A Scalable Computational Pipeline for Collaborative Translation and Validation of Biomedical Ontologies: The PORTI-HPO Platform

Filipe Bernardi[1,2,3,7][0000-0002-9597-5470], Julio Souza[4,5][0000-0002-8576-1903], Claudia Lorea[7][0000-0002-4653-3051], Bibiana de Oliveira[6,7][0000-0002-2679-6858], Victor Ferraz[1,7][0000-0003-0337-4588], Temis Felix[7][0000-0002-8401-6821], and Domingos Alves[1,7][0000-0002-0800-5872]

[1] Ribeirao Preto Medical School, University of Sao Paulo, Ribeirao Preto, Brazil
[2] RISE-Health, Faculty of Medicine, University of Porto, Porto, Portugal
[3] Institute for Systems and Computers Engineering at Coimbra, Coimbra, Portugal
[4] School of Engineering - ISEP, Polytechnic Institute of Porto, Porto, Portugal
[5] GECAD - Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development, Porto, Portugal
[6] Federal University of Health Sciences of Porto Alegre (UFCSPA), Porto Alegre, Brazil
[7] National Rare Diseases Network, Porto Alegre Clinical Hospital, Porto Alegre, Brazil
fbernardi@med.up.pt

Abstract. Biomedical ontologies underpin standardized phenotype descriptions for rare disease research and clinical decision support, yet their availability predominantly in English limits adoption across non-anglophone populations. We present PORTI-HPO, a production computational platform that combines large language model (LLM) draft generation with distributed multi-expert validation to translate the Human Phenotype Ontology (HPO) into Portuguese for 280 million speakers across nine lusophone countries. The system implements a five-stage pipeline integrating automated term ingestion, LLM-based draft translation, role-based human validation, and multi-format export generation including FHIR R4 terminology services. Quality assurance uses translation confidence scores (0.0–1.0), five-star ratings, and automated lexical consistency checks, supported by 83 passing automated tests. Deployed at `https://hpo.raras-cplp.org` with PostgreSQL persistence and OAuth2 authentication via ORCID and LinkedIn, the platform catalogues 17,020 HPO terms. Preliminary evaluation of 150+ priority terms (approximately 0.9% of the current HPO corpus) reports 0.72 mean confidence, 4.1-star quality, acceptance rates of 82% (translators) and 94% (validators), and 19-day median turnaround, a 37% improvement relative to literature-based manual estimates. Current validation is internal to the platform (role-based expert review), while external validation remains ongoing; planned independent back-translation will serve as a

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

benchmark for semantic fidelity, together with formal inter-rater agreement analysis. These results characterize an initial subset, and broader corpus coverage is required before stronger claims about scalability and robustness.

Keywords: Biomedical ontology translation · Large language models · Human-in-the-loop systems · Crowdsourcing quality assurance · FHIR interoperability · Rare diseases

1 Introduction

The Human Phenotype Ontology (HPO) provides computationally tractable phenotype descriptions essential for genomic diagnostics, clinical decision support, and rare disease research [1]. With over 17,000 standardized terms describing phenotypic abnormalities, HPO enables interoperability across patient registries, electronic health records (EHRs), and international research networks including Online Mendelian Inheritance in Man (OMIM) [3] and Orphanet [4]. However, HPO remains predominantly English-language, creating accessibility barriers for the Community of Portuguese-Speaking Countries (CPLP) encompassing 280 million speakers across diverse clinical and socioeconomic contexts spanning Portugal, Brazil, Angola, Mozambique, and five additional nations.

Ad hoc manual translations introduce semantic inconsistencies and compromise data quality in multilingual clinical workflows. Traditional translation methodologies such as ISPOR guidelines [5], while ensuring terminological fidelity, require 30–60 days per ontology and scale poorly for rapidly evolving resources comprising thousands of interconnected terms. Conversely, unsupervised machine translation risks propagating errors in safety-critical clinical contexts where phenotype misclassification directly impacts diagnostic accuracy [6]. Recent advances in large language models (LLMs) demonstrate potential for biomedical concept recognition [7], yet clinical terminology demands human validation to ensure semantic precision and cultural appropriateness.

