent lexical provenance	V01–V10
DO4	Participation-linked XP, quiz, badges, and leaderboard	G01–G07
DO5	Role-based administration, CSV exchange, and feedback	D01–D08, F01–F03

3.3 Source Data, Actors, and Implementation

The lexical base was adapted from Kamus Tondano–Indonesia [2]. Lexical records were represented using a Toulour headword, Indonesian meaning, optional example and description fields, and category metadata. The evaluated baseline contained a frozen snapshot of 4,584 approved lexical

<http://sistemasi.ftik.unisi.ac.id>

entries. Community proposals were kept distinguishable from this approved baseline so that later additions and corrections could remain attributable.

Four operational actor states were defined. A visitor can search and browse approved entries without authentication. A registered user can manage a profile, propose a new word or correction, inspect proposal status, answer the daily quiz, accumulate points and badge tiers, view the leaderboard, and submit feedback. A validator can inspect and decide pending lexical proposals. An administrator can maintain dictionary and category data, manage users and validator roles, process feedback, and import or export CSV data. Visitor is a public interaction state; registered user, validator, and administrator are persisted as role-controlled accounts.

The artifact was implemented as a server-rendered PHP application with MVC-like separation across routing, controllers, models/services, views, and database access; MariaDB supplies persistence. The inspected codebase contains 15 controllers and 27 PHP views. Benchmark and gold-standard modules used by the companion retrieval study were excluded from this article's functional scope. Figure 2 summarizes the operational architecture.

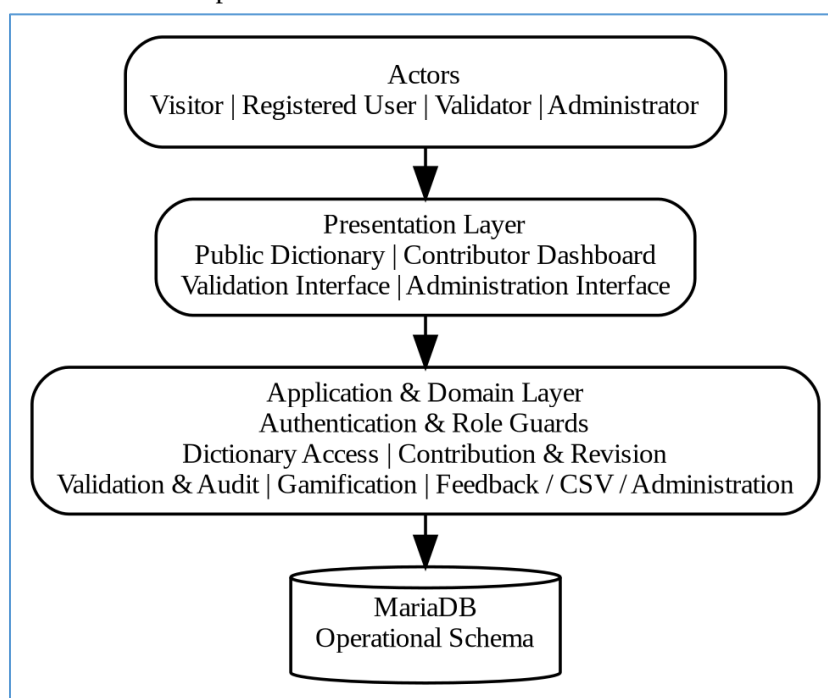


Figure 2 Role-oriented architecture and major operational layers of the artifact

3.4 Controlled Crowdsourcing, Provenance, and Gamification

Registered users can submit new lexical items or corrections. Proposals remain contributor-linked pending records and are excluded from approved public results; corrections also retain a link to the approved original. Validators review lexical and contextual fields before accepting or rejecting a proposal. Approval publishes a new word or applies an accepted correction while preserving before/after information in dictionary history; rejection preserves the decision trace without changing approved content. This reconstructable provenance is the operational meaning of auditability here, not a claim of immutable-ledger technology or formal security certification.

Gamification remains downstream of verified activities. The first successful daily login awards 1 XP, a correct daily quiz 2 XP, an approved new word 10 XP, an approved correction 5 XP, and a rejected proposal 0 XP. Point and quiz histories prevent duplicate credit, while accumulated XP maps to Tou (0–25), Kawanua (26–75), Taranak (76–150), and Tonaas (151+). Rewards affect recognition but not role authorization or lexical authority.

