WHITE PAPER

Stone Age to Space Age

The hidden discipline of mining-grade data normalisation.

Less searching. More strategising.™

PREPARED FOR	Geological practitioners — chief geologists, technical directors, exploration leads, JV technical reviewers, resource modellers, and the data engineers who support them
DOCUMENT	Paper two of two — companion to “The Trust Layer: From Co-Pilot to Autopilot in Mining Finance”
DATE ISSUED	July 2026

WHY THIS PAPER EXISTS Geology is a discipline of careful inference from incomplete data. Geological data normalisation is the same discipline applied one layer below — to the data the geology itself is built on. Both deserve the same rigour. Most platforms, most consultants, and most internal builds underestimate what the second discipline costs. This paper is the long answer.

About This Paper

This paper is written for the people who carry technical responsibility for what their organisations decide on the rocks — chief geologists, technical directors, exploration leads, JV technical reviewers, resource modellers, the data engineers who support them, and the geologists who run all of the above by themselves at smaller firms. Whatever the seat, the discipline is the same: drawing careful inference from incomplete, inconsistent, multilingual, multi-era data, and being able to defend that inference under JV review or an IC challenge.

It walks through what mining-grade data normalisation actually entails when the corpus is real — English or Cyrillic, born-digital or a degraded scan, filed this morning or decades old, multi-jurisdictional, sometimes structured for a regulatory regime that no longer exists — why the economics of doing this with generic AI get worse rather than better over the next three years, and what it takes for that work to become genuinely usable to a working geologist. The data problem and the usability problem are not the same problem. Most platforms underestimate the second one.

It uses concrete examples drawn from live engagements, anonymised under client agreements. It is a field report. A companion paper — "The Trust Layer: From Co-Pilot to Autopilot in Mining Finance" — covers the same architecture for mining capital allocators.

Contents

01 · The Grade That Wasn't

Picture a single drill intercept. A grade, an interval, a depth, a coordinate — the four numbers that, between them, tell a working geologist almost everything that matters about a single line in a corpus.

Start with the easy version. A born-digital English 43-101, printed perfectly, nothing degraded. Even here the four numbers are not the four numbers the model needs: the grade is quoted in ounces per ton where the model wants grams per tonne; the all-in sustaining cost is struck on a by-product-credit basis the peer set does not share; the reserve was restated under S-K 1300 and is genuinely a different figure from the one carried three years ago; and the whole thing still arrives as a picture of a page, the data trapped inside the image. No Cyrillic, no handwriting, no obsolete datum — and the normalisation discipline is already load-bearing. Now take the hard version.

Now picture the same intercept the way it actually arrived at our pipeline last month. The grade is reported in volumetric units — not g/t, not %, not ppm — because the report is from the late 1970s and the convention for that placer assemblage, in that country, at that time, was volumetric. The interval is not a sampled interval at all; it is a horizon-average expressed as a range — "grades from X to Y" — an arbitrary scale, not a true drill intercept. The depth is referenced to a collar elevation that is itself referenced to a regional datum current then but not now. And the coordinate is referenced to the Pulkovo Meridian — the geographical reference of the Russian Empire from 1844, sitting 30°19'34" east of Greenwich, which underpins the cartography of much of the former Soviet exploration corpus to this day.

Take that single intercept into a modern resource model and every one of those four numbers needs to be transformed before the model will accept it. Not converted — transformed. The unit conversion needs a custom constant for the volumetric placer assemblage. The interval requires a deliberate decision about whether to treat the horizon-average as a proxy for a sampled interval or to flag it as something else. The depth requires datum reconstruction. The coordinate requires a Pulkovo → WGS84 transformation, plus explicit handling of the historical Soviet "decree time" shift if any time-stamped record is involved. None of these transformations is impossible. Each is a place where a careless team will silently introduce a thirty-degree-of-longitude error.

The story is not unusual. It is the median morning at any data team trying to ingest a real corpus of historical and multilingual mining disclosures. The grade that was reported is not the grade the resource model needs. The discipline of getting from one to the other is not exotic and it is not optional — and it is the thing every off-the-shelf platform underestimates, because every off-the-shelf platform was built around the cleanly-disclosed quarterly figures of major operators in the major reporting jurisdictions. The rest of the industry has been left to fend for itself.

This paper is about the rest of the industry. It is also about the second discipline that lives one layer above the data work: making the result genuinely usable to a working geologist who is, almost by definition, not an IT specialist.

> **THE LANGUAGE IS THE LEAST INTERESTING PART** The goal was never to read Cyrillic. It was to make the language — and the century, and the file format — irrelevant to the answer. The Soviet-Cyrillic corpus is the hardest case we have worked, and we lean on it as proof: a pipeline that pushes past 99% on handwritten, stamped, mid-twentieth-century Cyrillic has earned the right to be believed on a clean English 43-101, an Australian annual report, a Spanish feasibility study, or the disclosure that dropped this morning. Cyrillic is the stress test, not the subject. The same architecture runs, at comparable accuracy, on every major resource language and on born-digital

English filings — where, as one of our own team put it, 'the printed-perfectly Western English stuff' extracts to effectively 100% at any scale at the first pass.

A 1982 Cyrillic drill-hole passport in the Pulse pipeline — the raw scan (left) and the same page with title, lithology, geophysics and assay regions detected and typed (right). The language is a variable; the layout discipline is the constant.

02 · The Five (main) Normalisation Problems

Read the five in order and a pattern comes into focus: each is theoretically solvable, none is solvable casually, and the cumulative cost of solving them all by hand is what makes internal builds stall at the eighteen-month mark.

LANGUAGE IS A VARIABLE, NOT THE SUBJECT. Four of the five problems below have nothing to do with the language the data is presented in. Units of measure, currency and economic convention, reporting-standard reconciliation, and a hundred and fifty years of shifting measurement practice are as present in a clean English 43-101, an Australian annual report carrying three currencies, or a Canadian S-K 1300 restatement as they are in a Cyrillic Soviet passport. Only the fourth — georeferencing and place-name collisions — is mostly former-Soviet in flavour, and even that is the extreme end of a universal problem. The language is where the reader's eye goes; it is the least interesting part of the work.

2.1 Units of measure

The most basic problem, and the one most often dismissed. Grade is reported in g/t, %, ppm, oz/t depending on commodity, jurisdiction, era — and, we have learned to expect, the personal preference of the technical author. A unit error at the front end becomes a deal error at the back end, and the deal error is not always caught.

During a recent platform review with the technical team at a tier-one royalty client, the conversation moved into the front end of the comps interface. The team's technical lead stayed on units — ounces-per-ton versus grams-per-ton — long after the rest of the conversation had moved on. They were right to. At decision-grade scale, a unit error is structural; an interface that surfaces the original disclosed value and the normalised value side by side is not a polish item. It is a defence.

Some of the harder unit problems we have built explicit handlers for: volumetric Soviet placer gold (Россыпи золота / Rossypi Zolota), reported in volumetric units that require a custom conversion constant rather than standard mass-grade conversion; Soviet sample types — Potswinae Proba, lithogeochemical analysis, channel sample, bulk sample — that do not map cleanly to JORC-equivalent categories and require explicit, governed, version-controlled equivalence rules; and historical "%" entries that are sometimes molar, sometimes mass, sometimes volume, depending on the assay laboratory and the year. Pulse holds the original disclosed value alongside the normalised value at every layer. The original is never overwritten; the normalised is never presented without the original one click away.

2.2 Currency and economic conventions

Currency normalisation looks trivial until it shows up. An Australian operator's annual report frequently carries multiple currency references inside a single document — head numbers in AUD, segment breakouts in USD, a divestment line in CAD or ZAR. Disambiguation requires page-level context, not a document-level assumption. Soviet-era cost reporting in roubles requires inflation-adjusted reconstruction before any modern peer comparison is meaningful — the rouble was redenominated multiple times across the relevant period, and Soviet inflation indices are not directly comparable to modern equivalents. We carry an explicit historical cost handler; teams without one tend to silently drop the cost lines and pretend they were never relevant. Royalty and streaming economics add their own definitional challenges: NSR, GRR, NPI, and percentage-of-gross variants are not interchangeable. A peer comparison that mixes them is not a peer comparison; it is a misleading bar chart.