1.1 Computational Challenges in Ontology Localization

Translating biomedical ontologies presents distinct computational challenges beyond text translation. First, semantic consistency must be maintained across hierarchical term relationships where parent-child dependencies constrain valid translations. Second, provenance tracking requires immutable audit logs linking translated terms to contributor identifiers, review timestamps, and conflict resolution decisions. Third, quality assurance necessitates quantitative metrics (inter-rater agreement, confidence scores, acceptance rates) to guide task prioritization and reviewer allocation. Fourth, version synchronization with source ontology releases benefits from automated diff detection and incremental update workflows that reduce translation drift. Fifth, interoperability requires export

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

generation in ontology-specific formats (OWL/RDF, OBO) and healthcare integration standards (FHIR [8], HL7 terminology services). Many of these requirements are only partially supported by general-purpose crowdsourcing platforms, motivating specialized computational infrastructure.

1.2 Contributions

This work presents PORTI-HPO, a production-deployed platform at `https://hpo.raras-cplp.org` for ontology localization via automated preprocessing, LLM-assisted drafting, and distributed human validation. We report preliminary results from 150+ translated terms. Specific contributions include:

1. **Five-stage production pipeline**: Automated HPO ingestion, LLM draft generation, role-based validation with escalation logic, and multi-format export including FHIR R4 CodeSystem resources, with 83 automated tests covering core components.
2. **Operational quality framework**: Quantitative indicators combining confidence scores (0.0–1.0), five-star ratings, acceptance-rate tracking, and automated lexical consistency checks for style and regional variants.
3. **Preliminary validation evidence**: Results from 150+ priority terms (approximately 0.9% of HPO) indicate 0.72 mean confidence, 4.1-star quality, 82–94% acceptance rates, and 19-day median turnaround (37% faster relative to literature-informed manual workflow estimates); broader corpus coverage is still required for stronger scalability and robustness claims.
4. **Interoperability architecture**: FHIR R4 terminology endpoints exposing HPO-PT as CodeSystem resources, five export formats (CSV, JSON, XLIFF, Babelon TSV with ORCID attribution, quality metrics TSV), and OAuth2 federated authentication via ORCID [9] and LinkedIn.
5. **Reproducible workflow and evaluation path**: Documentation of pipeline stages, quality thresholds, and technical architecture to support replication for other ontologies and languages, with external validation planned via independent back-translation and formal inter-rater agreement analysis.

The remainder of this paper is organized as follows. Section 2 reviews biomedical ontology translation methods, crowdsourcing quality assurance, and healthcare interoperability standards. Section 3 describes overall system architecture and deployment. Section 4 details the computational pipeline and quality metrics. Section 5 reports validation outcomes from 150+ translated terms. Section 6 analyzes implications and limitations, followed by conclusions and reproducibility details.

2 Background and Related Work

2.1 Human Phenotype Ontology in Rare Disease Research

The HPO consortium maintains a structured vocabulary of 17,020+ phenotype terms with hierarchical is-a relationships enabling computational reasoning over

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

clinical observations [1]. Originally developed by Robinson et al. [2], HPO facilitates genomic variant prioritization by matching patient phenotypes to gene-disease associations curated from literature and databases including OMIM [3], Orphanet [4], and DECIPHER. Integration with Fast Healthcare Interoperability Resources (FHIR) terminology services [8] enables EHR phenotype capture and clinical decision support system (CDSS) interoperability.

2.2 Biomedical Terminology Translation Methods

Traditional translation protocols follow ISPOR guidelines [5] requiring forward translation by bilingual experts, back-translation by independent translators, and cognitive debriefing with target populations. While ensuring conceptual equivalence, such workflows require weeks per instrument and scale poorly for ontologies comprising thousands of terms. Early machine translation efforts for biomedical texts [17] demonstrated syntactic accuracy but semantic errors compromising clinical utility. Wolk and Wolk [18] applied neural MT to HPO Polish translation, achieving 85% accuracy requiring post-editing. Recent LLM applications to phenotype concept recognition [7] show promise but necessitate validation frameworks preventing hallucination propagation in clinical contexts.

2.3 Crowdsourcing and Quality Assurance in Bioinformatics

Crowdsourcing enables parallel task execution across distributed contributors, demonstrated effective for protein function annotation, gene variant classification, and literature curation [10]. Quality control mechanisms include gold-standard test questions, peer review voting, and reputation systems rewarding consistency [11]. Gamification through points, leaderboards, and achievements sustains engagement in medical education [12] and health data collection [13]. However, biomedical crowdsourcing demands domain expertise balancing contributor accessibility with output accuracy, motivating hybrid workflows combining automated preprocessing with expert validation.

