OPEN DATA · ARTIFICIAL INTELLIGENCE · PUBLIC INFRASTRUCTURE

TECHNICAL REPORT · TD-TR-2026-03

onem-tunisia-mcp

A qualifier-aware MCP server over Tunisia's national energy time-series.

AUTHOR

Tarek Gasmi

PUBLISHED

June 2026

STATUS

Independently produced and published

TYPE

Technical report

ABSTRACT

onem-tunisia-mcp is a Model Context Protocol server that makes Tunisia's national energy statistics directly queryable by LLM clients. The source data is not an API but a corpus of recurring PDF reports published by the Observatoire National de l'Energie et des Mines (ONEM); the system's defining work is the discipline that carries a figure intact from a PDF table to an LLM's sentence. A coordinate-aware extractor lifts cells with per-cell provenance into a long-format DuckDB store whose schema makes every qualifier — calorific basis, period type, geography scope, aggregation role, data status — a first-class column rather than a derivable afterthought. The serving layer is built on one principle: a category error should be hard to even express. The server refuses to compare a PCI figure with a PCS one, an annual figure with a year-to-date one, or a total with its own components; it distinguishes “out of scope” from “no data”; and it never returns a bare number. A battery of fifteen relationship and arithmetic checks, a three-layer evaluation, and an adversarial multi-agent audit stand behind the store. This report documents the system as it ships today — a source-distributed local stdio server bundled with its built store — together with the rationale behind its principal decisions and its candid limitations.

tanitdata — autonomous studio · Tunis, Tunisia
tanitdata.org · contact@tanitdata.org

1. Introduction

Tunisia's **Observatoire National de l'Energie et des Mines (ONEM)**, within the Ministry of Industry, Mines and Energy, is the country's reference source for energy statistics: production and consumption of oil, gas, and electricity; the national energy balance; trade and royalties; renewable capacity. ONEM publishes this not as a data service but as a stream of **recurring PDF reports** — the annual *Bilan National de l'Energie*, the *Memento* ("Chiffres clés"), and the periodic *Conjoncture énergétique* bulletins — spanning 2010 to 2026. The figures are public and authoritative. They are also, for any computational use, effectively locked away.

The friction is not a difficult API; it is that **there is no API**. The numbers live in PDF tables whose layout drifts between editions, and they are surrounded by distinctions that look like formatting but are load-bearing: natural gas reported on a **lower** calorific basis (PCI) in one table and a **higher** one (PCS) in another, where PCI $\approx 0.9 \times$ PCS; a figure that is a full-year **annual** total beside one that is a **year-to-date** cumulation to a cutoff month; electricity sales quoted **local** in one line and **including exports** in another; a partition total printed next to its own components. A reader who treats any of these as interchangeable produces a number that is plausible, specific, and wrong. The reports also defer whole families — prices, transit volumes, refining detail — that a casual reader cannot distinguish from data that simply was not found.

onem-tunisia-mcp makes these statistics queryable through an AI assistant while keeping every one of those distinctions intact. It is a server implementing the **Model Context Protocol (MCP)**, the open standard by which LLM clients call external tools. A user runs the server locally, connects it to an MCP-compatible client — Claude Desktop, Claude Code, Cursor, VS Code, or any other — and asks questions in natural language; the model selects and chains the server's tools to find a series, fetch it with its full qualifier set, and answer with cited figures. The software is published as the repository `onem-tunisia-mcp`; this report refers to the software by that name and to **tanitdata** as the publishing studio.

Distribution model and audience. This report documents the system as it actually ships today: as **source code bundled with its built data store**, distributed for local use. A user clones the repository, installs three Python dependencies, and launches the server as a local **stdio** process that their MCP client spawns. There is no hosted endpoint and no network listener — the server communicates only over the standard input/output of the process its client started, and reads its own local database **read-only**. This bounds the audience honestly: the current form serves a technically capable user comfortable cloning a repository and editing a client configuration file. It removes the need to read PDFs or reconcile editions; it does not yet remove the need to run a local server.