2.3 Reporting standard reconciliation

JORC, NI 43-101, SAMREC, S-K 1300, and historical Soviet GKZ are not interchangeable. Each defines categories — measured, indicated, inferred, and their pre-modern equivalents — with subtly different criteria and confidence thresholds. A reserve restated under a new regime is genuinely different from the previous figure, even when the underlying geology has not changed. The Soviet GKZ classification deserves particular attention: the categories A, B, C1, C2, P1, P2, P3 are sometimes mapped one-to-one against JORC categories. They should not be — the GKZ confidence model was built around drilling-density and geological-assurance criteria that do not align with modern definitions. Equivalence is not equivalence; the reconciliation requires explicit handling, ideally with a per-deposit-type lookup. Pulse maintains an explicit reconciliation layer for these mappings, governed by domain experts and updated as standards evolve — a living artefact, not a one-time spec.

2.4 Georeferencing and place-name collisions

The Pulkovo Meridian. The Pulkovo Observatory, founded outside Saint Petersburg in 1839, sat at the centre of the Russian Empire's scientific cartography. Its meridian — 30°19'34" east of Greenwich — was adopted in 1844 as the geographical reference for all Russian mapping and remained in use through the Soviet period for many internal purposes. Without an explicit Pulkovo → WGS84 transformation, an asset can sit roughly thirty degrees of longitude from where the modern map says it is. This is not a subtle error. It places assets in the wrong country, produces geographically incoherent drainage analyses, and defeats every downstream geospatial workflow until corrected. The Pulse pipeline carries explicit handling for Pulkovo-referenced coordinates and the related Krasovsky 1940 datum, with confidence levels surfaced for any inferred transformation.

Beyond the historical datums. The datum problem is the extreme end of a universal one: most drill data arrives as eastings and northings with the projection or zone unstated, and a single project —

sometimes a single report — can carry more than one. Two holes in the wrong projection plot hundreds of metres apart. Correctly identifying the system each coordinate sits in is the load-bearing step; the arithmetic after it is trivial. Pulse detects the system, transforms to a common WGS84 reference, and flags any point that plots inconsistently with the rest of its dataset rather than mapping it as fact.

Soviet decree time. By Sovnarkom decree of 16 June 1930, all clocks in the Soviet Union were permanently shifted one hour ahead from 21 June 1930. The decree time persisted in various forms and influenced time-stamped records in geological and exploration archives. Cross-comparing time-stamped Soviet records against international datasets without correcting for it produces a quiet temporal drift that compounds across the period.

Place-name collisions. Dozens of villages, regions and rayons share names across the former Soviet republics. “Spasskoye” is everywhere. Disambiguation requires regional context, oblast metadata, neighbouring named features, and sometimes historical administrative-boundary reconstruction. A team that resolves place names by string match alone will conflate three deposits into one often enough that the conflation becomes a structural error in the resource model.

Latin-character map IDs in Cyrillic-only OCR contexts. A recurring extraction failure mode. Soviet-era OCR was Russian-only; map IDs and reference labels frequently use Latin characters that the OCR drops or misrecognises. The fix is a bilingual OCR pipeline with explicit Latin-character handling, not a post-hoc string-matching layer. We learned this the hard way; the fix is in production.

2.5 Measurement conventions across 150 years

Mathematical conventions and statistical treatments evolve. Outlier handling, capping, declustering, top-cut conventions, and the philosophical position of how much of a high grade to believe are not historically constant. A historical resource estimate is not just a number; it is a number produced by a method that itself has a date stamp. Three places where this most often surfaces: recovery and metallurgical assumptions baked into historical estimates frequently require explicit re-derivation rather than adoption — assuming a 1970s recovery onto a modern flowsheet is conservative at best and actively wrong at worst; drill-interval definitions, where Soviet reports often quote grade ranges across a horizon rather than a sampled interval — differentiating true drill intercepts from horizon averages is the single hardest extraction call in the historical corpus, and the one most likely to be silently mishandled by a vendor without a geologist in the loop; and capping and outlier handling, which vary enough that the same drill intercept produces different resource estimates depending on era and analyst. Without that lineage, a modern team is choosing which past methodological assumption to adopt without realising it.

03 · Going Past The Bar, Not Just Over It

Clients, of course, seek accuracy. Expectations, perceived as high, seem low to us. On typical extraction completeness and accuracy metrics (when contacting Pulse for a pilot/POC with Pulse), clients’ requests seem to average around 90%.

For us, such a threshold is unusable at scale. Ninety-per cent extraction accuracy on a technical corpus is simply not usable in day-to-day workflows.

The reason is structural. Errors at the 90% level cannot be filtered — there is no signal that distinguishes the correct rows from the incorrect ones without re-verifying every cell. A team running on 90% extraction has effectively done none of the validation work; they have done a fast pull and

inherited the burden of cell-by-cell verification on every row that matters. The work is still in front of them. Mining-grade extraction has to push past 99% on the metrics that drive decisions, and it has to do so on real corpora — not on the cleanly printed examples vendors choose for their benchmarks. And it has to know which 1% it could not process, and flag for human review.

A worked example, anonymised

A prospective client commissioned a pilot on complex Soviet-era geological reports — including maps — drawn from their historic exploration archive. The corpus was Cyrillic, multi-format, and historical — exactly the kind of corpus an off-the-shelf toolchain quietly fails on.

The brief: meet or exceed every threshold in the producer's scoring framework, on real data with real ground truth.

Seventy-one Priority-1 documents. One hundred and sixty-four input files in total — a blend of drill-hole passports, well geological-geophysical columns, well cross-sections, stratigraphic correlation panels and physical-properties sheets, each with its own structural conventions. The hardest documents — single drill-log passports — carried up to 20,000 data points apiece, on the order of 2,000 rows by 10 columns, at variable scan quality. The pilot was deliberately structured around solving the hardest items first.

The numbers, on the scored ground-truth corpus

METRIC	THRESHOLD	RESULT
Drill Hole ID accuracy	≥ 90%	Met — full match on Cyrillic and numeric IDs
Depth interval accuracy	≥ 90%	99.65% row recall; 99.71% interval-length equality
Assay value accuracy	≥ 90%	99.21% combined element-cell accuracy across 12 GT documents
Lithology pattern recognition	≥ 80%	Extracted on 37/37 assay-contributing docs; 96.3% of assay rows enriched
Other schema fields	≥ 80%	Met — year, collar elevation, datum, report ID, image ID
Precision	≥ 90%	99.06% aggregate
Recall	≥ 90%	99.65%
OCR — word accuracy	≥ 85%	~97% word / ~99% character on Russian text incl. handwritten, stamped, degraded
Table detection rate	≥ 90%	509 tables extracted across the PDF corpus
Processing failure rate	< 5%	0 of 164 hard failures

On the same corpus the pipeline extracted 11,249 drill-hole assay rows across 37 drill-log images, covering 34 distinct geochemical elements (majors Cu, Mo, Pb, Zn, Ag), and enriched 96.3% of those rows with full lithology — rock age, mineralisation, and lithological class alongside the grade data. The numbers were produced by combining computer vision, classical OCR (Tesseract), LLM-based extraction, and statistical reconciliation. No single technique would have reached 99%+ alone. The combination — validated through a benchmarking framework with fifty to two hundred hard samples per pipeline step, regression-tested on every change — is what made the result reproducible rather than lucky. That benchmarking discipline, not the model selection, is the part most internal builds skip.

> *"At this point, I'm not demanding 99% accuracy. I want to know if it's 70%, 80%, 90% — then we work in that context."*
>
> — Operating geologist, frontier exploration project (anonymised)

That quote captures what published vendor benchmarks miss. Confidence in the number matters more than the number itself. A vendor that says "99%" without describing the corpus, the metric, and the validation methodology is selling marketing. A vendor that says "Cu accuracy is 99.21% on this scored ground-truth subset, lithology accuracy is in the low-to-mid 90s by internal benchmark and pending external ground truth, multi-hole stratigraphic panels are deliberately scoped out and require a dedicated handler" is being honest.

04 · The Real Cost Nobody Is Pricing In

Everything in Section 03 — the 99%-plus accuracy, the worked example, the discipline behind it — costs something to produce. The question this section asks is what that cost does over the next three years, and why "once" is the most important word in the answer.

The reassuring story, and why it is the wrong conclusion

The price of a unit of AI — a token, the basic billing unit of a language model — has fallen faster than almost any input cost in modern computing history. Independent analysis from Epoch AI puts the decline at roughly an order of magnitude per year; Andreessen Horowitz named the trend "LLMflation." Read on its own, the conclusion is comfortable: AI gets cheaper every year, so the cost of processing a corpus takes care of itself.