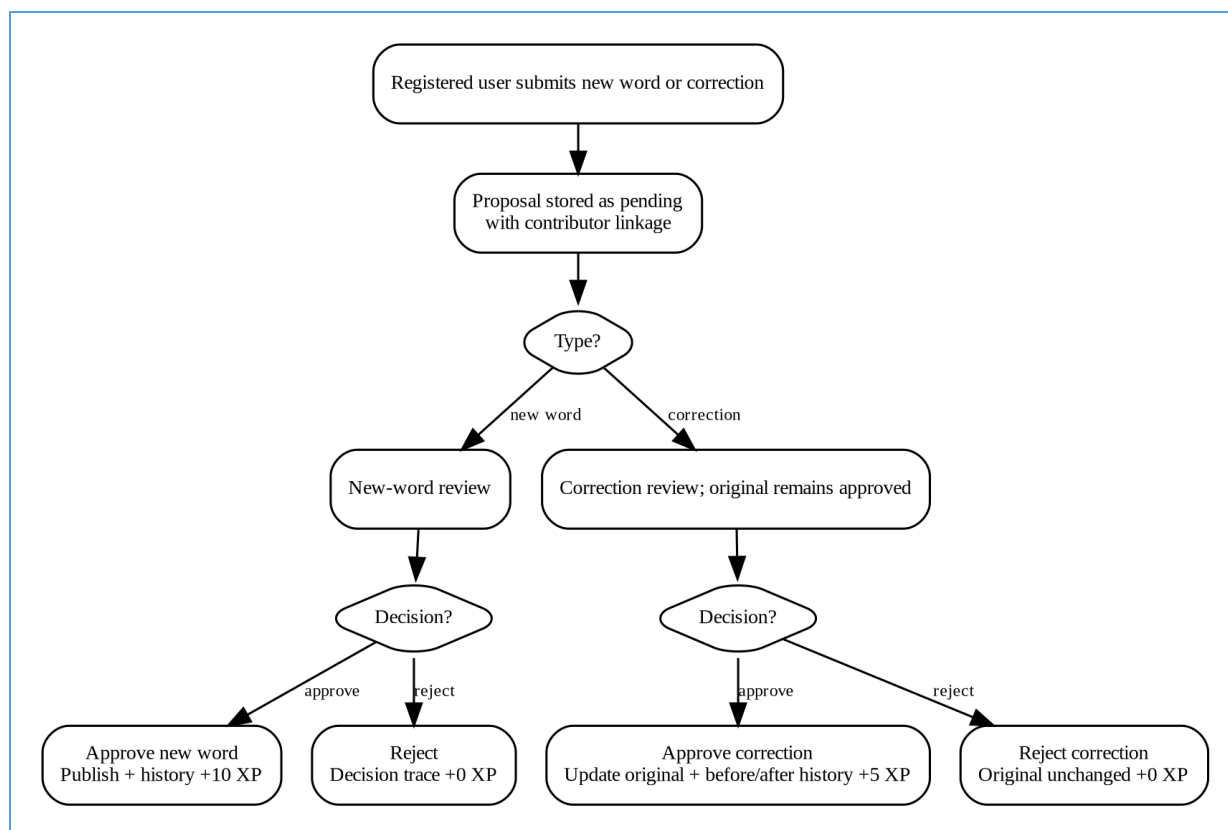


Figure 3 Controlled crowdsourcing, validation, provenance, and reward flow

3.5 Reduced Data Model and Demonstration

The reduced operational data model in Figure 4 focuses on the entities directly supporting governance and participation: roles, users, categories, dictionary, dictionary_history, user_points_history, user_quiz_history, and feedbacks. The dictionary table supports both approved entries and temporary proposals through status, contributor, edit, and original-entry linkage fields. The history table retains affected dictionary identifiers, validator identity, lexical before/after data, notes, and timestamps. Separate participation ledgers preserve points and quiz events.