The contribution here is not the protocol, which is a standard, nor the idea of querying a dataset, which is routine. It is the **end-to-end discipline** that lets a figure travel from an ONEM PDF cell to an LLM's sentence without losing the qualifiers that make it true: a coordinate-aware extractor that records cell-level provenance, a schema that treats every qualifier as a first-class column, a serving layer engineered so that a category error is hard to even express, and — standing behind all of it — a battery of arithmetic and relationship

checks, a three-layer evaluation, and an adversarial multi-agent audit.

2. Background

2.1 The Model Context Protocol

The Model Context Protocol is an open standard for connecting LLM applications to external tools and data sources. A client (the LLM application) launches or connects to one or more servers, each advertising a set of *tools* — named, typed, documented functions the model may call. The protocol is transport-agnostic; the transport relevant here is **stdio**, where the client spawns the server as a subprocess and communicates over standard input/output. onem-tunisia-mcp runs over stdio only.

MCP is treated here as settled infrastructure, not a research contribution. What matters for this report is the boundary it gives us: the server's job is to present a small, well-documented set of tools whose descriptions and return values are good enough that an LLM uses them correctly — and, crucially, *cannot* use them to assemble a qualifier-stripped wrong answer.

2.2 The ONEM corpus and its traps

The source material is a corpus of PDF reports in three families, registered with full provenance (source URL, SHA-256, publication date) in a manifest. The acquisition layer records **195 source entries** across the corpus; the time-series build draws on the canonical, machine-readable editions within it — principally the tabular Conjoncture series (81 canonical editions span late-2019 to 2026), the 2024 Memento, and the Bilan matrices.

Several properties of this corpus shape every later decision and recur throughout the report:

- **The data is in PDF table geometry, not text order.** Values must be read by coordinate, not by reading-order text: the Conjoncture tables interleave a *Réalisé* column with baseline and year-to-date columns, and a naïve text read mis-assigns them by one column — a trap that, left unguarded, silently swaps a current figure for a prior-year one.
- **Calorific basis is a genuine physical distinction.** Natural gas is reported on both PCI (lower/net heating value) and PCS (higher/gross), related by $1\,\text{PCI} = 0.9\,\text{PCS}$. The two are different numbers for the same gas and must never be blended or compared as if equal.
- **Period type is a trap.** The same indicator appears as an **annual** total and as a **year-to-date cumulation** carrying a cutoff month; the YTD value is not an annual value and the two share no meaning.
- **Scope and geography qualify the number.** Gas as "commercial dry" versus a broader "primary" aggregate; crude including or excluding primary LPG and condensate; electricity sales local versus including exports. Same indicator, different boundary.
- **Totals are not leaves.** A table prints partition components and a grand total; some

tables print *two* alternative partitions of the same total (gas demand by usage and by pressure). Summing the wrong combination double-counts.

- **Deferred is not absent.** Whole families — prices, trade values and volumes, refining and exploration KPIs — are not yet ingested. They must be reported as out-of-scope, never as "no such data," or the system launders a coverage gap into a false denial.
- **French is canonical; Arabic is a translation.** Arabic editions are registered but contribute **zero** ingested observations, to avoid double-counting a translated figure as an independent source.

3. The Data Pipeline: from PDF to Store

The system is two artifacts joined by a build: a **pipeline** that turns the PDF corpus into a modeled store, and a **server** that exposes that store to an LLM. The store ships pre-built, so an ordinary user never runs the pipeline — but it is fully re-runnable from the source PDFs, and its discipline is where the system's trustworthiness is established. Figure 1 shows the whole path.