It is the wrong conclusion. The per-token price is falling. The bill is not. Across 2025, enterprise AI spend rose sharply even as unit prices collapsed — the pattern one widely cited analysis called "cheaper tokens, bigger bills." Three forces drive the volume: reasoning models, which generate large quantities of internal "thinking" tokens before producing an answer; agentic workflows, which Gartner's 2026 analysis puts at five to thirty times the tokens of a single query; and context inflation — the cost of moving large documents through a model — which routinely adds 40–60% on top of raw inference. Beneath all of it, the current price of frontier AI is, by the providers' own disclosures, subsidised below cost. The floor rises eventually. Not "if." "When."

Why a technical corpus is the worst-case exposure

A Soviet-era geological corpus is, almost by definition, the most token-expensive material there is. A single drill-log passport can carry 20,000 data points. A frontier exploration archive easily runs to three million pages or more. Reading that material accurately — handwritten, stamped, degraded, multi-format, multi-era — demands the most capable model, the most reasoning tokens, the most context. It is the premium end of the cost curve, not the cheap end.

The naive way to "use AI on the archive" is to point a generic OCR-plus-LLM toolchain at it and re-run that processing whenever someone needs to query it. That architecture pays the most expensive bill there is, and pays it again on every query, by every geologist — re-deriving an extraction the team already paid for last week, because nothing was kept in a structured, validated form. Its cost scales with users, with queries, and with model pricing, and it is least reliable in exactly the places it is most expensive.

Why processing once, properly, is the cost answer

Pulse processes, normalises and categorises the corpus once. In the engagement described in Section 08, a corpus of roughly three million pages was processed end-to-end in about forty-eight hours after pipeline optimisation, and roughly 950,000 grade records were extracted once and stored as structured, validated, queryable data, ready for engagement on the latest AI models.

Every query thereafter is a deterministic lookup against that store — not a token-metered re-reading of three million pages. Because Pulse is algorithmic-first and LLM-downstream, the truth layer does not spend reasoning tokens re-deriving grades on demand; language models run downstream, over data that is already structured.

There is a second cost feature worth naming. The pipeline carries a self-monitoring quality layer that predicts the expected extraction count per page and flags pages where extraction underperforms — so that, as more capable models become available, you re-process selectively, the few hundred pages that need it, rather than paying to re-run the entire archive. Re-processing becomes a targeted cost, not a wholesale one.

> **THE COST ARGUMENT IN ONE LINE** Per-token prices may fall; token volumes are exploding and the price floor is subsidised. An architecture that re-derives the extraction every time it is queried is exposed to all of it. An architecture that processes the corpus once, algorithmically, and serves validated data thereafter is not. For an internal build, this is the part most teams miss: it is not only an eighteen-month project — it is a permanent, uncapped token liability that grows with every user and every upward move in model pricing.

The MCP turn — why the route to the archive changes the bill

For context, an MCP (Model Context Protocol) is simply a connector/plug in that connects Pulse's data with your own AI (e.g. Claude, ChatGPT, Grok, etc), in your own account. It is connected in 1 minute, similar to connecting your email, folders or diary to your AI.

The section so far has argued that processing once is cheaper than re-deriving on every query. There is a concrete mechanism that decides whether a team actually captures that saving: how the model reaches the archive. The generic approach points an OCR-plus-LLM toolchain at the documents and re-runs it whenever someone asks a question — the full text of a drill-log passport pushed through the model every time.

The Pulse approach exposes the structured, validated store through an MCP, a structured interface that an organisation's own AI queries directly. On a Soviet-era corpus — the most token-expensive material there is — the gap between the two is at its widest, because the naive route pays the premium end of the cost curve on every single query.

Three things drive token spend on a technical corpus, and MCP-mediated access collapses all three. Context inflation — moving a 600-page Russian feasibility study through the model — disappears, because the model receives a compact structured result rather than the scanned text. Reasoning tokens spent re-extracting grades from a table disappear, because the grades were extracted once, algorithmically, at ingestion. And agentic re-reading collapses to a deterministic lookup.

In Pulse's own benchmarking a single decision-grade question answered through the MCP moved roughly five thousand tokens; the same question put to an unaided model re-reading raw PDFs moved on the order of one hundred thousand to over a million — up to 200X the compute, for an

answer with no source page attached. The multiple is illustrative, not an audited constant; the direction is not.

> **THE MCP COST PRINCIPLE — GEOLOGICAL** Process the corpus once; query and compound it forever. The archive is the worst-case exposure on the token cost curve, which is exactly why processing-once is where the saving is largest — and why an MCP over the validated store, not a toolchain over the scans, is the route that captures it.

How a technical team consumes this has come to matter as much as the extraction. Across live engagements three surfaces recur, and a working geologist wants all three:

- the Pulse platform for organised data presentation, search, mapping, screening and 3D visualisation;
- the validated corpus wired into the team's own AI (Claude/ChatGPT/Grok, etc) via the MCP, so a natural-language question — 'summarise these ten reports: what year, what work, drill holes, geochemistry, geophysics' — runs in real time against an indexed, cited store rather than a folder of scans;
- and a served layer that lands in the GIS toolset as a WMS feed into ArcGIS, or any other format.

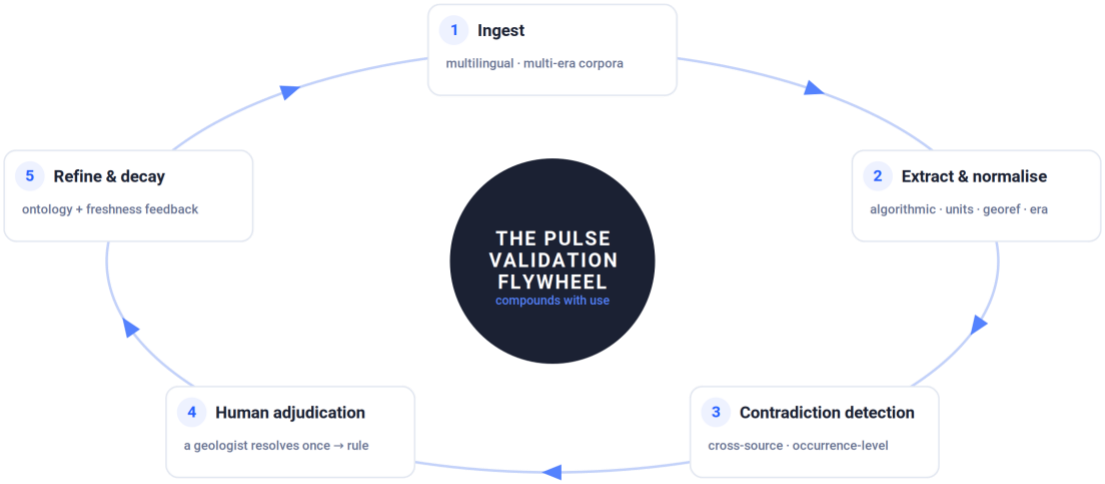
On the raw-MCP surface, the power users emerge: one Pulse client now writes custom porphyry-discovery prompts across four-and-a-half million pages of AI-native data. That surface is unbounded in power, and it is where the data-integrity discipline of the next section stops being optional. The default surface for the ninety per cent is chat, cited to the page; raw MCP is the power-user tier.

The field evidence is unambiguous on both cost and difficulty. An early client told us plainly that their blocker had been OCR and translation — they had hit page or megabyte limits with every off-the-shelf service they tried, and Pulse was already well past those; Pulse's own OCR counters have crossed more than thirty-five million pages.

On difficulty, a tier-one producer's three-week proof of concept reached about ninety-eight per cent on the drill logs it worked — with every row human-verified — while making the honest point that drill-hole extraction is not a three-week job but a multi-month build for a development team, and that table detection still missed one table in ten. The lesson is not that the work is cheap. It is that the expensive, hard part is paid once, by a specialist pipeline, rather than re-paid on every query by every geologist's model.

05 · The Pulse Validation Flywheel — Geological Lens

The same five-stage flywheel that underpins the platform for capital allocators underpins it for geological practitioners — with different stage emphasis. For a geological audience the load-bearing stages are extraction normalisation, occurrence-level entity resolution, and human-in-the-loop adjudication.



The Pulse Validation Flywheel — the five-stage loop that turns a raw corpus into a validated, self-improving record. A contradiction resolved once becomes a permanent rule, so the loop compounds with use rather than degrades.