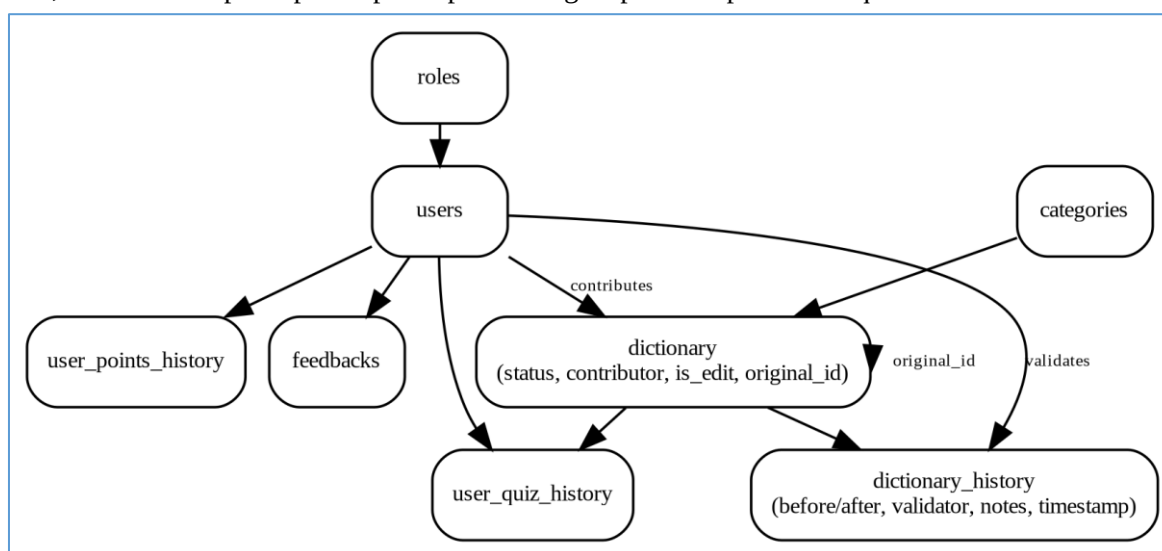


Figure 4 Reduced operational data model for governance and participation

Demonstration used two end-to-end traces. The first followed a new lexical item from contributor submission through pending status, validator approval, publication, history recording, and reward. The second followed a post-publication correction: the approved original remained public while the correction was pending; after approval, the accepted enrichment was applied, before/after information remained retrievable, and the correction reward was recorded. These traces demonstrated the intended link among contribution, decision, public state, provenance, and incentives.

3.6 Functional Suitability Evaluation

The ISO-based product-quality evaluation was restricted to Functional Suitability in ISO/IEC 25010:2023 [24], interpreted through functional completeness, correctness, and appropriateness. Completeness concerned availability of required functions; correctness concerned whether outputs, guards, and persistent states matched predefined expectations; and appropriateness concerned whether each role could complete its intended task without bypassing governance. The standard structures interpretation of functional evidence rather than producing a proprietary “ISO score.” User perception, security, and load behavior were evaluated separately.

Black-box execution used 50 predefined scenario-based cases with observable UI or persistent outcomes, consistent with ISO/IEC/IEEE 29119-4:2021 [25]. Cases covered search/browsing (8), account/profile (5), proposals (9), validation/history (10), gamification (7), administration/CSV (8), and feedback/authorization (3), including valid, invalid, duplicate, unauthorized, boundary, and state-transition conditions. Expected outcomes were fixed before execution.

Following FEDS, the final functional run was an artificial, summative evaluation of specified artifact properties [23]. The developer-researcher executed the locked protocol on one frozen build and database state. Evaluator-bias risk was mitigated through predefined expectations, observable UI/database evidence, role-specific preconditions, fixed PASS/FAIL rules, and retention of the failed case.

3.7 Supplementary User-Perspective Evaluation

A pilot user-perspective evaluation used an online questionnaire after participants tried the application. Eligibility required voluntary consent, age ≥ 18 years, and prior use of the Toulour Digital Dictionary during the evaluation. Forty submissions were received and 34 met the criteria. The valid sample spanned the predefined adult age bands from 18–24 to ≥ 55 years, several occupational groups, and varying familiarity with Toulour.

Usability was measured with the 10-item System Usability Scale (SUS), scored by the standard alternating transformation to a 0–100 total [26]. Acceptance was summarized through TAM-derived perceived usefulness (PU; four items), perceived ease of use (PEOU; four items), and behavioral intention (BI; three items) on five-point agreement scales [27]. Analysis was descriptive (n, mean, standard deviation, median, and range), with exact-wording counts for open responses. Exploratory internal-consistency estimates were used only as a limitation check; no SEM, hypothesis testing, or population-level inference was performed.

3.8 Supplementary Security Assessment

A supplementary automated application-surface security assessment targeted the local XAMPP deployment using OWASP ZAP 2.17.0 [28]. The workflow used the Dev Standard active-scan policy and traditional spider. Alerts were treated as indicators requiring verification, then triaged through source-code inspection, targeted hardening, clean-session repeat scans, and final context-scoped verification of the project application. Because the diagnostic baseline used broader localhost scope than the final report, severity-count changes are not interpreted as a pure remediation effect. This was not a manual penetration test or security certification.

3.9 Supplementary Local Load Evaluation

Local scalability was explored with Grafana k6 using a protocol-level, read-heavy HTTP workload [29]. After a 20-second warm-up at 1 virtual user (VU), non-overlapping stages used 1, 10, 25, and 50 VUs for 60 seconds each with a 20-second graceful-stop allowance. Each iteration issued GET requests to the homepage, search API, leaderboard, and about page, followed by one second of think time. The final run recorded average and p95 response time, throughput, and HTTP error rate. Results describe a short-duration local XAMPP assessment, not production capacity.