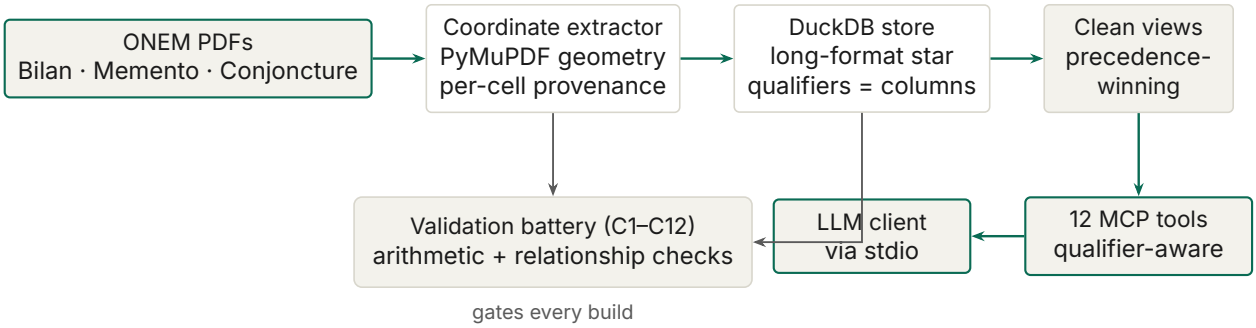

Figure 1. End-to-end pipeline. A coordinate-aware extractor lifts PDF cells, with cell-level provenance, into a long-format DuckDB star schema in which every qualifier is a column. A fifteen-check validation battery gates each build. The MCP server serves only from precedence-winning clean views, exposing twelve qualifier-aware tools to a local LLM client over stdio. The same qualifiers and provenance thread the entire path.

3.1 Coordinate-aware extraction

The extractor reads the PDF text layer with geometry (via PyMuPDF), clustering numeric fragments into cells by their coordinates rather than by reading order. This is not fastidiousness: reading-order text in the Conjoncture tables mis-assigns columns, and numeric fragments such as `5` and `024` must be merged by their edge gaps to recover `5 024` rather than read as a stray `5`. Born-digital text-with-geometry is the default; OCR is admitted only as a fallback and is flagged when used. Every observation records how it was obtained — an `extraction_method` and an `extraction_confidence` — and a cell-level `source_cell` ("row=… | col=…") so any served figure can be traced to the exact cell of the exact edition. Cells the extractor cannot align with confidence are marked low-confidence and excluded from the default surface rather than guessed.

3.2 The schema: a qualifier-preserving star

The store is a long-format ("tidy") star schema: one fact table, `observation`, with conformed dimensions (product, flow, sector, region, field, level, producer) and a set of **qualifier columns that are part of the fact, not metadata to be reattached later**. Calorific basis, period type and its cutoff month, scope, geography scope, aggregation role, and data status all sit on the observation row. The design choice is deliberate and is the hinge of the whole system: because a category error upstream becomes a wrong sentence downstream, the qualifiers that prevent it are carried as first-class, queryable fields all the way to the serving boundary. Figure 2 sketches the shape.

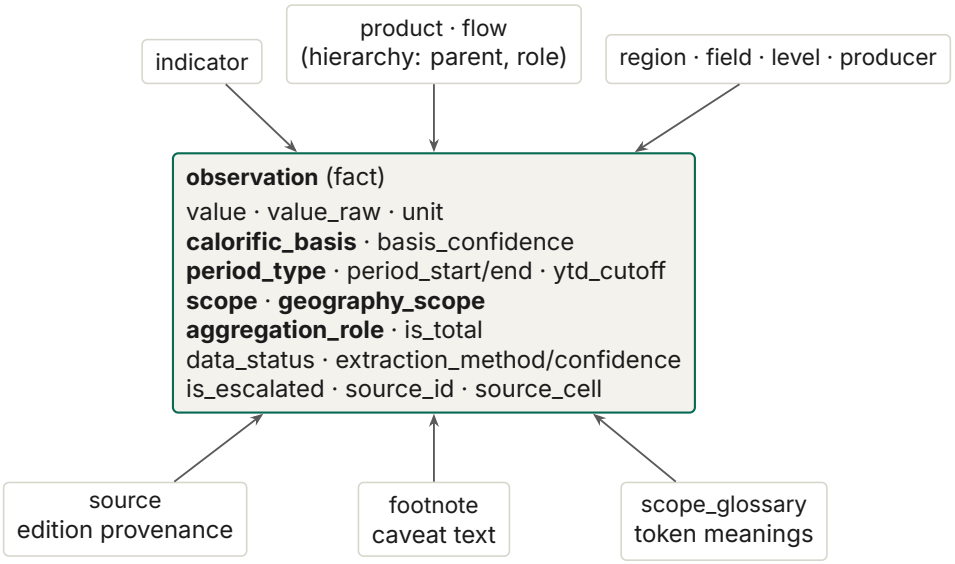

Figure 2. The qualifier-preserving star schema. Conformed dimensions surround a single fact table, but the distinctions that make a number true — calorific basis, period type, scope, geography scope, aggregation role, data status — are columns *on the fact*, not properties to be re-derived. A glossary defines every qualifier token; footnotes attach the source's own caveats to the cells they qualify.