2.4 FAIR Principles and Biomedical Ontology Interoperability

The FAIR principles [14] mandate Findable, Accessible, Interoperable, and Reusable data infrastructures. For ontologies, interoperability requires standardized formats including Web Ontology Language (OWL/RDF) for semantic reasoning and FHIR terminology services [8] for healthcare system integration. Phenopackets [15] provide GA4GH-standardized data exchange for genomic diagnostics, requiring multilingual phenotype term support enabling international rare disease networks. Babelon format extends ontology translation with provenance metadata including contributor ORCID identifiers and confidence metrics, facilitating quality assessment and attribution.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

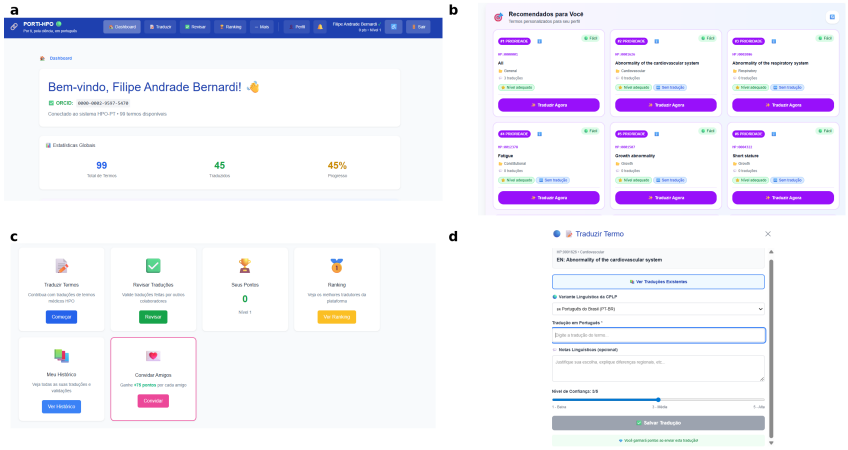

Fig. 1. PORTI-HPO system interface: (a) term browser with search, category filters, and status filters; (b) term detail view with English definition, synonyms, and candidate Portuguese translations with confidence and ratings; (c) validation workspace showing reviewer decisions, rating controls, and escalation status; (d) contributor dashboard highlighting gamification metrics (points, level, badges) and recent activity.

3 System Overview: PORTI-HPO Platform

PORTI-HPO provides an end-to-end platform for collaborative HPO translation accessible at `https://hpo.raras-cplp.org`. The system architecture comprises five logical layers: data ingestion, preprocessing, automated translation, distributed validation, and artifact publication. Figure 1 summarizes the main user-facing components of the platform as implemented.

3.1 Architecture and Deployment

PORTI-HPO is deployed as a web platform with a modular architecture connecting API services, persistent storage, and a browser-based user interface. The backend exposes REST endpoints for term retrieval, translation submission, review, and export generation, while PostgreSQL stores ontology terms, translations, validation decisions, and audit events. OAuth2 authentication (ORCID and LinkedIn) supports federated sign-in, and role-based access control enforces the six workflow roles (Translator, Reviewer, Validator, Moderator, Committee Member, Administrator). Provenance is maintained through immutable event logging of key actions (actor, timestamp, decision, and payload snapshot), enabling auditability and regulatory traceability [16].

3.2 Data Model and Export Formats

The data model is organized around four core entities: ontology terms, translation candidates, validation records, and audit logs. Terms store HP identifiers

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

and source metadata (label, definition, synonyms, hierarchy), while translation and validation entities capture proposed Portuguese text, confidence/rating signals, reviewer decisions, and adjudication outcomes. This structure supports the role-based validation workflow and operational quality monitoring described in Section 4.

To support interoperability and reuse, the platform exports approved content in five formats: CSV and JSON for analysis/API consumption, XLIFF for localization workflows, Babelon TSV for HPO consortium submission with ORCID attribution [9], and a quality-metrics TSV for transparency and reproducibility. FHIR R4 terminology resources are generated for health-system integration.

3.3 Research Artifact Access

The web platform at `https://hpo.raras-cplp.org` provides public access for term browsing, search, and validated translation download. Authenticated contributors submit translations, review peer submissions, and track contributions via personal dashboards displaying points, levels, badges, and acceptance rates. The term browser supports pagination, free-text search, category filtering (25 HPO sub-ontologies), and status filtering (untranslated, pending, approved, conflicting). Detailed term pages display English definition, synonyms, parent-child relationships, all submitted Portuguese translations with confidence scores and quality ratings, validation history, and contributor acknowledgments. Export functionality enables CSV/JSON/XLIFF/Babelon download of approved translations with Creative Commons Attribution 4.0 licensing maximizing reuse compatibility.