4 Results and Analysis

4.1 Implemented Artifact and Governance Behavior

The frozen artifact exposed distinct public, contributor, validator, and administrative interfaces. Approved entries were searchable, while pending and rejected proposals were excluded from public lexical content. Contributors could submit and track proposals, use profile and gamification functions, and provide feedback; validators handled pending decisions and history; administrators maintained trusted dictionary content, categories, users/roles, feedback, and CSV exchange.

The central governance result was preservation of the proposal-to-publication boundary. Submission or accumulated XP did not confer public lexical status. Corrections left the approved entry public until a validator accepted the proposed revision, thereby operationalizing the separation of participation incentives from lexical authority posed in RQ1.

4.2 Auditable Crowdsourcing and Participation Incentives

An observed trace using the entry *leset* made the lifecycle concrete. The new word was approved as *leset* (“tidak longgar”) at 18:38 WITA on 27 Aug 2026. A later correction retained the headword and meaning but proposed an example and contextual explanation. It remained separately attributable and pending while the approved entry stayed locked, then entered the validator queue as *Edit Kata* (Figure 5).

(a)

Kata dalam Bahasa Toulour *

Masukkan kata dasar atau kata berimbuhan asli suku Toulour.

Kategori Kata *

Arti / Terjemahan (Bahasa Indonesia) *

Contoh Penggunaan dalam Kalimat Toulour (Opsional)

Deskripsi / Penjelasan Kontekstual (Opsional)

(b)

Kata Toulour	Arti Bahasa Indonesia	Kategori	Status Validasi	Catatan Validator	Tanggal Usulan	Aksi
leset	tidak longgar	Kata Keterangan	⌛ MENUNGGU	—	27 Aug 2026, 19:00	✎ Edit
leset	tidak longgar	Kata Keterangan	✅ DISETUJUI	—	27 Aug 2026, 18:28	🔒 Terkunci

(c)

✎ Edit Kata

User A

✅ Setujui ❌ Tolak

Figure 5 Correction proposal and validation trace for *leset*: proposal fields, contributor states, and validator queue

Figure 5 shows that the correction remained contributor-attributable and non-public while pending; only the validator-controlled decision stage could authorize publication.

After approval at 19:01 WITA, word history preserved the prior blank example/explanation fields alongside the newly accepted values, validator identity, and timestamp (Figure 6), so the enrichment remained reconstructable rather than simply overwriting provenance.

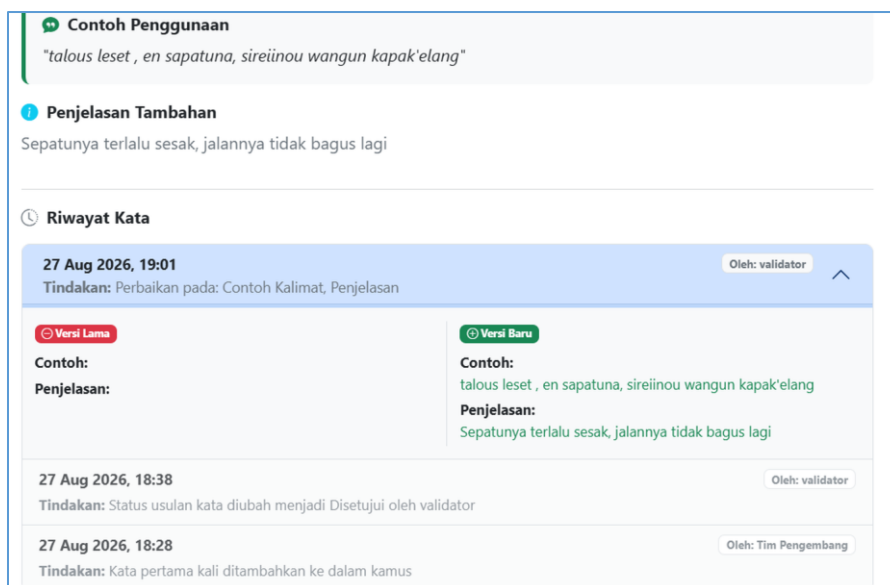


Figure 6 Accepted correction and preserved before/after provenance for leset

Figure 6 confirms that accepted enrichment remains inspectable together with the earlier field state, validator identity, and timestamp.