3.3 Aggregation roles and the double-count guard

Total-ness is a property of a *cell*, not of a dimension: the same product line is a total in a "DEMAND" row and a leaf in a sub-row. Each observation therefore carries an explicit `aggregation_role` — one of `leaf`, `grand_total`, `subtotal`, or `alternative_breakdown`. To total a partition one reads the `grand_total` or sums the `leaf`s — never both, and never two `alternative_breakdown` partitions of the same total (the gas-demand-by-usage versus by-pressure trap). This single classifier is what lets the serving layer and the validation battery both reason about aggregation without re-deriving it per query.

3.4 Cross-edition reconciliation, retained not erased

Because editions overlap, the same cell is often reported more than once, sometimes with different values (a provisional figure later revised, or two editions disagreeing). The build records **every** multi-source cell in a reconciliation log — **6,308** such (series, year,

period type, basis) cells — and flags the **787** that disagree beyond a rounding tolerance. Disagreements are **retained, never overwritten**: a documented precedence rule (Bilan > Memento > Conjoncture; final > provisional; later publication date wins) selects which value the clean view serves, while the alternatives stay queryable. The principle is that a store for analytical use should expose disagreement, not silently launder it into a single number.

3.5 An idempotent, gated build

The loaders are idempotent: an observation's identity is a deterministic upsert key over its indicator, dimensions, period, basis, and source version, so re-running the build adds only new or changed rows. The build is not considered done until the validation battery (Section 6) passes; the battery is a permanent gate, re-run after every ingest, not a one-time acceptance test. The shipped store holds **26,899 observations**; the precedence-winning, trustworthy default surface (the clean view) exposes **11,173**.

4. The MCP Server: Readability and Refusal

4.1 The tool suite

The server is built on the official MCP SDK's `FastMCP` framework and exposes **twelve tools**, organized as a discovery-then-retrieval path with explicit guardrails. Table 1 summarizes them.

Tool	Role	Purpose
`search_series`	discovery	Ranked semantic search over the catalogue; FR/EN label resolution, twin and out-of-scope signals.
`list_series`	discovery	List series, optionally by indicator; a deferred family returns an out-of-scope response.
`get_series`	retrieval	The full series for an id, every point with its qualifier envelope; clean by default, opt-in for provisional/low-confidence/escalated.
`get_observation`	retrieval	A single point (series + year [+ cutoff]), fully qualified.
`describe_series`	metadata	Definition, unit, basis, period type, scope, aggregation role, escalation, and the source's own footnotes.
`get_metadata`	metadata	Alias of `describe_series`.
`compare`	guardrail	Guarded comparison; *refuses* incompatible basis / period / scope / unit / aggregation role.
`get_conflicts`	provenance	Cross-edition disagreements for a series: who disagrees, the precedence winner, the retained alternatives.
`convert_units`	reference	Documented conversions only; PCI↔PCS treated as a flagged basis change.
`list_units`	reference	Units and available conversion factors.
`list_dimensions`	reference	Dimension vocabularies with FR/EN labels.
`get_scope_glossary`	reference	Definitions of qualifier tokens and the "never sum/equate across" rules.

Table 1. The twelve tools. Discovery resolves a series id; retrieval returns fully qualified values; the guardrail, provenance, and reference tools keep the distinctions intact and the caveats reachable.

4.2 The governing principle: make a category error hard to express

Every upstream convention exists so this layer cannot launder a qualifier-stripped value into a confident wrong answer. The serving layer is therefore designed not merely to *permit* a correct answer but to make an incorrect one **hard to even express**. Three mechanisms carry that principle (Figure 3).