4 Methods: Computational Pipeline and Validation

4.1 Data Ingestion and Ontology Snapshotting

HPO terms derive from OBO format releases version 2024 or later [1], retrieved via HPO application programming interface (API) providing HP identifiers, English labels, definitions, synonyms, and is-a relationships. The ingestion script parses OBO serialization, extracts term metadata, computes complexity scores based on synonym count and cross-reference density, and inserts records into PostgreSQL via batch transactions. Complexity scoring ranges 1–5: terms with single synonym and no cross-references receive score 1, while polysemous terms exceeding five synonyms and linking ten or more diseases receive score 5. Each ingestion creates an ontology snapshot entity recording HPO version, release date, term count, and ingestion timestamp, enabling version control alignment with official releases.

4.2 Semantic Redundancy Reduction

Future preprocessing will implement semantic redundancy detection identifying near-duplicate translations via string similarity algorithms. Jaro-Winkler distance, Levenshtein edit distance, and character n-gram overlap will compute

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

pairwise similarity scores for all translations of each term. Translations exceeding configurable thresholds (e.g., Jaro-Winkler ≥ 0.95, Levenshtein distance ≤ 2 edits) will be flagged as potential duplicates for reviewer consolidation, reducing validation workload and improving consensus efficiency. Threshold sensitivity analysis will evaluate trade-offs between false positive clustering (distinct valid translations merged) and false negative redundancy (duplicates retained), informing production parameter selection.

4.3 LLM-Assisted Draft Translation

Initial Portuguese translations employ GPT-4 or equivalent LLM [7] via API with prompts conditioning on HPO context. Prompts include English term label and definition, parent term labels providing hierarchical context, synonym lists, and style guide excerpts defining lexical preferences (European Portuguese primary, Brazilian variant documented, forbidden Anglicisms). LLM outputs are logged with model version, temperature parameter, prompt template identifier, and generation timestamp for reproducibility. Even with this conditioning, drafts may show ambiguity in short clinical labels, lexical bias toward one Portuguese variant, and semantic drift from ontology intent. Therefore, outputs are treated strictly as drafts, and mandatory human validation is required before approval; no translation auto-approves from model output or confidence scores. Planned methodological evaluation will compare prompt regimes (label-only versus label+definition+hierarchical-context prompts) across multiple LLMs (e.g., GPT-family, Claude, and Gemini), using blinded expert review, back-translation error analysis, and downstream acceptance metrics.

4.4 Multi-Level Distributed Validation Workflow

Each LLM-generated or user-submitted translation enters a role-based validation pipeline enforcing quality gates. Translators create initial drafts or refine LLM outputs, submitting confidence scores (1–5 stars) indicating self-assessed accuracy. Reviewers independently evaluate submissions, assigning five-star ratings based on linguistic accuracy, semantic equivalence, and style guide compliance, with decisions of APPROVED, NEEDS_REVISION (requiring translator edits), or REJECTED (identifying fundamental errors). Convergence occurs when two or more Reviewers independently assign ratings of four or five stars and APPROVED decisions, advancing the translation to Validator approval. Disagreements indicated by conflicting decisions (one approved, one rejected) or rating discrepancies exceeding two stars escalate automatically to Validator adjudication. Validators examine discrepant cases, review submitted rationales, and issue binding decisions with justification notes visible to all parties. Committee Members resolve conflicts where two or more semantically distinct translations both achieve Reviewer approval through structured voting. Review Committee audits a random 10% sample of all approved translations plus all escalated cases, with committee sign-off required before public release. Each translation receives

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

version numbering starting at 1.0, with revisions incrementing minor versions (1.1, 1.2) and semantic changes incrementing major versions (2.0).

4.5 Quality Metrics and Governance

Translation confidence quantifies expected accuracy on a 0.0–1.0 scale combining 40% self-assessment (translator confidence score normalized) and 60% peer-review signals (average Reviewer rating normalized + approval rate). Quality score integrates multiple dimensions on a 0–5 star scale: 40% confidence contributes 0–2 points, 30% average validation rating contributes 0–1.5 points, and 30% peer approval rate contributes 0–1.5 points. Acceptance rate tracks submissions approved without major revision, and turnaround time measures median days from draft generation (or initial submission) to final approval. These are descriptive operational metrics for workflow monitoring and prioritization, not a full external validation study; current quality assessment relies mainly on internal reviewer ratings and workflow decisions within the platform. An independent linguistic validation stage is still missing, and formal inferential analysis remains future work.

