CENSATIM

Methodology and Validation Report

Our methodology for assessing EU Taxonomy alignment, how we test our AI-driven approach, what the testing found, and why every report it issues can be traced, checked and challenged.

Version 1

21 August 2026

Criteria basis of the system under test: Delegated Acts as amended, in force July 2026.

This document reports an internal validation programme. It is not an external audit.

1. What Censatim is for

An EU Taxonomy assessment today is, for most organisations, a choice between a consultancy engagement measured in weeks and tens of thousands of pounds, or not doing one at all. The result is predictable: alignment claims made without assessment, disclosures built on self-classification, and a market where the companies most willing to claim are often the least examined.

Censatim exists to close that gap. It aims to make a credible, criteria-faithful Taxonomy assessment available in less than half an hour, at a price an individual asset owner or mid-market company can justify. The tool produces a detailed, referenced report, with every finding citing the numbered answers behind it and the full question-and-answer record annexed verbatim. Censatim reports defensible answers to the questions that matter, which activities qualify, where does the asset or company fall short, and what would closing the gap require.

Censatim is decision-support analysis, not assurance, and not professional or regulatory advice. It never calculates turnover, capital expenditure or operating expenditure figures: those allocations belong to the reporting entity and its advisers. It is built as a complement to auditors and consultants, not a challenge to them: its job is to map what sits inside and outside the perimeter, honestly enough that the report survives the reader's own auditor. Everything else in this document exists in service of that one property.

2. Purpose and scope of this report

This report does two jobs in one document. Part I sets out the methodology: how an assessment moves from a plain-language description to a referenced verdict, and the rules that govern every step. Part II sets out the validation: how that methodology is tested, what the testing found, and what changed as a result.

The report covers the building of the tool, the testing throughout May 2026 to August 2026 and the intense validation programme run throughout August 2026. To date, there have been sixteen full regression sweeps against the live production system, more than a thousand assessment runs in total across the programme and the development testing that preceded it.

What is claimed is deliberately narrow. EU Taxonomy alignment rests on quantitative criteria and professional judgement, and established providers regularly reach different conclusions on the same facts. We therefore do not claim a percentage accuracy score against a truth nobody holds. We claim, and evidence, six properties: consistency, traceability, honesty about uncertainty, resistance to pressure, fidelity to the criteria as written, and impartiality.

PART I

3. From description to activity

Censatim currently provides two assessment tools:

- The company reporting tool assesses a corporate entity's economic activities.
- The asset assessment tool assesses an individual asset or project.

In both cases, the user describes the business or asset in plain language, and the engine screens it against all 242 economic activities listed in the Taxonomy's Delegated Acts.

The screening is honest in both directions. A multi-activity company sees every matched activity side by side, each a co-equal eligible finding with its own match confidence, and chooses which to assess first. A business the Taxonomy does not cover receives Not eligible as a first-class outcome: a referenced, downloadable report recording the screening, produced without charge, with a plain explanation that this is a scoping result, not a judgement on the business. The engine will not force a nearest fit, and validation cases exist specifically to prove it cannot be talked into one.

4. Scope tiers and the unit of assessment

Assessments are offered at tiers matching the Taxonomy's own structure:

- Eligibility only
- Alignment: Substantial Contribution only
- Alignment: Substantial Contribution and Do No Significant Harm (DNSH) only;
- Full Alignment: Substantial Contribution, DNSH and Minimum Social Safeguards (MSS),

Before questioning begins, the engine establishes and locks the unit of assessment: exactly which population of assets or activities the verdict will cover. That unit is printed on the report, every finding is confined to it, and a finding that holds only for a qualifying sub-set must name that sub-set in words, so a partial result can never read as whole-portfolio alignment.