Occurrence-level entity resolution

The unit of resolution that matters in a real corpus is not the document and is not the project. It is the occurrence — the named geological feature, mineralisation, or deposit referenced in a report. A single occurrence may be referenced under multiple names, in multiple languages, across hundreds of documents covering decades. A team that does not deduplicate at occurrence level cannot build a chronological history of any given deposit; the same deposit appears as three different entries in the database.

Across one frontier-exploration corpus, occurrences were deduplicated and cross-linked across roughly fifteen thousand document bundles, with location, mineralisation and grade-context summaries generated per occurrence and chronological histories assembled per occurrence — enabling queries of the shape “show me every gold grade above 1 g/t associated with this named occurrence across every report it is mentioned in, in chronological order.” That query is the geological equivalent of source-line lineage. It is what makes a corpus auditable rather than searchable.

Grade-to-occurrence mapping

Document-level grade extraction is a starting point. Decision-grade work needs grade extraction mapped to the specific named occurrence the grade describes — not just to the parent document. The relational layer has to be maintained explicitly: a single grade record carrying original disclosed value, normalised value, sample type, depth and interval where present, source page reference, occurrence

ID, and any lineage from a contradicting disclosure of the same figure across the corpus. Anything less is a starting point dressed up as an answer.

Human-in-the-loop, by design

The flywheel's fourth stage — human adjudication — is where domain expertise enters the system permanently. When a contradiction surfaces between two disclosures of the same occurrence, a working geologist resolves it, and the resolution becomes a rule. Future extractions of the same pattern are governed by the rule rather than re-litigated. Over the life of an engagement, the system accumulates the team's judgement; the team accumulates capacity, because the same problems do not have to be solved twice. This is the architectural reason a properly-built validation loop compounds with use rather than degrades — and the architectural reason internal builds stall: the loop only spins at volume, and most teams do not see enough disclosures to feed it.

The data-integrity protocol — knowing what you don't know

Extraction accuracy is necessary and not sufficient. What makes a geological answer defensible under a JV technical review is the discipline that governs what the system is allowed to assert — and, just as importantly, what it must refuse to assert. The load-bearing behaviour is that the pipeline knows what it does not know: it flags an illegible cell rather than inventing a plausible number, surfaces a confidence interval on an inferred coordinate rather than fabricating a location, and stops at 'if we don't know, we don't know — it can't make it up.' That refusal is the geological equivalent of the trust layer's first principle: a database is a convenience layer, not ground truth, and a historical figure is only defensible when it is traced to a dated source and the transformation from disclosed to normalised is visible.

In practice this is a blocking discipline, not a disclaimer. A self-monitoring quality layer predicts the expected extraction count per page and flags the pages that underperform; 'ridiculous' outliers — a hundred grams per tonne over a hundred metres — are caught and quarantined rather than passed downstream where they erode a geologist's trust in the whole corpus. On the hardest documents, where the pipeline cannot yet guarantee quality unaided, the answer is not to guess: it is human-in-the-loop, row by row, at a hundred per cent validation — the exact discipline a tier-one producer required before it would trust drill-hole output. The completion claim is itself a figure: 'the extraction was checked' is a claim about the work and is verified against the work, because a false assurance that the numbers were checked is worse than any single wrong grade — it disarms the reviewer's scepticism at the moment it was warranted.

Verification that reaches outside the document

Self-policing catches what a figure contradicts inside the corpus. The harder discipline is confirming a figure against what sits outside the document it was disclosed in. Where a decision warrants it, Pulse cross-references an extracted value against the rest of the validated record — other reports on the same occurrence, adjacent projects, tenure and regional context — and, beyond that, proprietary algorithms and AI agents verify independently against sources the disclosure itself does not contain. A number is not trusted because it was stated cleanly; it is trusted because the wider record agrees with it. And because every layer is normalised into the same structure, each new disclosure Pulse ingests makes every figure already in the system more defensible — the record does not just grow, it corroborates itself.

BENEFIT BEFORE BELIEF — THE GEOLOGIST'S LADDER A conservative technical team benefits before it adopts anything it would call AI. The day a Russian report is OCR'd, translated and source-referenced to the page, the search for where a grade lives is gone — a fully manual workflow, faster, no model in sight. From there the rungs hold only because the one beneath is built: source-referenced translation, then hybrid search, then 99%+ extraction, then spatial linkage, then the corpus wired into your own AI through the MCP. The rungs cannot be skipped, which is exactly why the data foundation is not optional.

A NOTE ON SECURITY AND THE INGESTION MODEL Some teams will not be comfortable routing proprietary disclosures through an MCP-style integration with a third-party AI front-end — even when the front-end is theirs. That is a fair concern. The same pipeline runs in encrypted private ingestion zones for clients that need it (including on the client's own servers, in-house): documents enter a permissioned environment, are processed under that environment's controls, and never sit alongside another client's data. Architecture and outcomes are identical; the perimeter is tighter.

06 · Where The Asset Actually Is — Georeferencing and GIS

Geospatial work is its own discipline. It overlaps with normalisation, but it is not the same thing. Three questions, all load-bearing: when a coordinate is stated, is it correct after transformation? When a coordinate is not stated, where is the asset? And once the answer to either is in hand, does it land in the GIS toolset the geologist already works in?

When a coordinate is stated

Normalise it. Pulkovo → WGS84. Krasovsky 1940 → WGS84. Local datum → WGS84. Always preserve the original. Surface the transformation logic. Carry confidence levels for any inferred element. This sounds rote, and it is the kind of work vendors who treat georeferencing as a library function get subtly wrong on real data. The Soviet datum landscape is not unified — different oblasts adopted different conventions at different times, and individual field reports were sometimes referenced to a local control network that does not appear in any standard cartographic library. Production-grade georeferencing requires a fallback chain, a confidence-level reporting layer, and explicit flagging of any coordinate the system could not transform with high confidence.

Detection before transformation. The transformation maths is the easy part; the load-bearing step is correctly identifying which coordinate reference system a given set of coordinates is actually in — projection and zone, datum, or a local control network — before any conversion runs. Drill data most often arrives as eastings and northings with the zone unstated, and a single project, sometimes a single report, can carry a mix of systems. Get the identification wrong and two adjacent holes plot hundreds of metres apart. Once the system is correctly identified, conversion to a common WGS84 reference is deterministic.

Verify against everything, not just itself. A dataset is not trusted to be internally consistent just because each coordinate transformed cleanly. Verification runs in layers, and the layers compound.

Internal consistency. Each coordinate is checked against the rest of its own dataset. A collar that plots anomalously far from its neighbours — the signature of a mis-projected or mixed-projection record — is flagged and withheld rather than surfaced as valid.

Cross-reference against the corpus. Because every asset, tenement outline, adjacent project and report Pulse holds is normalised into the same structure, a coordinate is also checked against the

public record around it — the licence boundary it should fall inside, the other reports filed on the same ground, the regional and deposit-type context. A point that clears its own dataset but contradicts the ground around it is still caught.

Independent verification beyond disclosure. Where the decision warrants it, proprietary algorithms and AI agents go past what the document itself discloses to triangulate the location independently — not adopting the stated figure on trust, but confirming it against sources the report does not contain.