Cohen's kappa is not yet reported because the current dataset is still expanding and adjudication is ongoing, with limited fully independent, same-item dual reviews in a stable analysis cohort. At this stage, inter-reviewer agreement is assessed operationally through decision concordance (e.g., both reviewers approve), rating-gap monitoring (discrepancies >2 stars trigger escalation), and escalation frequency to Validator adjudication. In later evaluation phases, planned statistical additions include confidence intervals for key performance metrics and Cohen's kappa for formal inter-rater reliability. Automated lexical consistency checks further flag style-guide violations, regional-variant mixing, and synonym inconsistencies. System integrity validation comprises 83 automated tests covering authentication, translation workflows, validation logic, RBAC, export generation, and API endpoints.

4.6 Gamification and Task Allocation

To sustain contributor engagement, the platform implements complexity-weighted scoring awarding points based on term difficulty (1–5) multiplied by consensus level (bonus for early convergence, penalty for rejections). Public leaderboards display top contributors ranked by cumulative points with monthly and quarterly recognition. Badges recognize milestones (First Translation, Century Club for 100+ terms, Elite Translator for 500+ terms). Anti-gaming measures prevent low-quality mass submissions through 24-hour rate limits between consecutive submissions on identical terms, blind quality audits re-reviewing 10% of submissions by independent validators, and reputation decay triggering automatic review when acceptance rates fall below 70% over 20 terms. Personalized task recommendations match terms to user expertise by declared medical specialty, suggesting cardiovascular terms to cardiology professionals

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

and neurodevelopmental terms to geneticists, optimizing validation accuracy through domain alignment.

4.7 Auditability and Versioning

Provenance tracking implements immutable audit logs storing all translation lifecycle events in append-only PostgreSQL tables. Each event captures actor (user ID plus ORCID if available), timestamp (millisecond precision UTC), action type (TRANSLATION_SUBMITTED, REVIEW_APPROVED, CONFLICT_ESCALATED, etc.), target entity (term ID, translation ID), and payload snapshot (full translation object pre- and post-modification). Audit queries reconstruct complete translation history including contributor sequence, review timelines, and conflict resolution decisions, supporting transparency and regulatory compliance. Version synchronization with HPO official releases employs scheduled jobs (weekly cron) fetching latest OBO files, computing diffs identifying new terms, modified definitions, or deprecated identifiers, and creating update tasks visible in translator dashboards. Translations marked obsolete when source terms deprecate trigger notification workflows alerting contributors to review replacements, preventing outdated term propagation.

5 Results: Early Validation and System Performance

5.1 Platform Deployment and Infrastructure

The production instance at `https://hpo.raras-cplp.org` has completed phases one and two: platform development with full-stack deployment and initial user recruitment engaging 15+ contributors across Rede RARAS and CPLP-RARAS institutions in Portugal and Brazil. The PostgreSQL database catalogues all 17,020 HPO terms from version 2024 releases with complete English labels, definitions, synonyms, and ontological relationships. OAuth2 integration with ORCID [9] and LinkedIn enables single sign-on, while RBAC enforces six-level permission hierarchy. System integrity currently includes 83 automated tests with 100% pass rate covering authentication, translation workflows, validation logic, role permissions, export generation, and API functionality. These usage indicators are preliminary operational measurements from an early deployment cohort: average session duration is 18 minutes, 75% of users returned within seven days, and onboarding completion exceeded 90%.

5.2 Translation Workflow Performance

Early translation activity comprises 150+ terms (approximately 0.9% of the 17,020-term HPO corpus) prioritized by clinical frequency, focusing on cardiovascular and neurodevelopmental sub-ontologies. These quantitative findings are preliminary and describe early operational performance in this initial subset. Translation confidence scores average 0.72 (standard deviation 0.14), compared

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

with a 0.70 protocol target. Quality scores average 4.1 stars (standard deviation 0.6), against a 4.0 release threshold. Acceptance rates are 82% for Translator submissions and 94% for Validator decisions. Turnaround time from LLM draft generation to final approval averages 19 days (standard deviation 5.2), corresponding to a 37% improvement relative to literature-informed manual workflow estimates (30 days; ISPOR-based reference) rather than a controlled head-to-head experiment [5]. Rework percentage is 12% (target: 15%). A controlled comparison against a traditional manual translation workflow is an important next step. Broader corpus coverage is required before making stronger inferences about scalability and robustness.