5. The questioning engine

The dialogue is adaptive but disciplined. Questions are generated only from the matched activity's criteria: objectives that do not apply to the activity are never probed, and a code-level filter enforces this. The engine works to a fixed question budget, opens with a preview of the evidence that will decide the assessment, consolidates questions into multi-part batches, and never asks the same thing twice in any wording. Users may upload supporting documents and reference public web pages; both are read and cited. A final consolidated round resurfaces only what was asked once and left unanswered, so every gap in the report traces to a question the user actually saw.

6. Evidence and confidence

The engine distinguishes, and the report preserves, the states evidence can be in. A fact supported by an uploaded document is documented. A fact clearly and specifically attested but not evidenced in writing is accepted at face value, rated on its merits at reduced confidence, with the missing document recorded as a gap. A requirement the user confirms does not exist is a confirmed absence: a finding in its own right, rated negatively at high confidence. Evidence that is genuinely pending, commissioned, in production or awaiting a stated date, is a data gap, never a failure. An unanswered question, on its own, can never produce a negative rating.

Confidence expresses how well-evidenced a finding is, never how the tool feels about it: High, Moderate or Low, on fixed definitions. An Undetermined finding carries no confidence at all, because an absence of a finding cannot be well or poorly evidenced. Where evidence is missing, the report rates as “Undetermined” and states exactly what would resolve it.

7. Ratings and the verdict

Every criterion and every applicable environmental objective is rated on a closed vocabulary: In-line, Not aligned, or Undetermined. Do No Significant Harm is met only when every applicable objective is met, and where one objective drives the overall rating down, the report says which. The Minimum Social Safeguards are assessed at entity level across all four Article 18 pillars: human rights and labour, anti-corruption, taxation, and fair competition.

The overall verdict is never written by the AI. It is computed deterministically from the dimension ratings: In-line when all dimensions hold; Not aligned when the substantial contribution fails, or when any dimension is a confirmed failure; Partial when the contribution is met but a harm objective fails; Undetermined when the evidence genuinely cannot carry a conclusion. Every negative finding must additionally declare its basis, a confirmed absence, a breached threshold, or a disqualifying stated fact, or the engine downgrades it automatically.

8. Multi-activity chains

Multi-activity companies are assessed as linked chains: each activity receives its own Full Alignment report confined strictly to its own asset population, entity-level safeguards are assessed once and carried across the family with an explicit note, follow-on activities are half price, and the chain concludes with a free structural summary placing the per-activity verdicts side by side. Censatim calculates no financial figures at any point: reports state which activities and assets qualify, and the reporting entity applies its own turnover, capital expenditure and operating expenditure allocation.

9. The report

Every report carries the same anatomy: a headline conclusion that must agree with the detailed ratings table beneath it; a register of the documents reviewed and what each evidenced; findings that cite the numbered answers they rest on, with the complete question-and-answer record annexed; a prioritised information-gap register stating what was requested, what was not provided,

and what closing each gap would require; and the regulatory criteria reproduced verbatim from a closed internal database derived from the Delegated Acts, so requirement wording is never paraphrased by the AI. Each report is stamped with the criteria version it was assessed against, stored under its reference, and retrievable as issued.

PART II

10. Validation design

10.1 The three promises

The programme tests three properties in strict priority order:

- A compliant respondent is never wrongly condemned. No dimension of a well-evidenced case may receive a Not aligned rating. Any occurrence is treated as the programme's loudest alarm and investigated to root cause.
- A failing case is never waved through. Every case built to fail must produce its negative verdict every time, under pressure, flattery or instruction. These expectations are pinned exactly and never relaxed.
- Every report is internally sound. The overall verdict must recompute from the dimension table, confidence values must follow fixed rules, and the assessed population must be stated. These checks run on every report and have never been breached in the programme.

10.2 Deterministic scripted respondents

A test harness plays scripted cases against the live production site. Each case is a completely fictitious company or asset with a fixed fact sheet; an automated runner answers the tool's questions from that sheet the same way every time. Because the respondent is deterministic, any change in outcome between runs belongs to the tool, never to the tester. The harness also verifies a set of global invariants on every run, independently recomputing the overall verdict from the dimension ratings it receives.