No bare numbers. Every value is returned inside a qualifier envelope — basis, period type and cutoff, scope, geography scope, aggregation role, data status, and full provenance — alongside the relevant glossary entries. A model cannot obtain a figure without the information needed to state it correctly.

Refusal across incompatible qualifiers. The `compare` tool refuses to line up series that differ on calorific basis, period type, geography scope, unit, or aggregation role — returning a structured refusal that names the incompatibility, rather than a silently wrong difference. Comparing a PCI figure with a PCS one, an annual with a year-to-date, or a grand total with its own components is not a discouraged action; it is an *unavailable* one without an explicit, warned override.

Out-of-scope is not no-data. The server distinguishes three outcomes that a naïve layer would collapse into one: a **value**; an **empty-in-scope** result (the series exists but no figure matched the filters); and **out-of-scope** (the family is deferred / not ingested). A query for prices or transit volumes returns an explicit out-of-scope signal that cites the coverage documentation — never an empty result that reads as "the data does not exist."

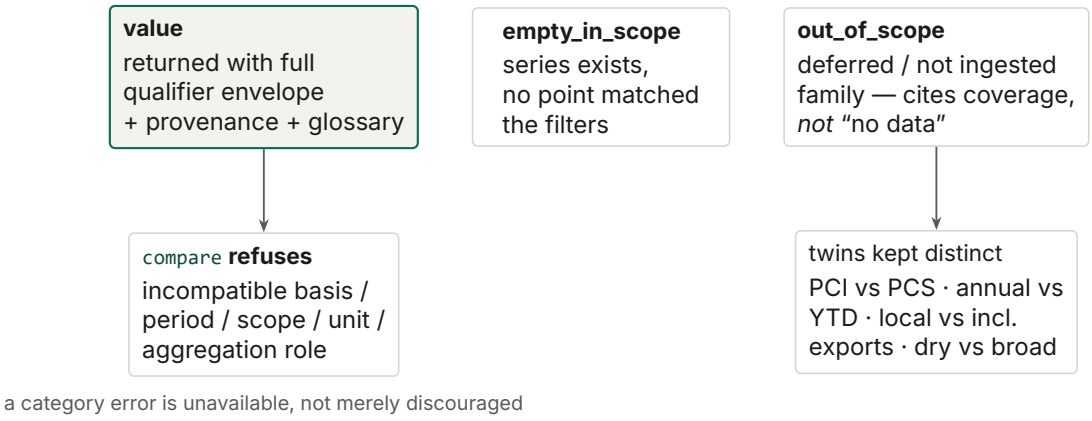

Figure 3. The qualifier guard. Three distinct outcomes (value, empty-in-scope, out-of-scope) are never collapsed; `compare` converts the common category errors from "possible to avoid" into "impossible to express without an explicit override"; and twin distinctions are preserved end-to-end.

4.3 Read-only posture and clean-view serving

The server opens the store **read-only** and serves only from the precedence-winning clean views, never the raw observation tables. The read-only handle is not cosmetic: a stale writer left holding the file mid-write had caused an incident during development, and the read-only contract guarantees the server can never be that writer. Low-confidence and escalated observations are excluded from the default surface and reachable only through explicit opt-in parameters, each returned with a loud warning flag.

4.4 Surfacing the source's own caveats

Some answers turn entirely on a footnote — for instance, an apparent contradiction between rising gas purchases and a falling royalty is explained by a documented over-withdrawal "in the course of regularization," a caveat ONEM prints and a careful analyst would need. `describe_series` surfaces these footnotes, resolved to their full text, attached to the very cells they qualify. A standing audit ensures caveats that belong to live series are actually linked to observations rather than stranded unreachably in the footnote table.

5. Design Decisions and Rationale

5.1 Qualifiers are columns, not afterthoughts

The most consequential decision is to carry every qualifier as a first-class column on the fact, from extraction to serving. The alternative — store the number, reattach basis

and period and scope later from context — is exactly how a pipeline produces confident wrong answers at scale. Making the qualifiers structural means the double-count guard, the comparison refusal, and the twin distinctions are all enforced against stored fields, not reconstructed heuristically per query.