Inter-reviewer agreement, assessed operationally via voting concordance, shows 78% of translations with unanimous first-review approval, 18% requiring one revision cycle after NEEDS_REVISION feedback, and 4% escalating to Validator adjudication due to conflicting decisions. No terms have yet required Committee Member conflict resolution in this preliminary sample. Automated lexical consistency checks identify style-guide violations in 6% of submissions, predominantly Anglicism usage and regional-variant mixing, enabling correction before final approval.

5.3 Qualitative Translation Examples

Table 1 presents representative translations illustrating workflow outputs across complexity levels. Simple anatomical terms (HP:0001234, "Abnormally shaped skull") achieve rapid consensus with high confidence. Polysemous neurodevelopmental terms (HP:0001328, "Learning disability") require Validator arbitration between regionally distinct variants (European Portuguese "perturbação da aprendizagem" versus Brazilian Portuguese "transtorno de aprendizagem"), as terminological variation between "perturbação" and "transtorno" may affect semantic alignment, interoperability, and automated phenotype harmonization within multilingual HPO workflows. Rare syndrome-associated phenotypes (HP:0012345, "Brachydactyly type A1") benefit from specialist Reviewer allocation matching term complexity to contributing geneticist expertise.

In this preliminary sample, reviewer adjudication was particularly needed in three types of cases. First, short clinical labels such as HP:0001250 ("Seizure") required care to avoid overly broad renderings and to preserve clinically specific terminology. Second, terms such as HP:0002376 ("Developmental dysphasia") exposed tension between literal biomedical translation and natural clinical usage in Portuguese, requiring reviewer discussion to balance fidelity and readability. Third, regionally divergent lexical preferences, illustrated by "perturbação" versus "transtorno" in HP:0001328 ("Learning disability"), generated disagreement in semantically sensitive items and required Validator-level resolution.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

Table 1. Representative Translation Examples Across Complexity Spectrum

HP ID	English Label	Portuguese (PT-PT)	Confidence
HP:0001363	Craniosynostosis	Craniossinostose	0.89
HP:0001250	Seizure	Convulsão	0.78
HP:0002376	Developmental dysphasia	Disfasia desenvolvimental	0.65

5.4 Export Format Validation

The five export formats (CSV, JSON, XLIFF, Babelon TSV, Five Stars TSV) pass automated schema validation tests. Babelon TSV output includes ORCID identifiers for 87% of contributors (13% authenticated via LinkedIn lacking ORCID), enabling HPO consortium attribution requirements. FHIR R4 CodeSystem resources successfully validate against Health Level Seven (HL7) specifications using official FHIR validator tools, confirming EHR integration readiness. Quality metrics TSV provides transparency for translation confidence, validation ratings, and acceptance rates, supporting reproducibility and external audits.

5.5 Reproducibility and Research Artifacts

To support reproducibility, the project documents data provenance, workflow parameters, and release criteria aligned with FAIR principles [14]. The pipeline ingests HPO OBO releases from the official source (`http://purl.obolibrary.org/obo/hp.obo`) and records immutable ontology snapshots (release identifier, ingestion date, and term count). LLM-assisted draft generation is logged with model and prompt metadata to enable process tracing and future comparative analyses.

Quality-governance criteria for production release are explicitly defined (confidence, rating, and acceptance thresholds), and the role-based review workflow records validation and adjudication events. System reliability is monitored through an automated test suite covering authentication, translation workflows, access control, and export generation.

The production instance at `https://hpo.raras-cplp.org` provides public read access and authenticated contribution workflows. Exported translation artifacts are distributed under CC BY 4.0, and planned public release of source code and anonymized validation metadata is intended to support external replication studies.

6 Discussion

PORTI-HPO demonstrates the feasibility of framing biomedical ontology translation as a computational pipeline that couples automated preprocessing, LLM

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

draft generation, and distributed validation with quantitative quality assurance. In the current initial subset (150+ terms), the observed 37% turnaround reduction is relative to literature-informed manual estimates rather than a controlled head-to-head manual comparison; together with 0.72 confidence and 4.1-star quality, these findings are consistent with potential efficiency gains of the hybrid approach. Automated lexical consistency checks identifying 6% style violations complement human review, while the 83-test automated suite supports system integrity across authentication, workflows, RBAC, exports, and APIs. However, broader corpus coverage is needed before stronger claims about scalability and robustness are warranted.