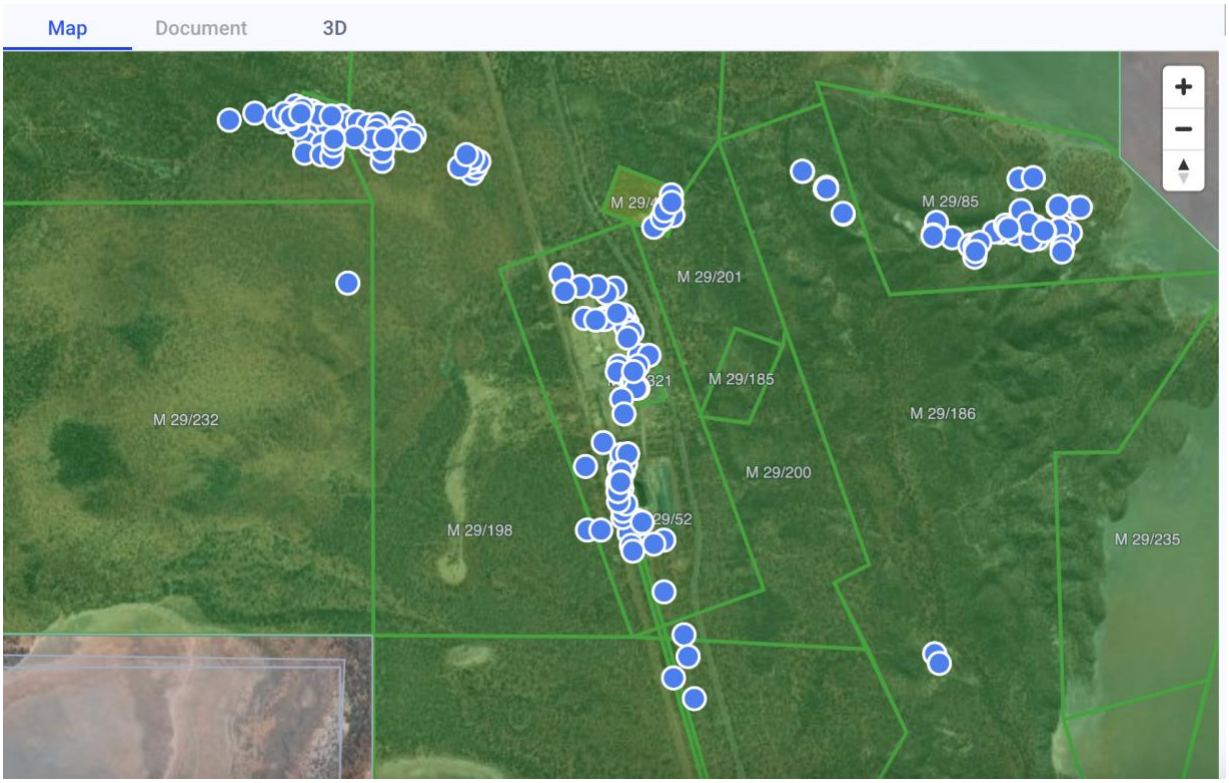
HOW AND WHY IT COMPOUNDS No single check is doing the heavy lifting. Each layer is only possible because the one beneath it is already normalised into the same structure — and each new dataset Pulse ingests makes every coordinate already in the system more defensible, not just the new one. The corpus does not only get bigger; it gets more self-consistent. That is the difference between a georeferencing library that transforms a coordinate once and a trust layer that keeps re-earning it.

When a coordinate is not stated

Infer it. Carefully. Then, check until a resolution is found. Trained inference can locate an asset from the textual content of disclosures — regional context, deposit-type cues, named geographic features, drainage references, adjacent project mentions, and the occasional verbal description of the form “approximately 12 km north-northwest of the named settlement in the named oblast.” Done correctly, the inference does not invent coordinates; it triangulates a location and returns a confidence interval, which is surfaced to the user. A high-confidence inference is treated differently from a low-confidence one in every downstream workflow. This is the work that makes a sparse historical archive geographically tractable. Without it, half the corpus drops out of any spatial workflow because no coordinate is stated; with it, the corpus can be filtered, mapped and analysed alongside the disclosed-coordinate fraction.

When the coordinate exists, where does it go next?

This is the question most data platforms forget to answer. A correctly transformed coordinate that lives only inside the platform is not very useful to a geologist who works in ArcGIS, QGIS, or LeapFrog all day. The expectation, fairly, is that mining-grade data plugs into the GIS toolset they already trust — not that the toolset gets replaced. Pulse outputs are designed to integrate, not displace. Coordinate sets export as shapefiles, GeoJSON, or KML. Map outlines export with metadata and lithological-unit attribution attached. Drill collars export as point layers with their full validated attribute table — grade, interval, sample type, source page, occurrence ID. Lithological units extracted from raster geology maps are cross-referenced to the report text that names them, and that cross-reference is preserved through the export.



Drill collars extracted from digitised reports, georeferenced to WGS84 and rendered over tenement outlines — the same validated points export as shapefiles, GeoJSON or KML into ArcGIS, QGIS or LeapFrog.

STATE OF PLAY Not every visualisation is yet a one-click integration. Some workflows still require a deliberate export step. The roadmap is to remove those steps in order of how often the team actually uses them. We will not claim seamless when it is not. What matters is that the data, when it arrives, is correct and traceable; the convenience layer above that follows. Satellite-imagery interpretation — identifying open pits, tailings facilities, and surface signatures of mining activity at scale — is a logical extension of the inference layer and sits on the roadmap. It is not yet deployed. We mention it for completeness, not for credit.

07 · The Interface Gap — from queryable data to geologist-usable workflow

There is a gap between “data is ready” and “geologists can use it.” It is its own discipline, and most platforms do not address it, because most platforms were designed by data engineers for data engineers.

A working geologist at a frontier exploration team put it directly in a recent scoping conversation: being geologists, and not being across the IT side, what they struggled with most was not the underlying data work — it was the interface of how to actually access the AI-translated documents and visualise the results spatially. That sentence carries weight. The data work, however hard, can be solved at the data team’s desk. The interface gap can only be closed at the geologist’s desk.

Below is the five-step workflow the way a working geologist actually experiences it. Each step has its own automation discipline, and each is somewhere a team can be capacity-bound, partially supported, or fully equipped.

The five-step practitioner workflow — from a raw scan to a layer that queries natively inside ArcGIS, QGIS or LeapFrog. Each rung holds only because the one beneath it is built.

Step 1 — Translate

What good looks like: full-corpus OCR with structural reconstruction (tables, headers, footnotes preserved), bilingual outputs (original alongside English), and translation that respects technical terminology — "Redkozemelne elementy" translates as "rare earth metals," not "rare ground earths"; "Rossypi Zolota" is preserved as a technical term, not flattened to "gold deposits." Original Cyrillic is one click away.

What gets left behind: teams that paste a 600-page Russian feasibility study into a generic translation tool and end up with a document the next step cannot use — translation that loses table structure has destroyed the data.

Step 2 — Search

What good looks like: hybrid search across keyword and semantic layers, with citations to the page and table cell that produced the answer; cross-language queries (an English question against a Russian corpus) handled correctly; misspellings of Russian or place names returning results, not zero hits. The geologist trusts the search because every answer comes back with a source they can open.

What gets left behind: keyword-only search over OCR'd text — it works for any term the geologist remembers exactly, and fails for everything else, which is most of the questions worth asking.

Step 3 — Extract

What good looks like: the Section 03 numbers — 99%+ accuracy on the metrics that drive decisions, original disclosed value preserved alongside the normalised one, sample-type tagging, outlier flagging, source-line lineage on every cell, drill grades mapped to the specific named occurrence they describe.

What gets left behind: teams that treat extraction as a one-off transcription job. Extraction without structure is just retyping. Structure without lineage is just guessing. Both fail under JV technical review.

Step 4 — Spatially link

What good looks like: drill collars as point layers with attribute tables, map outlines as polygon layers with metadata, lithological units as both raster classification and text-cross-referenced features; mineral occurrences with grade-context summaries, location descriptions and chronological histories attached. The geologist moves from a search result to the spatial view to the source document and back, without losing context.

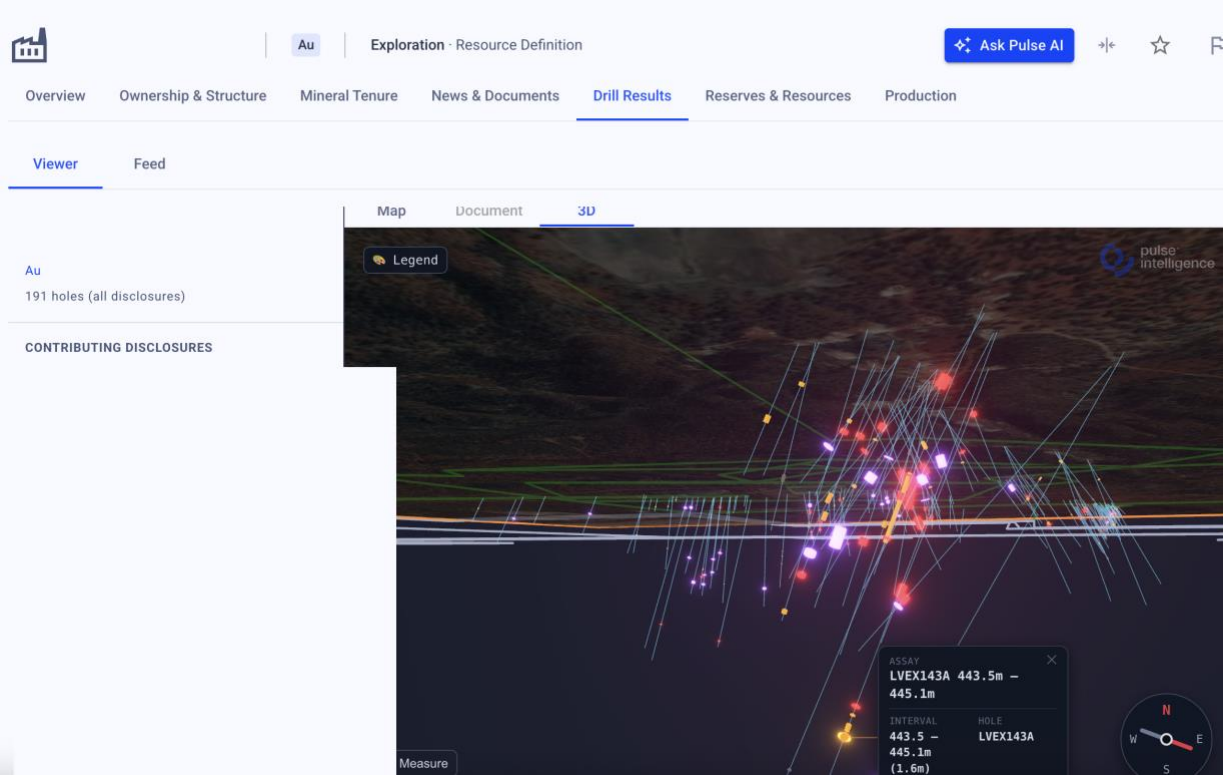
What gets left behind: workflows that treat the spatial layer and the text layer as separate systems — the geologist ends up jumping between two windows and reconciling by eye.