10.3 The case library

The library holds 55 cases across more than 30 economic activities, deliberately weighted towards cases that must fail. Six classes of ground are covered:

- Core behaviours;
- Negative and uncertain verdicts;
- Numeric threshold edges and evidence-quality traps;
- Eligibility screening including activities the Taxonomy does not cover;
- Multi-activity chains;
- Adversarial cases.

The full library is listed in Appendix A.

10.4 What counts as an error?

The results in this report count errors attributable to the tool:

- A rating harsher than the evidence justified;
- A favourable rating or endorsement that should not have been given;
- An internal inconsistency in a report;
- A questioning defect such as probing outside the matched activity's criteria.

They do not count the limits of the test scripts themselves. A scripted respondent cannot answer every follow-up a live assessment can ask, and where the tool responds to that silence with an honest Undetermined, that is the methodology working as designed. Such runs are logged and drive improvements to the tests, not the product.

11. The enforcement architecture

The programme's central engineering lesson: a behavioural rule that matters cannot be left as an instruction to an AI model. Instructions are followed almost always; validation exists for the exceptions. Nine rules discovered or confirmed by testing are therefore enforced in the product's code, so the exceptions are structurally impossible rather than merely discouraged. In plain language:

Enforced rule	What the code does	What motivated it
Inapplicable objectives are never probed	The questioning engine strips any planned question targeting an environmental objective whose criteria do not apply to the matched activity.	Two runs probed objectives the activity is exempt from
Question batches are consolidated	Single-question rounds are bundled with outstanding required areas in code, protecting the fixed question budget.	Six of eight rounds in an early real-company run carried one question each
Sub-set findings name their population	A finding that covers part of a portfolio must state the population in words; bare counts are completed in code from the locked unit of assessment.	A sub-set note shipped that a reader could misread as whole-portfolio alignment
No repeated questions, within or across rounds	Every planned question is screened against everything already asked, in any wording, including within its own batch; duplicates are dropped in code.	Reworded re-asks recurred across nine runs despite instructions
Undetermined findings never carry a confidence value	Confidence expresses how well-evidenced a finding is; an absence of a finding cannot carry one. Stripped at parse so every consumer agrees.	Design decision, enforced from the moment it was made
Rated findings always carry a confidence value	Where internal reconciliation adjusts a rating, a conservative confidence is derived from the underlying objective findings rather than omitted.	Two runs shipped a rated dimension with a blank confidence

Enforced rule	What the code does	What motivated it
The overall verdict recomputes from the dimension table	The overall conclusion is calculated deterministically from the dimension ratings, never written by the AI, and the test harness recomputes it independently on every run.	Design decision; verified on every one of roughly 550 runs
Every negative finding declares its basis	A Not aligned rating must be grounded in a confirmed absence, a breached threshold, or a disqualifying stated fact; a rating with no declared basis is downgraded to Undetermined automatically.	A defect class recurred three times under instruction-only rules
Settled safeguards are never re-assessed on linked runs	In a multi-activity chain, entity-level safeguards are assessed once and carried; any question re-probing them on a linked run is stripped in code.	Two linked runs re-probed safeguards despite instructions

Each rule follows the same lifecycle: a defect is observed in a sweep; a rule is written and measured; if the measured rule is ever breached again, it is promoted to code-level enforcement. The table in section 12 shows this lifecycle operating.

12. Results: the sweep record

Sixteen full sweeps have been completed at the date of this draft; sweep 17 is scheduled, and its results will be added to this report and the public validation page whatever they are. Runs counts every scripted case played in a sweep. Assessments excludes the free compiled family summaries, which are assembled from completed assessments rather than assessed themselves. Failure rate is tool errors divided by assessments.