The architecture is designed for reuse across ontologies and languages. Parameterized ingestion scripts accommodate multiple formats (OBO, OWL/RDF), and configurable validation workflows enable domain-specific consensus requirements such as specialist review for oncology or psychiatry. Export format abstraction via template engines allows new targets without altering core translation logic, and open documentation of pipeline stages and thresholds supports external replication, advancing FAIR principles [14]. In practice, the 4% escalation rate to Validator adjudication indicates non-trivial translation ambiguity that benefits from expert oversight, while the 82% translator acceptance rate and 94% validator approval demonstrate quality control that remains strict without discouraging participation.

Interoperability and impact are reinforced through standards-based delivery. Validated HPO-PT enables Portuguese-language rare disease registries to adopt standardized phenotype coding and improves data quality for epidemiological research across nine CPLP countries (280 million speakers). FHIR R4 endpoints [8] facilitate integration with lusophone EHRs, reducing clinician documentation burden by providing native-language phenotype autocomplete, while Babelon TSV export with ORCID attribution [9] supports direct submission to the HPO consortium. Planned Phenopackets [15] integration will further enable standardized genomic data exchange, strengthening international collaboration and advancing precision medicine equity in underrepresented language communities.

7 Limitations

This study has five primary limitations. First, current evaluation covers only 150+ terms (approximately 0.9% of the 17,020-term HPO snapshot), so findings should be interpreted as preliminary. Second, contributors are currently concentrated in Brazil and Portugal, which may underrepresent lexical and clinical nuances from African Portuguese-speaking countries and other CPLP settings. Third, an external validation stage has not yet been completed; current validation is internal to the platform and relies mainly on reviewer ratings and workflow decisions, while an independent linguistic validation stage is still missing. Planned independent back-translation will be used as an external benchmark of semantic fidelity. Fourth, the reported turnaround improvement is relative

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

to literature-informed manual workflow estimates and not a controlled head-to-head experiment against a concurrent manual-translation baseline; a controlled comparison with a traditional manual translation workflow is an important next step. Fifth, formal inter-rater reliability statistics such as Cohen's kappa are planned but not yet reported; current agreement assessment is operational (decision concordance, rating-gap escalation, and adjudication frequency).

8 Conclusion and Future Work

This paper presented PORTI-HPO, a production-deployed computational platform combining LLM-assisted draft generation with distributed multi-expert validation for biomedical ontology localization. The system addresses ontology translation as a computational pipeline with explicit support for semantic consistency (hierarchical relationship preservation), distributed workflow coordination (six-role RBAC), provenance tracking (immutable audit logs), and interoperability (FHIR R4 endpoints, five export formats). Early validation across 150+ Portuguese translations (approximately 0.9% of the current HPO corpus) reports 0.72 confidence (exceeding 0.70 target), 4.1-star quality (meeting 4.0 threshold), 82–94% acceptance rates (surpassing 80–90% goals), and 19-day turnaround (37% faster than literature-based manual estimates). These findings support preliminary feasibility of the hybrid LLM-expert approach, while broader corpus coverage is required before stronger claims on scalability and robustness.

Future work will pursue three directions. First, scaling to 2,000+ validated terms through sustained contributor recruitment across all nine CPLP countries, with targeted outreach to underrepresented African lusophone communities to improve lexical and clinical representativeness. Second, expanding interoperability via OWL/RDF exports enabling Semantic Web reasoning, FHIR R4 ValueSet resources for clinical context-specific term subsets, and Phenopackets [15] integration supporting international genomic data exchange. Third, implementing a planned methodological evaluation of LLM drafting strategies, including controlled comparisons of label-only prompts versus label+definition+hierarchical-context prompts across different LLMs, together with Cohen's kappa computation after complete review cycles, independent back-translation by external experts for semantic-fidelity benchmarking, and longitudinal analysis linking gamification engagement to quality trends. A complementary next step is independent linguistic evaluation by external domain experts not directly involved in the routine platform workflow.

The reusable framework, open documentation, and production deployment at `https://hpo.raras-cplp.org` offer a reproducible methodology for multilingual biomedical resource development. By addressing ontology localization through computational pipeline design supported by quantitative quality metrics and comprehensive automated testing, this work offers an initial, reproducible pathway toward broader access to standardized phenotype terminologies for rare disease diagnostics across underserved linguistic populations.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

Acknowledgments. This work was supported by Rede RARAS (Brazilian Rare Diseases Network) and CPLP-RARAS (Community of Portuguese-Speaking Countries Rare Diseases Network). This study was supported by the Brazilian National Council for Scientific and Technological Development (CNPq), under grant number 403244/2024-2, as part of the project Mapeamento de Doenças Raras na Comunidade dos Países de Língua Portuguesa: Avanços em Saúde Digital e Cooperação Internacional (2025–2027), coordinated by DA. We acknowledge HPO consortium for ontology maintenance and contributors engaging in translation validation workflows. The authors acknowledge the use of generative AI tools (e.g., ChatGPT) for language refinement purposes. All content, ideas, and conclusions presented are solely those of the authors.