Gamification was linked to verified events rather than validation authority. The observed ledger recorded +10 XP for the approved new word, +5 XP for the approved correction, +1 XP for daily login, and +2 XP for a correct daily quiz. The 50-XP account was assigned to Kawanua and appeared in an XP-ranked leaderboard (Figure 7). These observations establish rule implementation, not effects on motivation, learning, retention, or contribution frequency.

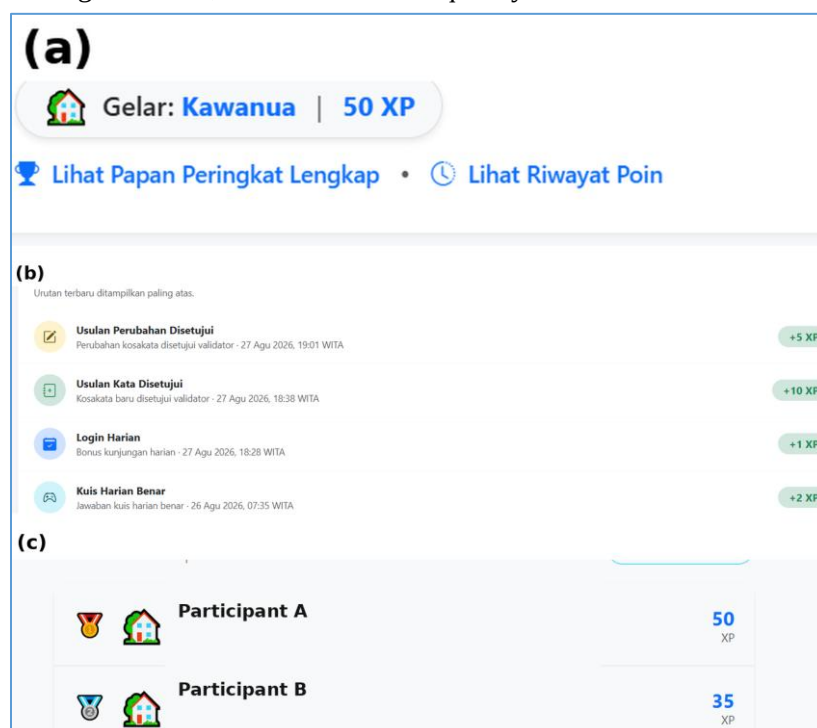


Figure 7 Gamification evidence: XP-based tier, event-linked rewards, and XP-ranked leaderboard (participant names anonymized)

4.3 Black-Box Functional Results

Table 3 summarizes the frozen 50-case execution: 49 passed and one failed (98.0%). The proportion is test-execution evidence, not an ISO/IEC 25010 numerical quality score. Table 4 highlights representative governance and boundary cases.

Table 3 Functional test results by area

Functional area	Cases	Pass	Fail
Search & browsing	8	7	1
Account & profile	5	5	0
Crowdsourcing proposals	9	9	0
Validation & history	10	10	0
Gamification	7	7	0
Administration & CSV	8	8	0
Feedback & authorization	3	3	0
TOTAL	50	49	1

Table 4 Representative governance and boundary cases

ID	Purpose	Status
C02	Pending proposal non-public	PASS
C09	Proposal ownership guard	PASS
V03	Approval + validator history	PASS
V06	Before/after provenance	PASS
V10	Invalid-decision safety	PASS
G02	Duplicate login reward	PASS
D08	Admin self-protection	PASS
S08	Oversized query	FAIL

The sole failure, S08, combined two search-boundary conditions. Empty input could not be submitted, whereas a query exceeding 150 characters was processed rather than rejected. It terminated safely with the normal not-found state and did not crash the application, but the predefined expectation required rejection; S08 was therefore retained as FAIL. The other seven search/browsing cases passed.

4.4 ISO/IEC 25010 Functional Suitability Interpretation

Functional completeness was supported because all defined areas—lexical access, account/profile, contribution, validation/history, gamification, administration/CSV, feedback, and authorization—were present and exercisable. S08 concerned boundary handling within an available function rather than a missing module.

For functional correctness, 49 scenarios matched predefined outcomes across authorization, state transitions, persistence, reward accounting, and administration. S08 shows that correctness was not perfect at the search-input boundary.

For functional appropriateness, contributors could submit but not self-publish, validators could decide proposals and leave provenance, administrators could maintain trusted content and governance functions, and gamification could recognize participation without altering lexical authority. This supports the intended role workflow but is distinct from user-efficiency or usability evidence.

4.5 User-Perspective Evaluation Results

Of 40 questionnaires, 34 responses (85.0%) met the eligibility criteria; three respondents were younger than 18 and three had not tried the application. Valid respondents covered all five predefined adult age bands and multiple occupational groups. Mean SUS was 77.79 (SD 8.27; median 78.75; range 60.0–92.5). On the 1–5 scale, PU averaged 4.10 (SD 0.36), PEOU 4.07 (SD 0.45), and BI 3.90 (SD 0.43) (Table 5).