5.2 Ship the built store *and* the pipeline

The repository ships the built DuckDB store so the server works immediately, and the full acquisition-and-build pipeline so the store can be regenerated from the ONEM PDFs. Shipping only the pipeline would make every user re-run a multi-edition extraction; shipping only the store would make the figures unauditable. Shipping both serves the immediate user and the skeptical one, and treats the derived store honestly as a reproducible artifact rather than an opaque drop.

5.3 Retain conflicts rather than resolve them away

A single "correct" number per cell would be easier to serve and less honest. The store retains every cross-edition disagreement and exposes it through `get_conflicts`, with precedence deciding only which value is served by default. For statistics that are revised edition to edition, the disagreement is itself information.

5.4 Out-of-scope as an explicit, first-class signal

Treating "not ingested" as distinct from "does not exist" is an anti-misinformation decision, not a UX nicety. A deferred family that returns empty would let a model state, in good faith, that Tunisia has no such data — a false and consequential claim. The explicit out-of-scope signal, citing the coverage documentation, closes that path.

5.5 Keyword and label resolution, not embeddings

Discovery is keyword and label based, with accent- and case-folding and a small French/Arabic-dialect synonym layer for high-value terms, rather than a vector index. The folding and synonyms handle the common real failures (accented spellings, a dialect term for a royalty) without the opacity and maintenance surface of an embedding layer that is not yet justified at this scale. Where a plain-keyword query under-matches, the relevance signal is reported rather than papered over.

5.6 A conservative redistribution posture

The system attributes ONEM as the data source, states plainly that it is an independent and unofficial project, and ships a downloader that fetches the PDFs from ONEM rather than re-hosting the government's documents wholesale. The built store is distributed as a derived artifact under that attribution. This is the conservative reading of reuse, chosen deliberately pending any clarification of ONEM's terms.

6. Cross-Validation: Arithmetic and Relationship Checks

The store is gated by a battery of **fifteen checks** that validate not just that the data parsed, but that it satisfies the **physical and structural relationships** the domain requires. They are the system's quantitative conscience, and they are deliberately **layered and scale-aware**: a figure that is contaminated can still sum correctly, so a paired-series check and a row-wise check catch different failure modes, and tolerance bands widen for small values where rounding dominates. Table 2 lists the battery; all hard checks pass on the shipped store, with two checks reporting advisory reconciliation detail.

Check	Relationship asserted	Trap it closes
C1	Energy-balance identity (production + imports − exports ± stock ≈ gross inland)	A mis-extracted balance line that breaks the accounting identity.
C2	Gas PCI $\approx 0.9 \times$ PCS on paired series	PCI/PCS confusion at the series level.
C2b	*Row-wise* PCI $\approx 0.9 \times$ PCS for the same field+period	Per-row basis contamination (a PCS value stored under PCI) that C2 and rollups miss.
C3	Components ≈ total where a total exists	Detail/total double-count.
C5	December YTD ≈ matching annual *Réalisé*	A year-end cumulation diverging from the annual figure it should equal.
C6	A series_key never mixes period_type	Annual and YTD merged into one series.
C6b	A series_key never mixes unit or basis	PCI and PCS (or different units) merged into one series.
C6c	YTD and annual never share a series_key	The period-type twin collapsing.
C4	Cross-edition values logged and reconciled	A silent overwrite hiding edition disagreement.
C7	No observations from non-canonical (AR) sources	Double-counting a translated edition.
C8	Realized coverage matches the known holes	An unrecorded coverage gap.
C9	Every observation carries unit/period/source; gas carries a basis	A bare, unprovenanced, or basis-less figure.
C10	Leaf rows reconcile to the pinned grand total	Incomplete partitions; non-reconciling editions are excluded as low-confidence.
C11	Leaves $\leq 1.15\times$ grand total	Overlapping partitions summed together.
C12	Exactly one grand total per group	Leaves with no total, or a subtotal masquerading as the total.

Table 2. The validation battery. Each check encodes a physical law (C1, C2/C2b), an aggregation relationship (C3, C10–C12), or an anti-trap structural invariant (C5–C9). The checks are layered — the row-wise C2b exists precisely because a contaminated row can still satisfy the series-level C2 and the rollup C10.