Step 5 — Visualise and query in the GIS toolset

What good looks like: shapefiles, GeoJSON, KML, raster GeoTIFFs, all carrying the validated attribute tables; spatial queries ("all drill intercepts above 1 g/t Au within 5 km of this licence boundary") working natively against the layer; filters by commodity, grade range, occurrence, source document, and time period, all available without leaving the GIS environment.

What gets left behind: data that is technically validated but operationally unusable — the geologist who has to download a CSV, reformat it, re-georeference it, and import it into QGIS to see anything spatial has just been told to do an hour of work for a five-minute question.

> **THE INTERFACE PRINCIPLE, IN ONE LINE** Mining-grade data infrastructure that does not show up in the geologist's working toolset is not, operationally, mining-grade. The discipline is not done at the data team's desk. It is done where the inference is.



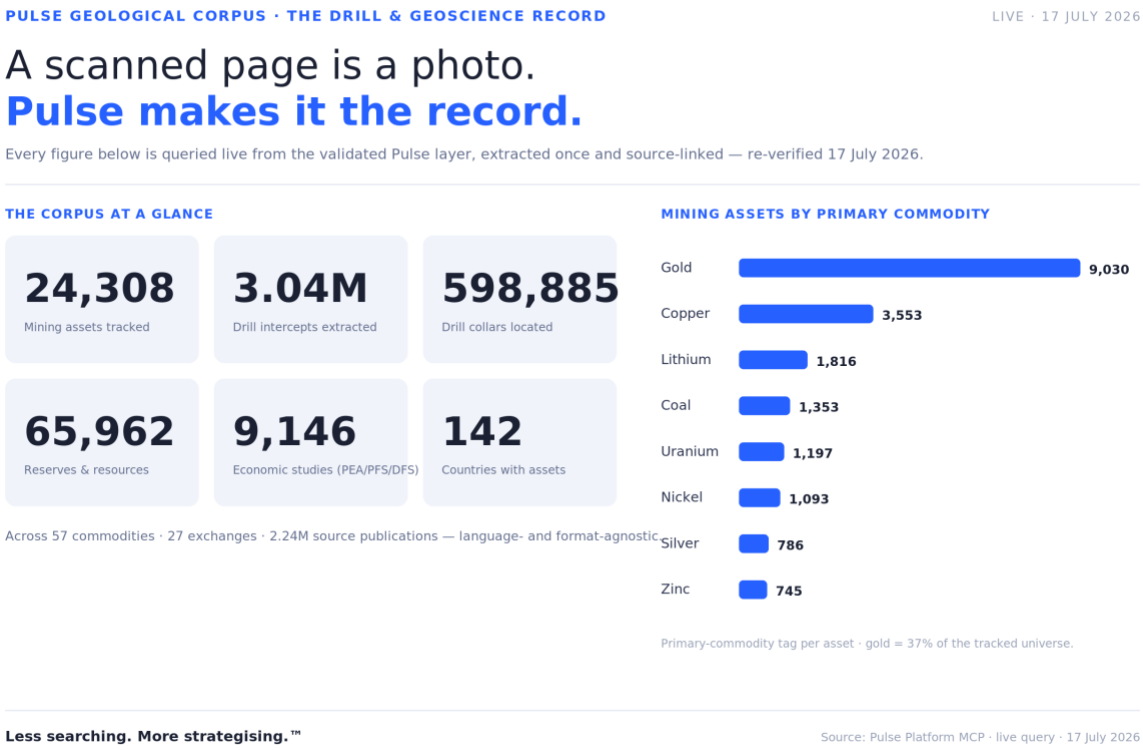
Drill results from the extracted corpus, grade-coloured and interrogable in 3D inside the platform — every intercept clicks through to the source page it was disclosed on.

08 · What We See In The Field

We chose five anonymised archetypes drawn from live or recent engagements. The numbers are real. They deliberately span the range — the Cyrillic extreme, a clean English producer corpus, the live daily disclosure flow, and a national open-file archive in English — because this paper argues that the pipeline does not care which.

Three of the five are Soviet-era Cyrillic, because that is the hardest case and the truest demonstration of capability; the other two are English and born-digital, because that is most of the market, and the same pipeline runs on both without a line of special-casing.

Over the last two years, we have digitised more than 35 million pages, including all forms of geological data, all formats of technical reports and daily market data and documents (the latter two daily, in real time).



The Pulse geological corpus, queried live and re-verified on 17 July 2026 — every figure extracted once and source-linked, not rounded or inferred.

Archetype A — an exploration archive in Central Asia

A late-stage exploration project working a Soviet-era archive across an asset of national-scale strategic importance. The corpus arrived as roughly fifteen thousand document bundles — around three million pages — with 537,000 attachments, of which 320,000 were image-format (maps, charts, photos, plates). Russian-first; the team's own internal toolset until then had been a script-based fuzzy keyword search with frequency-ranked results.

The foundation was rebuilt from the ground up: full-corpus OCR with structural reconstruction; bilingual outputs; a structured data layer underneath the text; one hundred elements normalised across multi-language naming conventions, including Soviet-specific terminology that does not translate one-to-one; approximately 950,000 grade records extracted with sample-type tagging and outlier flagging; occurrence-level entity resolution with chronological history across the entire corpus. Full corpus end-to-end processing time after pipeline optimisation: approximately forty-eight hours. Above the original proposal scope: an MCP integration enabling the team to query the full corpus via natural language through their preferred AI front-end; ring-fenced search environments for JV partners

(so a partner with access to a sub-set of report IDs sees only the reports they are entitled to see); and a self-monitoring quality layer that predicts expected extraction counts per page and flags pages where extraction underperforms for future re-processing as more capable models become available.

Archetype B — a tier-one global producer

The pilot anchored Section 03. The headline numbers bear repeating in the case-study context: 71 documents, 11,249 drill-hole assay rows, 34 distinct geochemical elements, 958 OCR pages at ~97% word / ~99% character accuracy on Russian text including handwritten, stamped and degraded scans, 509 tables extracted to structured CSV. On the scored ground-truth corpus: 99.21% element accuracy, 99.65% recall, 99.06% precision, 99.71% interval-length equality. Zero hard processing failures across 164 input files. What this archetype shows that the previous one does not: the pipeline performs to threshold on a different scoring framework, applied by a different technical team, against a different corpus. The numbers are not corpus-specific. They are a property of the system.

Archetype C — an explorer in active POC scoping

An exploration team holding a portfolio of Soviet-era licence ground in a frontier jurisdiction, with five concurrent problems articulated in their own technical language during scoping: translation of Soviet reports at scale; AI-assisted natural-language search across the translated corpus; recognition and extraction of drilling data; spatial linkage between drillholes, map outlines and the reports they reference; and spatial search natively in ArcGIS. Their honest framing of where they were stuck: as geologists, not IT specialists, the hard part was not the underlying data engineering — it was the interface where the AI-translated documents and the spatial visualisation met their workflow. The POC scope settled on a small, deliberately-limited subset — around five percent of the documents available — with the goal of running every step end-to-end on a high-quality output. The principle: do less, do it well, surface the edge cases on a small corpus before scaling.