Table 5 Pilot user-perspective evaluation summary

Measure	n/items	Mean	SD / range
SUS	34 respondents	77.79 / 100	SD 8.27; 60.0–92.5
Perceived Usefulness (PU)	4 items	4.10 / 5	SD 0.36
Perceived Ease of Use (PEOU)	4 items	4.07 / 5	SD 0.45
Behavioral Intention (BI)	3 items	3.90 / 5	SD 0.43

Open responses most often valued vocabulary search, accessible meanings, examples, and clear results; recurrent improvement requests concerned first-time-user guidance, example searches, and clearer return navigation. Exploratory internal consistency was modest for SUS ($\alpha=0.60$) and PEOU ($\alpha=0.62$) but weak for PU ($\alpha=0.05$) and BI ($\alpha=0.34$). TAM results are therefore treated only as descriptive user-perception indicators.

4.6 Supplementary Security Results

The broad-localhost ZAP baseline reported 19 distinct alert types/groups (1 High, 5 Medium, 8 Low, 5 Informational). The High alert was a medium-confidence SQL Injection indicator; source review found native PDO prepared statements, while malformed numeric identifiers were being coerced. Validation was tightened and the alert was not reproduced in later clean-session scans. Hardening also added global CSRF protection, safer error handling, hardened session-cookie attributes, and browser security headers.

Table 6 Security verification checkpoints

Checkpoint	Scope	High	Medium	Low	Info
Initial baseline	Broad localhost	1	5	8	5
Post-remediation verification	Broad localhost	0	4	3	5
Final verification	Context: project	0	3	3	4

Final context-scoped verification reported 0 High, 3 Medium, 3 Low, and 4 Informational alert types/groups. The Medium findings concerned CSP script-src unsafe-inline, CSP style-src unsafe-inline, and missing Subresource Integrity attributes. The absence of a High alert does not establish that the application is vulnerability-free; results are limited to the discovered endpoints, ZAP configuration, local environment, and test session.

4.7 Supplementary Local Load Results

The final non-overlapping k6 run reached 50 VUs with 1,128 completed and 0 interrupted iterations; no HTTP request failures were observed. Table 7 shows that throughput increased strongly from 1 to 25 VUs but only modestly from 25 to 50 VUs while response time rose sharply at the highest load.

Table 7 K6 results by measured load level

VU	Requests	Avg response	p95 response	Throughput	Error rate
1	176	96.20 ms	252.31 ms	2.93 req/s	0.00%
10	1,052	322.98 ms	1,210.81 ms	17.53 req/s	0.00%
25	1,568	732.79 ms	2,669.23 ms	26.13 req/s	0.00%
50	1,656	1,683.25 ms	6,390.41 ms	27.60 req/s	0.00%

At 25 VUs, throughput was 26.13 requests/s, average response time 732.79 ms, and p95 2,669.23 ms. At 50 VUs, throughput rose only 5.6% to 27.60 requests/s while average response time reached 1,683.25 ms and p95 6,390.41 ms. This pattern is consistent with increasing contention or saturation in the tested local deployment, not an absolute production-capacity limit.

4.8 Discussion, Comparison, and Threats to Validity

The leset trace illustrates why the contribution is an integration claim rather than a collection of isolated features: one lexical object moved through contributor submission, validator-controlled state change, public presentation, later enrichment, before/after history, and reward logging. Prior studies already report role-based approval and other relevant components, including Kuene's multi-round approval [15] and work on digitization, community consultation, collaborative dictionary building, automated construction, contextual provenance, and crowd validation [7]–[14], [16]. The contribution here is their application-level coupling with participation incentives that remain separate from lexical authority and with traceable functional plus supplementary evaluation. It is not a claim of global firstness.

The separation of incentives from lexical authority allows verified participation to be recognized without turning activity volume into an acceptance criterion. The user pilot adds a complementary perspective: SUS and descriptive TAM indicators show favorable perceptions among the evaluated adults, while open responses identify onboarding and navigation improvements. Variation in age, occupation, and Toulour familiarity broadens the evidence beyond developer-only testing, but 34 participants cannot establish population-wide acceptance, long-term adoption, or a causal gamification effect.