A worked catch. The value of layering is concrete. An early build stored a field's PCS

gas value (327) in its PCI column. The series-level C2 passed — the aggregate ratios still looked right — and the rollup C10 passed — the contaminated row still summed to its total. Only the **row-wise C2b**, asserting PCI $\approx 0.9 \times$ PCS for the *same field and period*, exposed the contamination, which traced to a block boundary read at the wrong vertical position. The lesson — that a single aggregate check is not enough, because errors that cancel in aggregate survive it — is built into the battery's structure.

7. Evaluation

Beyond the build-time battery, the served *behavior* is evaluated by a three-layer harness that scores by **failure mode** rather than a single accuracy number. The deterministic layers are the headline result; the model-in-the-loop layer is a richer, stochastic layer on top.

- **Layer 1 — retrieval fidelity (deterministic).** For each golden series, the value and every qualifier must survive the tool round-trip intact against direct database ground truth. **105 of 105 checks pass.**
- **Layer 3 — adversarial guards (deterministic gate).** Twin-conflation and double-count attempts must be refused; out-of-scope must be honest; the qualifier envelope must be complete. **10 of 10 pass, zero gating failures.**
- **Layer 2 — behavioral (model-in-the-loop).** A real LLM, given only the tools, answers realistic questions; trajectories are scored per failure mode (PCI/PCS conflation, period mixing, scope confusion, double-count, no-data-vs-out-of-scope, provisional-as-fact, and others). This layer is the methodology by which served reasoning is probed; because it requires a live model backend it is stochastic and is reported as a method here, with the deterministic Layers 1 and 3 as the standing, credential-free gate.

The unit and acceptance suites that accompany the layers pass at **39/39** and **26/26** respectively. A worked **investigative specimen** — a multi-series question about the Tunisia–Algeria gas royalty, requiring the model to separate purchases from royalty, surface the regularization footnote, and refuse the out-of-scope transit figure — exercises the full surface end to end.

> **THE EVALUATION THAT CORRECTED THE SYSTEM**
>
> The evaluation is not ceremonial. The behavioral layer flagged a unit conversion that the deterministic tests had been *accepting*: the PCI→PCS direction was inverted, returning a value that was confidently wrong. The decisive evidence was arithmetic ground truth — for the same gas, PCS exceeds PCI, so a PCI→PCS conversion must *increase* the figure — which a stale expected value had encoded backwards. The fix corrected the stored factor in both directions and added a property-based check (PCI→PCS must increase, PCS→PCI must decrease, and the ratio must match the database's own same-gas magnitudes), so the regression cannot recur. The principle: validate against ground-truth relationships, never against a previously blessed number.

8. Quality Assurance: an Adversarial Multi-Agent Audit

The store and server were hardened through an **adversarial, multi-agent review** in which independent reviewers — structural, semantic, and behavioral — examined the system from different angles and were explicitly tasked to *distrust* claims of correctness, including "correct by construction" arguments. The discipline that emerged, and that the project now follows, has three rules: **surface, do not silently resolve** (log an ambiguity or conflict rather than guess); **verification beats assertion** (a double-count one reviewer had dismissed as impossible was found real by a second, independent pass); and **findings become rulings become canon** (a human ruling settles a contested modeling question, after which it is authoritative).

The audit's recurring lesson was a class of bug worth naming: **metadata that never reaches the served surface**. Escalation flags set in principle but not propagated to the data; a decisive footnote present in the corpus but linked to **zero** observations and so invisible to the tool that should surface it. Each was caught by an independent pass asking not "is the value right" but "does the meaning reach the consumer," and each is now closed and guarded against recurrence. The system carries **900 observations** explicitly flagged as escalated — uncertain items isolated and presented provisionally rather than quietly folded into their neighbors.

9. Limitations and Known Issues

This section is deliberately complete; for an archival document, candid limitations are a credibility feature.

Extraction is validated against the database, not re-adjudicated against the PDFs. The validation battery and evaluation establish internal consistency and relationship correctness; they do not re-prove every cell against the source PDF. Spot-checks against source were performed and were accurate, but a systematic cell-by-cell audit against the PDFs is future work, not a completed guarantee.