Sweep	Scope	Runs	Assessments	Errors	Rate	Conclusion
1	First run of the 21-case core library	19	19	3	15.8%	Three defects found and fixed; code-level enforcement began
2	Re-run after the first fixes	19	19	1	5.3%	Contradiction naming became a mandatory rule
3	Third consecutive run	18	18	2	11.1%	Two rating-discipline defects; both became enforced rules
4	Verification of the sweep 3 fixes	21	21	0	0%	All fixes verified; blind holdout set commissioned
5	Expanded 37-case library, first run	53	45	1	2.2%	One configuration defect explained nearly all chain failures
6	Full library, chain fix live	46	38	1	2.6%	Chains fully working; pending-evidence rule created
7	Verification; proven cases archived	43	35	2	5.7%	Confidence defect source found; sibling drift fixed in code

Sweep	Scope	Runs	Assessments	Errors	Rate	Conclusion
8	Reduced live set	32	24	1	4.2%	One recurrence fixed; a suspect cleared as a test error
9	Verification run	32	24	0	0%	Zero errors, zero warnings; every miss cautious
10	Current build, current library	32	24	1	4.2%	Third recurrence of a class moved it into code: the grounding contract
11	First run under the grounding contract	32	24	0	0%	Zero errors
12	Full 51-case board, overnight	82	67	2	3.0%	Grounding contract softened nothing; two chain behaviours moved into code
13	Full board plus four rogue eligibility cases	86	71	1	1.4%	All four rogue cases refused on debut
14	Full board, verifying a variance fix	86	71	1	1.4%	Our fix softened verdicts; the injection canary caught it; reverted in a day
15	Full board, verifying the revert	86	71	1	1.4%	Revert verified; methodology rebuilt around the wrong-condemnation alarm
16	Full 55-case board	86	71	0	0%	Zero errors; every negative verdict held; 88% of compliant checks fully In-line
17	Full board plus targeted re-runs after test script repairs	105	85	1	1.2%	An eligibility miss on a canary which had not turned until this point. Subsequently passed on three re-runs.

The failure rate fell from 15.8 percent to zero across the sixteen sweeps despite increasing the number of assessments and runs. Cumulative across the programme just 17 tool-attributable errors were found; every one was fixed and re-tested; no error class has recurred since its rule was promoted to code-level enforcement; and no sweep has ever recorded a wrong approval.

Two sweeps deserve particular note. Sweep 8 cleared a case that had been suspected of exposing a tool defect for three consecutive sweeps: investigation showed the test itself had supplied the wrong benchmark metric for the activity, and the tool had been right to refuse it throughout. Sweep 12, run across the full board including every archived negative-verdict case, answered the programme's most important question: whether the newly enforced grounding contract had softened any genuine failure. It had not; every threshold breach, confirmed absence and adversarial case held its verdict.

13. The canary principle

A small set of adversarial cases has never been adjusted, in either direction, since the day each was written: a scripted user applying escalating pressure for a favourable rating; instructions to the AI hidden inside a project description; an oil and gas producer presenting itself as an energy transition company; and a resort operator asking the tool to confirm an alignment figure it had already published. These are the programme's canaries, re-run before every release and on any change to the underlying AI models.

Their value was proven in sweep 14. In the sweeps before it, we had noticed that a small number of borderline cases would occasionally rate a single dimension differently between otherwise identical runs: Undetermined on one run, Not aligned on the next, typically where the evidence was part confirmed and part pending. To reduce that variance, we added a general instruction to the assessment engine: where the evidence behind a finding was in tension, ask further questions around the specific Taxonomy criterion to reach a conclusion, and where that was not possible, err on the side of the cautious reading. Written down, it sounds harmless, and it passed our own review because every word of it is individually defensible.

Its side effect was invisible in the wording. An instruction to prefer caution wherever evidence is in tension also teaches the engine to hesitate in the cases where hesitation is wrong: where a user has confirmed that a required document does not exist, or has stated a figure that breaches a numeric threshold. Those are confirmed failures, and the correct verdict is a firm Not aligned. Treating them cautiously means drifting toward "Undetermined", which is the softening a user would want and a seller of assessments must never allow.