Threats to validity remain. Functional testing was executed by the developer-researcher, although predefined expectations, frozen scenarios, observable evidence, and the retained failure reduce evaluator-bias risk. The user evaluation was a short-exposure, 34-person pilot rather than a representative community study, and weak PU/BI internal consistency limits TAM interpretation to descriptive indicators. ZAP assessed a local XAMPP application surface rather than conducting a manual penetration test; baseline and final scans also differed in scope breadth. The k6 assessment was short, read-heavy, local, and based on one clean final run without distributed generation or resource telemetry, so it describes a latency/throughput trend rather than production capacity. The study also concerns one language pair and one architecture; broader transfer requires larger, longer, and production-context deployment evidence.

5 Conclusion

This study applied DSRM to design and evaluate an auditable crowdsourced and gamified Toulour–Indonesian digital dictionary. The artifact combines approved lexical access, contributor proposals, validator-controlled decisions, persistent revision provenance, participation-linked XP, quizzes, badge tiers, feedback, and administration. RQ1 is answered by separating participation incentives from lexical authority: users may propose content and receive recognition, but only authorized validation decisions determine approved public data. Comparison with prior low-resource lexical platforms positions this as an integration-and-traceability contribution rather than novelty of crowdsourcing, role-based validation, provenance, or gamification individually.

For RQ2, 49 of 50 predefined black-box cases passed; S08 retained an oversized-query boundary limitation. Under ISO/IEC 25010:2023 Functional Suitability, the evidence supports the defined functional scope and its governance workflows without claiming a proprietary ISO score. For RQ3, the 34-person pilot yielded mean SUS 77.79 and descriptive PU, PEOU, and BI means of 4.10, 4.07, and 3.90. Final application-scoped ZAP verification observed no High-risk alert but retained three Medium findings, while staged k6 testing observed no HTTP request failures through 50 VUs but increasing latency and flattening throughput at higher concurrency. These supplementary findings remain limited to a pilot sample and local deployment. Future work should include larger and longer Toulour-community studies, accessibility evaluation, production-hosted/manual security testing, and longer or distributed performance testing with resource telemetry.

Acknowledgement

The authors acknowledge the compilers of Kamus Tondano–Indonesia, the language validators, and members of the Toulour community who support lexical preservation. This work was supported and funded by The Institute for Research and Community Service (LPPM), Sam Ratulangi University, for the 2026 fiscal year under Rector’s Decree Letter No. 543/UN12.27/LT/2026 and Research Grant Contract Agreement No. 543.

References

- [1] S. Bird, “*Local Languages, Third Spaces, and Other High-Resource Scenarios*,” in Proc. 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 7817-7829, DOI: 10.18653/v1/2022.acl-long.539.
- [2] A. Wantalangi, G. J. Tendean, S. Wuntu, A. Kawengian, and E. Kaligis, Kamus Tondano-Indonesia. Jakarta, Indonesia: Pusat Pembinaan dan Pengembangan Bahasa, 1985.
- [3] T. C. Brickell, Tondano (Toundano): *A Grammar Sketch of an Endangered Minahasan Language*. Routledge, 2022, DOI: 10.4324/9780429508165.
- [4] G. I. Winata et al., “*NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages*,” in Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 815-834, DOI: 10.18653/v1/2023.eacl-main.57.
- [5] S. Cahyawijaya et al., “*NusaWrites: Constructing High-Quality Corpora for Underrepresented and Extremely Low-Resource Languages*,” in Proc. IJCNLP-AACL, 2023, pp. 921-945, DOI: 10.18653/v1/2023.ijcnlp-main.60.
- [6] S. Cahyawijaya et al., “*NusaCrowd: Open Source Initiative for Indonesian NLP resources*,” in Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 13745-13818, DOI: 10.18653/v1/2023.findings-acl.868.
- [7] M. Maxwell and A. Bills, “*Endangered Data for Endangered Languages: Digitizing Print Dictionaries*,” in Proc. 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2017, pp. 85-91, DOI: 10.18653/v1/W17-0112.
- [8] M. Saltzman, “*Jejueo Talking Dictionary: A Collaborative Online Database for Language Revitalization*,” in Proc. 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2017, pp. 122-129, DOI: 10.18653/v1/W17-0117.
- [9] N. F. Hasan and M. S. A. Iribaram, “*Digitalisasi Kamus Bahasa Daerah Papua menggunakan Metode Rapid Application Development*,” MATRIK, Vol. 21, No. 3, pp. 710-720, 2022, DOI: 10.30812/matrik.v21i3.1688.
- [10] B. Hunt, L. Schwartz, S. Schreiner, and E. Chen, “*Community Consultation and the Development of an online Akuzipik-English Dictionary*,” in Proc. AmericasNLP, 2023, pp. 134-143, DOI: 10.18653/v1/2023.americasnlp-1.14.
- [11] C. Pramatha, G. P. W. Rajeg, M. Dalrymple, and I. W. Arka, “*Preserving the Enggano language: A Digital Dictionary Approach*,” in Proc. Australasian Conf. Information Systems, 2024, paper 9.
- [12] A. L. Daigneault and G. Anderson, “*Citizen Linguists and Decolonial Lexicography: Co-Creative Dictionary-Building in Grassroots Digital Language Documentation*,” in Proc. 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2025, pp. 20-28.
- [13] H. Khalilia, J. Otterbacher, G. Bella, S. Darma, and F. Giunchiglia, “*Crowdsourcing Lexical Diversity*,” Frontiers in Artificial Intelligence, Vol. 8, 2025, DOI: 10.3389/frai.2025.1648073.
- [14] J. Janetzki, G. De Melo, J. Nemecek, and D. Whitenack, “*GUIDE: Creating Semantic Domain Dictionaries for Low-Resource Languages*,” in Proc. 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP), 2024, pp. 10-24, DOI: 10.18653/v1/2024.sigtyp-1.2.
- [15] S. Walker, W. Wu, B. Torres Fischer, and L. Kimura, “*Kuene: A Web Platform for Facilitating Hawaiian Word Neologism*,” in Proc. 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages, 2025, pp. 182-187. [Online]. Available: <https://aclanthology.org/2025.computel-main.21/>
- [16] K. Narayanan R et al., “*Field to Model: Pairing Community Data Collection with Scalable NLP through the LiFE Suite*,” in Proc. 4th Workshop on NLP Applications to Field Linguistics, 2025, pp. 76-84. [Online]. Available: <https://aclanthology.org/2025.fieldmatters-1.7/>