Archetype D — the live market-data flow (English, born-digital and scanned)

Not every corpus is an archive. The largest one Pulse runs is the live disclosure flow — the announcements, quarterlies, MD&A and technical reports that hit the exchanges every day, overwhelmingly in English. The same extraction pipeline that reconstructs a 1970s passport parses the 43-101 filed this morning: units normalised, reporting standard tagged, every figure traced to the page it came from. Across the platform that has meant more than two-and-a-quarter million publications ingested and 9,146 economic studies extracted with page-level source links — an NPV you can click straight through to the page of the study it was disclosed in. What the archive engagements prove about the past, this one proves about the present: the language and the file format are variables; the discipline is the constant. And it runs daily, not once.

Archetype E — a national open-file archive in English (WAMEX)

The clearest proof that none of this is about Cyrillic is a corpus with no Cyrillic in it at all. Western Australia's open-file exploration record — WAMEX — runs to more than 180,000 reports (c10 million pages): plus decades of historical drill results and geochemical history, filed in English, scanned, and effectively unsearchable as data. The same pipeline that handles a Soviet archive turns it into structured, validated, spatially-linked intelligence — drill collars, intercepts and geochemistry tied to occurrences and coordinates, queryable in the geologist's own GIS. No translation step, no Cyrillic OCR, no Pulkovo transform; just the format problem — a scanned page is a photo, not information — solved at national scale.

09 · Why Your Existing Toolset Can't Do This

Existing tools are excellent at what they were built for. The work in this paper is not what they were built for.

Generic OCR is not technical-document OCR

Basic OCR is the equivalent of handing a 600-page feasibility study to a junior intern on their first day and asking them to type it out. They will recognise some letters, skip the tables, scramble multi-column pages, mangle technical terms, and drop entire sections. The output is roughly 5% usable text on a real technical corpus — adequate for a clean restaurant menu, useless for a feasibility study. Advanced OCR with structural reconstruction — layout-aware extraction that rebuilds tables, preserves columns, retains headers and footnotes, fixes broken characters, and reconstructs spacing — sits at 97%+ on Russian text including handwritten, stamped and degraded scans, in production today. The difference is not incremental; it is the difference between a system the AI can understand and a system it cannot.

Generic LLMs hallucinate technical figures

Conversational language models are powerful tools, and they cannot be trusted as the truth-extraction layer of a mining-grade pipeline. They produce confident answers that are subtly wrong. They cannot apply deterministic extraction rules. They cannot consistently cite their sources. They do not know your reports, your assets, your terminology, or the specific definitional rules of your reporting regime. Asked to find drill grades above 1 g/t in a 200-page feasibility study — Spanish, Portuguese, Russian, Mandarin, English, it does not matter — they will produce a list, and some of the entries will be invented. A defensible architecture uses language models downstream of the truth layer — for summarisation, conversational interfaces, and answer assembly — after structured data has been validated. Truth is extracted by deterministic algorithmic logic. The model is the brain; the trust layer is the memory; the trust layer must be built.

Why generic enterprise AI inside Office 365 isn't this

A version of this comes up in nearly every scoping conversation: "Surely Copilot can do this within the Microsoft ecosystem?" Copilot is a competent conversational layer over text — it will read a Russian feasibility study and answer questions about it, and for general document Q&A on clean material it is genuinely useful. It is not a substitute for the disciplines this paper has walked through. Copilot does not maintain a definitional ontology for mining reporting standards. It does not reconcile JORC, NI 43-101, GKZ. It does not transform Pulkovo coordinates to WGS84. It does not preserve the original disclosed value alongside a normalised one with the transformation logic visible. It does not surface contradictions between two disclosures of the same figure across a 2.2 million-publication corpus. It does not maintain a benchmarking framework with 50–200 hard samples per pipeline step. It does not run encrypted private ingestion zones. And — per Section 04 — it carries the full token-cost exposure of the re-derive-every-query architecture, because that is exactly what it is.

QGIS, LeapFrog, ArcGIS, Adobe, Excel

All excellent for what they were built for. None is a substitute for a data platform with a domain ontology, a validation layer, and a freshness loop. A team that runs technical due diligence on Excel does not have a technical due diligence platform; it has Excel. The distinction matters when the corpus crosses the threshold of what a person can read, validate and reconcile inside a working week. The

right relationship between Pulse and these tools is integration, not replacement: the geologist works in QGIS or ArcGIS or LeapFrog as before; Pulse arrives as a layer they can pull from.

WHAT WE SHARE, WHAT WE DON'T We describe the architecture of the Pulse trust layer at the layer level: definitional ontology, contradiction detection, freshness decay, human-in-the-loop adjudication, benchmarking framework with 50–200 hard samples per pipeline step. The implementation — specific prompts, schemas, vendor stack, and the resolutions encoded as rules — stays internal. That asymmetry is deliberate: it is what makes the flywheel defensible.

10 · A Self-Diagnostic For Technical Teams

Eight questions. Answered honestly, against your own organisation, in the next ninety seconds. Each "no" is a wedge.

#	QUESTION	IF NO…
1	Have you ever ingested a Cyrillic, Spanish, Portuguese or French technical report and made it queryable in your team's working language without losing the table structure?	Half the corpus that will matter in the next decade is invisible to your current toolchain.
2	Can you take a drill intercept from a 1970s report and reconcile its interval definition to a JORC-equivalent measured category, with the reasoning visible?	You are inheriting historical methodology without auditing it.
3	Can you trace any single grade in your resource model back to the report page and table cell it was extracted from?	Your model has lineage gaps. Defending it under JV technical review is fragile.
4	Do you know which of your historical assets carry coordinates referenced to Pulkovo, Krasovsky 1940, or another non-WGS84 datum?	You are likely carrying one or more thirty-degree errors you do not know about.
5	When a peer restates a reserve, how long until it flows through your peer-comparison model?	If days or weeks rather than hours, the model is misleading you for that period.
6	Could a JV partner audit your extraction lineage end-to-end if asked tomorrow?	Your audit trail is tribal knowledge. Tribal knowledge does not survive staff turnover.
7	When validated data is ready, does it show up in the GIS toolset your team already works in — ArcGIS, QGIS, LeapFrog — as a layer, or behind a separate login?	The interface gap is open. The data is technically usable and operationally not.
8	Do you know what your AI-on-documents cost looks like in three years — and which way it moves?	Per Section 04, you are carrying an unpriced, uncapped liability.

Three or more "no" answers indicate the team is operating on infrastructure the corpus has outgrown. Five or more indicates the cost of bad data has crossed from a productivity problem to a technical-defensibility problem.

11 · The Discipline Beneath The Discipline

Geology is a discipline of careful inference from incomplete data. The geologist takes drill intercepts, structural measurements, geochemical signatures, and a modelled understanding of how the deposit formed, and produces an estimate of what is there. The estimate is defensible because the method is rigorous, the assumptions are explicit, and the data is curated.

Geological data normalisation is the same discipline applied one layer below — to the data the geology itself is built on. A drill intercept is itself a derived figure; behind it sits a sampling decision, an instrument, a laboratory, a unit convention, an interval definition, a coordinate, an era. A team that normalises that derived figure with the same rigour the geologist applies to the inference is doing the work the corpus deserves. A team that does not is leaving load-bearing gaps in the foundation of every estimate that follows.

Above that data layer sits the second discipline this paper has named: making the result genuinely usable to a working geologist. Solving the data layer without solving the workflow layer leaves the geologist with a queryable corpus they cannot operate. Solving the workflow layer without the data layer leaves them with a usable interface over information they cannot trust. Both have to be done. Both deserve respect as disciplines.

> **A PRACTICAL TEST** The cleanest way to evaluate any candidate normalisation pipeline — ours or anyone else's — is to put a real drill intercept through it. Take one from a report that has given your team trouble. Ask the vendor to return the full normalisation trace: original disclosed value, normalised value, every transformation step, the governance behind each, the source page reference, and the lineage from any contradicting disclosure across the corpus. If the trace comes back complete, defensible and reproducible, the discipline is in place. If not, the answers above are marketing.

This paper has been a long answer to a short question: what does it actually take to do that work well, at scale, on real data? The short answer is two years of blood, sweat and tears, and an architecture that compounds rather than degrades.

12 · About Pulse Intelligence

Pulse Intelligence is an AI infrastructure company that replaces manual technical and investment workflows in capital-intensive industries, with a deep focus on mining, where we have more than twenty years of experience.

