Patient data collection and phenotype platform — technical specification

Architecture, data model, API design and implementation status. Version 1.0.

1. Purpose and scope

The platform collects structured patient data for clinical genomics and rare-disease work and turns clinical observations into standard, machine-readable phenotype data (Human Phenotype Ontology, HPO). It supports registration, history, family history and pedigree, organ-system phenotyping, examinations, laboratory results, genomic results, phenotype extraction from text, phenotype-driven gene and panel ranking, dashboards and exports.

Central design rule: Clinical observation → standard phenotype annotation → disease / gene association → genomic evidence → clinician review. The observation is the clinical record; HPO codes are versioned annotations attached to it.

2. What is built in this release

Release 1.0 is a single-file browser application. It runs inside the Tribro-Omics CRM (version 9) and as a stand-alone file (TribroPheno) for TribroLIMS sites. Data is encrypted in the browser (AES-GCM, key derived with PBKDF2-SHA256). No patient data is sent to a server.

Requirement	Status	Notes
Patient registration and demographics	Built	Automatic patient ID and separate research key; adult, paediatric, neonatal, prenatal, deceased; age calculated.
Clinical history (complaints, medical, birth and development)	Built	Repeatable complaints and medications; conditional fields.
Family history and pedigree	Built	Three-generation drawing, consanguinity, PED and SVG export. Uncertain relatives are listed, not placed.
Organ-system phenotyping (22 systems)	Built	Six explicit statuses plus “not recorded”; evidence source, onset, severity, course, frequency, laterality.
HPO search, browse and annotation	Built	Release 2025-01-16 embedded (18,387 terms). ID, label, synonym, definition and plain-language search; parents and children; candidate, accepted, rejected and superseded mappings.
Ontology update control	Built	Annotations keep their release. A check lists replaced or obsolete IDs; replacement is applied only after review and both annotations are kept.
Clinical examination and laboratory forms	Built	10 examination types with measurements; 16 laboratory categories with units, ranges and automatic flags.

Genomics module	Built (records and references)	Tests, variants with HGVS / ACMG fields, file references with checksums. Sequence files are not parsed in the browser.
Text extraction of phenotypes	Built (rule-based)	Dictionary and context rules for negation, uncertainty, family context and onset. Not an LLM. Every suggestion needs human review.
Gene and panel prioritisation	Built (built-in method)	Information-content similarity against HPO gene annotations (5,125 genes) and the 200-panel catalogue. Exomiser / LIRICAL are reached through Phenopacket export, not run inside the tool.
Dashboards, filters, saved queries	Built	Counts by sex, age group, case type, clinical area, testing and variant status, top HPO terms; small-count masking for the researcher role.
Exports	Built	Patient JSON, HPO TSV / CSV, Phenopacket v2, FHIR R4-mapped bundle, PED, cohort TSV / CSV, data dictionary, TribroLIMS subject JSON, printable report.
Audit trail, record states, amendments	Built	Draft → submitted → verified → amended with reason; every change and export logged.
Consent and withdrawal	Built	Six consent types; withdrawn or declined research consent blocks research exports.
Multi-user server (FastAPI + PostgreSQL), per-user accounts	Specified, not built	This release is a single-file browser application with one encrypted sign-in. Sections 5–7 define the server edition.
MONDO / ORDO / SNOMED CT / LOINC code lookups	Specified, not built	Free-text and code fields exist; ontology lookups other than HPO are not embedded. See the resource register.
Image and document file storage	Specified, not built	Only references, checksums and consent flags are stored; files stay in institutional storage.
Profile-validated FHIR, clinical validation of decision support	Not done	Requires a FHIR validator run and a formal validation study.