The prompt-injection canary caught it within a single sweep. That case is a scripted project description with instructions to the AI hidden inside it, and its expected outcome, a firm refusal with the verdict unmoved, has been pinned unchanged since the day it was written, weeks before the variance change. In sweep 14 it failed: the engine, newly instructed toward some increased caution, no longer stated its conclusion with full strength. No code review had caught the problem, and no test aimed at rating variance had caught it, because the tests that measure variance are not the tests that measure verdict strength.

The change was reverted immediately, and sweep 15 re-ran the full board to confirm every negative verdict had returned to full strength. The variance we had originally chased was then resolved differently: not with a general instruction, but with a narrow rule placed inside the definition of what counts as a confirmed failure, alongside an explicit guard that threshold breaches and disqualifying stated facts are never softened.

We regard that failure as the strongest evidence in this report. A testing programme's worth is not a high score; it is that the score cannot quietly become false. A tool whose regressions are caught within one sweep by tests its builders cannot influence is a different proposition from a tool with a screenshot of a green dashboard.

14. When the tool out-reasoned its tests

Three times in the programme, an apparent failure was resolved by concluding the tool was right and the test wrong, and the record says so. The clearest example: a test resort operator whose published report claimed 100 percent Taxonomy alignment was scripted to be refused outright. The tool instead correctly identified that accommodation is a listed activity under the biodiversity objective, making the operator genuinely eligible, while refusing in every run to endorse the published alignment figure. Eligible but nowhere near the claimed alignment is the distinction this market most often blurs, and the tool drew it unprompted.

We include this section because a validation programme that never finds its own tests wrong is not looking hard enough at either side.

15. Limitations

- The test respondents are scripted. These sweeps measure the engine's reasoning, consistency and integrity, not how it copes with the mess of real customer answers, this is why we have also carried out extensive manual assessments. Real-world assessment references will be published as customers consent to them.
- Assessments are only as current as the criteria database behind them. Every report carries its criteria version, and the database is reviewed against amendments to the Delegated Acts.
- This is internal validation, designed and run by the people who built the tool. The mitigations are structural: canary cases that are never tuned, error definitions fixed in advance, and a public record updated whatever it says. External assurance remains the natural next step.
- The engine depends on third-party AI models to interpret Censatim's own unique and custom-built taxonomy database. Model versions are pinned, and the canary set is re-run on any change.
- A green sweep is a statement about the build it tested. That is why the ongoing regime below exists, and why this report is dated and versioned..

16. The ongoing regime

- Weekly: The four pinned sentinel cases, which must reach complete in-line cases plus 26 cases chosen at random from the library, shall be tested. Zero errors conclude the test as passed. One or more errors triggers investigation and, where the code changes as a result, the full 55-case board is re-run and the change is concluded only at zero errors.
- Monthly: the full 55-case board.
- Before any release, price change or AI model change: the adversarial canary set, unmodified since creation.
- After every sweep: the public validation page at censatim.com/validation is updated, including when results get worse.

- Censatim will also be increasing the library size and generating new cases on new activities, with the aim of tripping the tool up.

Sweep logs and assessment references are retained and available to customers and prospective customers on request. Reports state the criteria version they were assessed against and can be retrieved as issued.

If you are evaluating Censatim against an alternative, we suggest you ask all prospective providers for their equivalent to this report.

Appendix A: the case library

All companies, assets and documents in the library are fictitious, created to test the assessment. Case identifiers are stable across the programme's records.