Our infrastructure and technology automate the data extraction, validation, normalisation and structuring work that technical teams and analysts currently do by hand — cutting workflows that take ten to thirty hours down to under an hour, with full source-line lineage on every figure.

Platform scale (live)

- 4,550+ mining companies tracked — 3,099 public, 1,292 private, 162 government
- 24,308 mining assets tracked and monitored (17,744 primary + 6,564 sub-assets)
- 2.24M+ publications ingested and processed. 35M+ pages
- 27 stock-exchange integrations (NYSE, NASDAQ, LSE, TSX, TSXV, ASX, JSE, HKEX, Euronext, SIX, and others)
- 57 commodities tracked, 27 with live daily price feeds
- Language- and format-agnostic: English (born-digital and scanned), Russian / Cyrillic / FSU, Mandarin, Spanish, Portuguese, French, Arabic — Cyrillic is the accuracy proof, not the limit

Demonstrated technical capability

- 99.21% element accuracy / 99.65% recall / 99.06% precision / 99.71% interval-length equality on the hardest case worked to date — a Cyrillic Soviet-era ground-truth corpus (anonymised tier-one engagement)
- ~97% word / ~99% character OCR accuracy on Russian text including handwritten, stamped and degraded scans; comparable performance on the other major resource languages, available under NDA
- Multilingual normalisation across 100+ commodity / element naming conventions
- Occurrence-level entity resolution with chronological history across 15,000+ document bundles in production
- Same pipeline on the live daily disclosure flow and on national open-file archives in English — e.g. 180,000+ WAMEX reports (historical drill results and geochemistry) — alongside 2.24M+ publications and 9,146 economic studies extracted with page-level source links
- GIS-native outputs (shapefile, GeoJSON, KML, raster GeoTIFF) with full attribute lineage preserved through export

Companion paper

“The Trust Layer: From Co-Pilot to Autopilot in Mining Finance” — written for the capital allocators that the same practitioners serve. Available from the Pulse Intelligence team.

CONTACT	Will Coetzer, Director & Co-Founder — will@pulseintelligence.com
WEB	pulseintelligence.com

Less searching. More strategising.™

Sources for Section 04 (token economics): Epoch AI, LLM inference price trends; Andreessen Horowitz, “Welcome to LLMflation”; Gartner, 2026 analysis of agentic AI token consumption; VentureBeat, “Cheaper tokens, bigger bills”; Deloitte, “AI tokenomics: a CFO’s guide”; and Pulse platform benchmarking versus an unaided frontier model, June 2026.

Appendix A · The quantified ledger — re-verified July 2026

The narrative above uses illustrative figures to make an argument; this appendix separates them from the numbers meant to be defensible. Every platform-scale figure here was re-pulled live from the Pulse Platform MCP on 16 July 2026 under the data-integrity protocol, and supersedes any earlier figure in prior versions of this paper. Extraction and engagement figures are dated to their source and were produced on real ground-truth corpora, not clean benchmark examples. The compute and cost multiples are illustrative — drawn from Pulse's own benchmark instrumentation and client engagements, not audited constants — and are marked as such.

MEASURE	PULSE — VALIDATED LAYER	GENERIC AI / OFF-THE-SHELF · BASIS
Tokens per decision-grade query	~5,000 — one compact, cited, structured result	~100,000–1,000,000+ — re-read of the scans · Pulse benchmark, Jun 2026 (illustrative)
Relative compute per query	1×	up to ~200× · same benchmark (illustrative)
Off-the-shelf OCR ceiling on a dense technical scan	99%+ with structural reconstruction	~5% usable text / 90% ceiling — every cell re-checked by hand · engagement observation
Extraction accuracy — tier-one producer ground truth	99.21% element · 99.65% recall · 99.06% precision · 99.71% interval	not comparable · anonymised producer pilot, 2026
Extraction accuracy — Soviet drill passports (human-reviewed)	98.9% assay (2 errors / 180 rows); 100% row capture	not comparable · anonymised drill-passport engagement, 2026
Drill-log extraction — verified logs	~98% (AI + 100% human row validation)	not production-ready in a 3-week POC; multi-month build · tier-one producer POC, 2026
OCR on handwritten / stamped / degraded Cyrillic	~97% word / ~99% character	generic OCR fails on this material · producer pilot, 2026
Archive throughput	~3M pages → ~950k grade records in ~48 h; ~35M pages OCR'd to date	re-read on every query · Central Asia engagement; company instrumentation
Live platform scale (16 Jul 2026)	4,550+ companies · 24,308 assets · 2.24M publications · 3.04M drill intercepts · 9,146 economic studies	— · Pulse Platform MCP, re-verified

A NOTE ON DEFENSIBILITY Platform counts are live and grow week to week; the figures above are floors dated 16 July 2026. Extraction figures are stated with the corpus, the metric and the validation method, because a percentage without those three is marketing. The compute and cost multiples are directional, not audited.

APPENDIX B · Your Data, Your Boundary — Security, Residency and Exit

Geological disclosures are among the most sensitive data a company holds — pre-competitive drill results, JV-restricted technical work, and sovereign archives that predate the company that now owns them. A platform that ingests them earns trust only if the client's own security team can satisfy itself that sensitive data never leaves a boundary the client controls. Pulse is built for that review, not around it. The commitments below are architectural, not aspirational; where something is a roadmap item, we say so.

Isolation by design, not by policy

Every client runs in a dedicated environment — its own database, storage and compute — not a shared multi-tenant pool. Namespace isolation contains any incident to a single environment, which makes the cross-tenant data-leak that has embarrassed more than one AI product simply not applicable to Pulse. The strongest control is the one the architecture enforces whether or not anyone remembers the policy.

It can run on your own servers

Pulse offers three deployment models: hosted by Pulse in an isolated per-client environment; deployed by Pulse into the client's own cloud account; or handed over as pre-built images the client runs entirely independently. Data residency follows the model — the European Union by default (AWS, Stockholm), any AWS region on request, or the client's own infrastructure. For a team that will not route proprietary disclosures outside a perimeter it controls, the data, the database, the encryption keys and the AI access can all sit inside that perimeter. This is the answer that wins resource-sector and sovereign reviews, and it is true in concrete form, not in principle.

Encryption and access

Data is encrypted in transit (TLS 1.2+) and at rest (AES-256) under managed keys, with client-managed keys (BYOK) on the roadmap. Access is least-privilege by default: role-based permissions, multi-factor authentication, short-lived credentials, and no shared service accounts — with single sign-on (SAML/OIDC) available on request.

How the AI touches your data — two layers, two rules

Your private project data — the disclosures, the drill data, the archives you bring — is processed inside your isolated environment, is encrypted and opt-in, and is never shared with or exposed to another client. It stays yours; how it may be used beyond serving you is set by your agreement, not presumed by us. Separately, Pulse gets sharper for everyone from a shared layer built only from validated public records and anonymised, aggregated usage signals — never from your confidential content. Language-model calls are stateless, and you can point the AI at your own keys. Because Pulse exposes an MCP, it is built to the isolation reviewers now expect of one: the MCP operates inside your environment, scoped to your corpus, with credentials kept separate from the model and every answer carrying a citation back to a source page.

You own it, and you can leave with it

The client owns all of its data. Full export in open formats is available at any time, deletion on request, and — under the fully client-controlled deployment — the system keeps running even if Pulse were to discontinue. "No lock-in" is an architectural fact here, not a sales line: the exit is designed in, so choosing Pulse is never a one-way door.

Governance, stated honestly

Pulse is GDPR-compliant and built on AWS Well-Architected foundations, with monitoring, defined incident-response targets, and daily backups; a Data Processing Addendum is available on request, and the full internal Trust Framework is shared under NDA. Independent third-party certification — SOC 2 Type II — is a roadmap item, not a badge we hold today. We state it that way on purpose: claiming an attestation you do not have is the one thing that destroys a security review, and this paper's argument is that honesty about limits is what survives one.

> **SAID HONESTLY** The load-bearing claims here are architectural — per-client isolation, an on-your-own-infrastructure option, encryption in transit and at rest, and a guaranteed exit — and they are true today. The operational specifics behind them (response-time targets, key-rotation cadence, backup and recovery intervals) live in our internal Trust Framework, which we share under NDA rather than pin to a public page. We would rather a reviewer find a limit stated here than discover it themselves.

In one live engagement, the deployment was scoped explicitly around a dedicated client environment — the client's data, the client's encryption boundary, the client's own database — with the live MCP querying their corpus inside it. The security posture is not a page we wrote; it is a boundary clients already run behind.