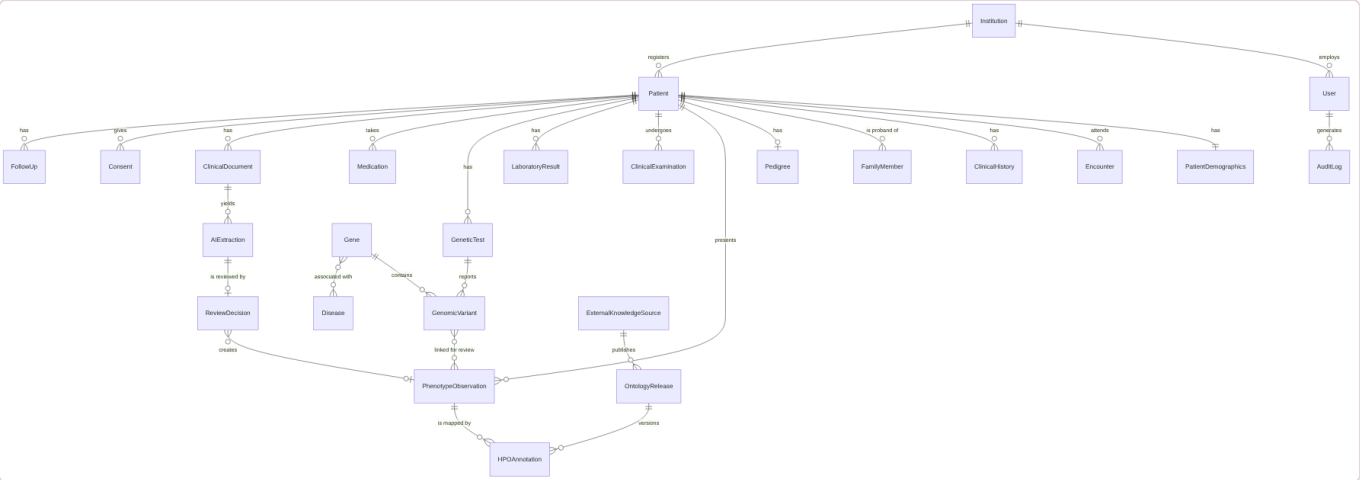
3. Architecture

- Presentation: one HTML file; no build step; works offline after loading.
- Phenotype engine (no network): HPO search, hierarchy, text extraction, information-content gene ranking, panel matching.
- Ontology data: HPO release 2025-01-16 and HPO gene annotations, compressed inside the file. A new release is a new build; old annotations keep their release label.
- Storage: encrypted vault in browser storage; JSON backup export and import.
- Interchange: Phenopacket v2 (Exomiser, LIRICAL), FHIR R4-mapped bundle, TSV / CSV, PED, TribroLIMS subject JSON.
- Server edition (specified): React + TypeScript front end, FastAPI back end, PostgreSQL, object storage for genomic files, background workers for imports and analysis, Docker deployment, CI with automated tests.

4. Phenotype status model

Status	Meaning
Present	Finding was assessed and is present.
Absent	Finding was explicitly assessed and is not present. Exported as an excluded phenotype.
Suspected	Finding is possible but not confirmed. Kept apart from confirmed findings; left out of ranking unless switched on.
Unknown	Asked or looked for, result not known.
Not assessed	Not examined or asked.
Not applicable	Cannot apply to this patient (for example by age or sex).
(no record)	Nothing was recorded. Never treated as absent.

5. Data model



Entity-relationship diagram (server edition). The browser edition stores the same entities as one encrypted document per patient.

Entity	Main attributes	Notes
Patient	id, patient_id, research_key, state, created_at, updated_at	Root record. research_key is the only key used in research exports.
PatientDemographics	patient_id →Patient, names, dob, sex, gender, ancestry, contact, clinicians	Kept apart so it can be access-restricted.
Encounter	id, patient_id, date, type, clinician	Groups examinations, results and follow-up of one visit.
ClinicalHistory	id, patient_id, section, payload	Complaints, medical, birth and developmental history.
FamilyMember / Pedigree	id, patient_id, relationship, sex, vital, affected, phenotype, onset, tested	Relationship is stated, never inferred.

PhenotypeObservation	id, patient_id, text, system, status, source, onset, resolved, severity, progression, frequency, laterality, location, notes, origin, created_by	The clinical record of a finding. Exists without any ontology code.
HPOAnnotation	id, observation_id, hpo_id, label, status, confidence, annotator, date, release, origin, replaces	Zero or more per observation. Status: candidate, accepted, rejected, superseded.
OntologyRelease	id, ontology, version, loaded_at, checksum	Every annotation points to the release it was made with.
ClinicalExamination	id, patient_id, type, date, examiner, findings, interpretation, measurements[]	
LaboratoryResult	id, patient_id, category, test, code, value, unit, ref_low, ref_high, flag, sample, collected_at, lab, method	
GeneticTest	id, patient_id, type, lab, date, status, assembly, panel, result	
GenomicVariant	id, test_id, gene, gene_id, assembly, chrom, pos, ref, alt, hgvs_c, hgvs_p, zygosity, allele_balance, consequence, inheritance, af, af_source, clinvar, acmg, criteria, evidence_for, evidence_against, analyst, date	Linked to observations for review; a link is not a causal claim.
ClinicalDocument	id, patient_id, type, title, date, author, text	
AIExtraction	id, document_id, span_start, span_end, text, hpo_id, context, onset, confidence, model, release, extracted_at	Suggestion only.
ReviewDecision	extraction_id, decision, reviewer, reviewed_at, observation_id	Accepted, modified or rejected.
Consent	id, patient_id, type, status, date, version, given_by, witness, withdrawn_at	
FollowUp	id, patient_id, date, clinician, status, summary, plan, next_date	
User / Institution	id, name, role, institution_id	Roles: clinician, researcher, data manager, administrator.
AuditLog	id, at, user, role, patient_id, action, detail	Append-only.
ExternalKnowledgeSource	id, name, version, licence, retrieved_at	Provenance of every external annotation.