Deferred families are out of scope, by design and for now. Prices, trade values and volumes (including pipeline transit), refining and exploration KPIs, product imports, and capacity are defined but not ingested. The server reports them as out-of-scope; it cannot answer them. This is honest, but it is a real coverage boundary.

Deferred editions and low-confidence cells. Several older editions (parts of the Memento and Bilan series, pre-2019 Conjoncture) are not ingested or are ingested at low confidence and excluded from the default surface; their figures may still arrive indirectly through later editions' historical columns. The reconciliation log records **787** cross-edition disagreements that are retained and served by precedence but not independently re-adjudicated against source.

Escalated, unresolved items. A small number of modeling questions only ONEM can settle are flagged escalated and isolated — notably the calorific basis and scope of the Bilan's broad gas aggregate, and a contested field identity (Barka versus Maâmoura-Baraka). These are presented provisionally, never reconciled into other series, pending confirma-

tion.

Discovery is keyword/label based. Search is accent- and case-folded with a small dialect synonym layer, but it is not embedding- or fully synonym-aware; a user's term may not match ONEM's. The relevance signal mitigates this but does not remove it.

The behavioral evaluation layer is stochastic and backend-dependent. Layers 1 and 3 are deterministic and form the standing gate; Layer 2 requires a live model and is reported as methodology. A single model run is not a guarantee of behavior across models or runs.

Local, source-distributed, single-machine. The system serves a technically capable user who can clone a repository and configure a client. It removes the need to read PDFs; it does not yet remove the need to run a local server.

10. Future Work

Concrete next steps, in rough priority order:

- **Ground-truth the retained collisions and low-confidence cells** against the source PDFs as those editions are recovered, converting "retained by precedence" into "re-adjudicated against source."
- **Ingest deferred families** — prices, trade values and volumes, refining and exploration — extending coverage from the physical-quantity core toward the economic layer, with the same qualifier discipline.
- **Recover deferred editions** — the rotated and drifted Memento and Bilan layouts, and the narrative-template Conjoncture editions — behind per-edition calibration.
- **Lower the barrier to running the server** — a packaged install or launcher to extend reach beyond the technical user.
- **Broaden the tested surface** — more end-to-end behavioral coverage across models, and a systematic source-vs-store extraction audit.

11. Conclusion

onem-tunisia-mcp turns a stream of ONEM PDF reports into something an LLM can query in natural language without losing what makes the numbers true. Its value is not the protocol it speaks or the dataset it wraps, but the discipline between them: coordinate-aware extraction with cell-level provenance, a schema in which every qualifier is a first-class column, cross-edition disagreement retained rather than erased, a serving layer that makes a category error hard to even express, and — standing behind the store — a layered battery of arithmetic and relationship checks, a three-layer evaluation, and an adversarial multi-agent audit. The system as documented here is a local, source-distributed stdio server bundled with its built store; its limitations are real and stated plainly, and its guarantees are the kind that hold up under cross-examination.

About the Author

Tarek Gasmi

FOUNDER, TANITDATA

Tarek Gasmi is Founder of tanitdata and Head of Data and AI at datadoit, with an academic affiliation at the University of Manouba. His work focuses on AI governance, digital infrastructure, sustainability, and technology policy, with particular attention to transitional economies and the geopolitical economy of AI systems.

Citation and License

Suggested citation. Gasmi, T. (2026). *onem-tunisia-mcp: A Qualifier-Aware MCP Server over Tunisia's National Energy Time-Series.* Technical report TD-TR-2026-03. tanitdata, Tunis.

Version. v1.0.0, June 2026.

References

Model Context Protocol — open standard specification and SDKs. modelcontextprotocol.io.

Observatoire National de l'Energie et des Mines (ONEM), Ministère de l'Industrie, des Mines et de l'Énergie, Tunisie. energiemines.gov.tn.

onem-tunisia-mcp source repository, github.com/tanitdata/onem-tunisia-mcp (source tree, `README`, `LICENSE`, the build pipeline, and the evaluation suite).