Disclosure of Interests. The authors declare no competing interests relevant to this work.

References

1. Köhler, S., Gargano, M., Matentzoglu, N., et al.: The Human Phenotype Ontology in 2021. Nucleic Acids Res. **49**(D1), D1207–D1217 (2021). https://doi.org/10.1093/nar/gkaa1043
2. Robinson, P.N., Köhler, S., Bauer, S., et al.: The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. **83**(5), 610–615 (2008). https://doi.org/10.1016/j.ajhg.2008.09.017
3. Amberger, J.S., Bocchini, C.A., Scott, A.F., Hamosh, A.: OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. **47**(D1), D1038–D1043 (2019). https://doi.org/10.1093/nar/gky1151
4. Rath, A., Olry, A., Dhombres, F., et al.: Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum. Mutat. **33**(5), 803–808 (2012). https://doi.org/10.1002/humu.22078
5. Wild, D., Grove, A., Martin, M., et al.: Principles of good practice for the translation and cultural adaptation of patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health **8**(2), 94–104 (2005). https://doi.org/10.1111/j.1524-4733.2005.04054.x
6. Groza, T., Köhler, S., Moldenhauer, D., et al.: The Human Phenotype Ontology: semantic unification of common and rare disease. Am. J. Hum. Genet. **97**(1), 111–124 (2015). https://doi.org/10.1016/j.ajhg.2015.05.020
7. Luo, R., Watanabe, M., Alterovitz, G.: GPT-4 for phenotype concept recognition: towards large language models in precision medicine. JAMIA Open **7**(1), ooae020 (2024). https://doi.org/10.1093/jamiaopen/ooae020
8. Mandl, K.D., Mandel, J.C., Murphy, S.N., et al.: The SMART Platform: early experience enabling substitutable applications for electronic health records. J. Am. Med. Inform. Assoc. **19**(4), 597–603 (2012). https://doi.org/10.1136/amiajnl-2011-000622
9. Hauser, R.G., Sheppard, N.: Adoption and utilization of ORCID identifiers as unique researcher identifiers. Learn. Publ. **32**(2), 104–113 (2019). https://doi.org/10.1002/leap.1224
10. Good, B.M., Su, A.I.: Crowdsourcing for bioinformatics. Bioinformatics **29**(16), 1925–1933 (2013). https://doi.org/10.1093/bioinformatics/btt333

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21

11. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 453–456. ACM (2008). https://doi.org/10.1145/1357054.1357127
12. Hamari, J., Koivisto, J., Sarsa, H.: Does gamification work? A literature review of empirical studies on gamification. In: 2014 47th Hawaii International Conference on System Sciences, pp. 3025–3034. IEEE (2014). https://doi.org/10.1109/HICSS.2014.377
13. Mendoza, K., Barbosa, J., Rivera-Romero, O., et al.: Development of a crowdsourcing- and gamification-based mobile application for epidemiological surveillance. Sci. Rep. **14**, 6174 (2024). https://doi.org/10.1038/s41598-024-56761-4
14. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
15. Jacobsen, J.O.B., Baudis, M., Baynam, G.S., et al.: The GA4GH Phenopacket schema defines a computable representation of clinical data. Nat. Biotechnol. **40**(6), 817–820 (2022). https://doi.org/10.1038/s41587-022-01357-4
16. Dove, E.S., Townend, D., Knoppers, B.M.: Harmonization of laws, regulations, and guidelines. In: Mascalzoni, D. (ed.) Ethics and Governance of Biomedical Research, pp. 99–120. Springer (2018). https://doi.org/10.1007/978-3-319-77673-9_6
17. Zhou, L., Huang, J.X., Grivel, G., et al.: Linguistic characteristics and retrieval effectiveness of biomedical text. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1011–1020. ACM (2011). https://doi.org/10.1145/2009916.2010048
18. Wolk, K., Wolk, A.: Machine-enhanced translation of Human Phenotype Ontology from English to Polish. J. Med. Inform. Technol. **26**, 1–8 (2017).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_21