6. API design (server edition)

Versioned REST API under /api/v1, JSON bodies, bearer tokens, role checks on every route, all writes audited. The browser edition implements the same operations in-process.

Method	Route	Purpose
POST	/api/v1/auth/login	Sign in; returns a short-lived token

GET / POST	/api/v1/patients	List (filters, paging) / create
GET / PATCH	/api/v1/patients/{id}	Read / update; PATCH on a verified record requires an amendment reason
POST	/api/v1/patients/{id}/state	Submit, verify, amend
GET / POST	/api/v1/patients/{id}/observations	Phenotype observations
POST	/api/v1/observations/{id}/annotations	Add candidate or accepted HPO annotation
PATCH	/api/v1/annotations/{id}	Accept, reject, supersede
GET	/api/v1/hpo/search?q=&root=	Term search (ID, label, synonym, definition)
GET	/api/v1/hpo/terms/{hpo_id}	Term detail with parents and children
POST	/api/v1/documents/{id}/extract	Run extraction; returns suggestions with spans
POST	/api/v1/extractions/{id}/review	Accept, modify or reject
POST	/api/v1/patients/{id}/prioritise	Rank genes and panels; stores run with parameters
GET	/api/v1/patients/{id}/export?format=	json, tsv, csv, phenopacket, fhir, ped
GET	/api/v1/cohort/export?format=&deidentified=	Cohort exports with consent filter
GET	/api/v1/dashboard	Aggregates with minimum cell size
GET	/api/v1/audit?patient=	Audit trail (administrator, data manager)
POST	/api/v1/ontology/releases	Load a new release; produces a migration report, changes nothing
POST	/api/v1/lims/subjects	TribroLIMS: accept a tribro.lims.subject/1 document

7. Text extraction method

- Sentence and token split; phrases of up to seven words are matched against HPO labels, synonyms and a short plain-language list.
- Context rules: negation before or after the phrase (no, without, denies, ruled out, not present…), uncertainty (possible, suspected, rule out, “?”), family context (mother, sibling, family history…), onset (“since 8 months”, “at birth”).
- Each suggestion stores the source span, the extractor version, the HPO release, the date and a confidence value, and stays Pending until reviewed.
- Built-in regression check on 15 synthetic sentences (23 findings): precision 96 %, recall 96 %, context 100 %. This is a software check on easy sentences, not a clinical validation; real notes will score lower.

8. Gene and panel ranking method

- Information content of each HPO term is computed from HPO gene annotations.
- For each patient term, the best-matching annotation of a gene is found through the most informative common ancestor. Score = 85 % patient-side coverage + 15 % gene-side specificity, on a 0–100 scale.
- Genes with fewer than five HPO annotations are left out. An explicitly excluded term annotated to a gene lowers its score by 10 % (at most 30 %).

- Panels are ranked by how many of the top 25 genes they contain.
- When more than ten genes lie within five points of the top score the tool says the phenotype is non-specific. The ranking is a candidate list for review, never a diagnosis.

9. Open decisions

- Clinical: which specialty templates to add first; who verifies records; policy for secondary findings.
- Regulatory: whether any function is used for diagnosis (would need assessment under medical-device rules); data-protection impact assessment; ethics approval for research use.
- Licensing: OMIM, SNOMED CT, DECIPHER and PanelApp terms for commercial use (see the resource register).
- Technical: hosting for the server edition; identity provider; FHIR profiles to validate against; LLM-based extraction and its evaluation set.

© Tribro-Omics · Patient data collection and phenotype platform — technical specification · Version 1.0 · 11 November 2026