- [17] B. Morschheuser, J. Hamari, J. Koivisto, and A. Maedche, “Gamified Crowdsourcing: Conceptualization, Literature Review, and Future Agenda,” *International Journal of Human-Computer Studies*, Vol. 106, pp. 26-43, 2017, DOI: 10.1016/j.ijhcs.2017.04.005.
- [18] H. Ogawa, H. Nishikawa, T. Tokunaga, and H. Yokono, “Gamification Platform for Collecting Task-Oriented Dialogue Data,” in *Proc. 12th Language Resources and Evaluation Conference*, 2020, pp. 7084-7093.
- [19] G. Eryiğit, A. Şentaş, and J. Monti, “Gamified Crowdsourcing for Idiom Corpora Construction,” *Natural Language Engineering*, Vol. 29, No. 4, pp. 909-941, 2023, DOI: 10.1017/S1351324921000401.
- [20] F. Genovese, M. M. Bolognesi, A. Di Iorio, and F. Vitali, “The Advantages of Gamification for Collecting Linguistic Data: A Case Study using Word Ladders,” *Online Journal of Communication and Media Technologies*, Vol. 14, No. 2, e202426, 2024, DOI: 10.30935/ojcm/14443.
- [21] D. Torunoğlu Selamet, A. Şentaş, and G. Eryiğit, “Gamified Crowd-Sourcing for Word Sense Disambiguation of Turkish,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 24, No. 11, Art. No. 130, pp. 1-23, 2025, DOI: 10.1145/3767160.
- [22] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *Journal of Management Information Systems*, Vol. 24, No. 3, pp. 45-77, 2007, DOI: 10.2753/MIS0742-1222240302.
- [23] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A Framework for Evaluation in Design Science Research,” *European Journal of Information Systems*, Vol. 25, No. 1, pp. 77-89, 2016, DOI: 10.1057/ejis.2014.36.
- [24] ISO/IEC 25010:2023, *Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Product Quality Model*, 2nd ed. Geneva, Switzerland: International Organization for Standardization, 2023.
- [25] ISO/IEC/IEEE 29119-4:2021, *Software and Systems Engineering—Software Testing—Part 4: Test Techniques*, 2nd ed. Geneva, Switzerland: International Organization for Standardization, 2021.
- [26] J. Brooke, “SUS: A ‘Quick and Dirty’ Usability Scale,” in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds. London, U.K.: Taylor & Francis, 1996, pp. 189–194.
- [27] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *MIS Quarterly*, Vol. 13, No. 3, pp. 319–340, 1989, DOI: 10.2307/249008.
- [28] ZAP by Checkmarx, “ZAP 2.17.0,” 2025. [Online]. Available: <https://www.zaproxy.org/docs/desktop/releases/2.17.0/>
- [29] Grafana Labs, “Grafana k6 Documentation: Load Testing Websites,” [Online]. Available: <https://grafana.com/docs/k6/latest/testing-guides/load-testing-websites/>