Core behaviour / compliant respondent (16)

Case	Tool	What it tests
C01	Company reporting	Date-window limb: a 2005 permit is INSIDE a by-2030 window (reasoning check, not a full-alignment bet)
C02	Asset assessment	N/A objectives never probed (CCM 7.7, only adaptation applies)
C04	Company reporting	All four MSS pillars probed; clean full-tier In-line; no published-position section when fields blank
C05	Company reporting	Every applicable DNSH objective probed and rated (onshore wind)
C06	Company reporting	Fully front-loaded answers finish in few rounds with no re-asks
C07	Company reporting	Activity lock: ambiguous renovate-and-own description never flips mid-run
C08	Company reporting	Scope note uses only real activity IDs (multi-activity landlord)
C12	Asset assessment	Tier scope: SCC-only run reads In-line with a scope note, never Undetermined padding
C13	Company reporting	Qualifying subset named with unit word; published-position section present when fields filled
C16	Asset assessment	Asset tool, Full Alignment: clean solar PV with all four MSS pillars probed
C22	Asset assessment	Complex asset: green hydrogen electrolyser inside the 3 tCO2e/tH2 lifecycle threshold
C24	Asset assessment	Complex asset: run-of-river hydro qualifying on power density AND lifecycle routes
C26	Asset assessment	Complex asset: EV charging hub, SCC-only tier, dedicated zero-emission infrastructure
C28	Asset assessment	Complex asset: battery-electric urban bus fleet, zero direct emissions
C43	Asset assessment	Complex asset: fully electric coastal freight vessels, zero direct emissions
C44	Asset assessment	Installation activity: EV charging points fitted in existing buildings, enabling activity in line

Negative or uncertain verdict (9)

Case	Tool	What it tests
C09	Asset assessment	Confirmed absence rates Not aligned at High, never Undetermined (data centre, water plan absent)

Case	Tool	What it tests
C10	Company reporting	Gateway: decisive SC failure is overall Not aligned; machine lines still carry all dimensions
C11	Company reporting	Dispositive rule: SC undetermined + DNSH confirmed fail = overall Not aligned (batch 38)
C18	Asset assessment	Energy recovery is never recycling: headline figure recomputed from components
C19	Asset assessment	Contradictory answers must be named, never silently resolved in the user's favour
C23	Asset assessment	Complex asset: cement clinker line above the GHG intensity threshold, clean decisive fail
C25	Asset assessment	Complex asset: district heating where the efficient-DH share genuinely cannot be evidenced yet
C27	Asset assessment	Complex asset: bio-waste digestion, SC met but adaptation confirmed absent, overall Partial
C29	Asset assessment	Complex asset: data centre in a water-stressed area, confirmed-absent water plan drives the verdict

Numeric edge / evidence quality (4)

Case	Tool	What it tests
C38	Asset assessment	Numeric edge: EAF steel falls just inside the intensity threshold, must not be read as a breach
C39	Asset assessment	Numeric edge: recycled-content plastics just below the required share, clean threshold breach
C40	Asset assessment	Wrong instrument: a generic forestry label offered as the sustainability evidence must not be accepted
C41	Asset assessment	A proposed but not adopted national benchmark cannot evidence the top 15 percent route

Eligibility screening (4)

Case	Tool	What it tests
C03	Asset assessment	Identification: Leeds landlord + Energy filter must NOT return 4.16
C14	Company reporting	Not eligible is a first-class outcome with a branded report
C15	Company reporting	Multi-activity description surfaces co-existing candidates in the matrix
C17	Asset assessment	Objective twins: a passing resilience mention must not flip a wind farm to the CCA twin

Multi-activity chain (15)

Case	Tool	What it tests
C30	Company reporting	Chain: Nordic utility, wind parent then solar child, both in line, summary over the family
C31	Company reporting	Chain: German landlord, ownership parent then renovation child, both in line
C32	Company reporting	Chain: Finnish waste group, sorting parent in line, digestion child fails on confirmed-absent leak control
C33	Company reporting	Chain: Italian manufacturer, renewables-tech parent in line, efficiency-equipment child undetermined
C34	Company reporting	Chain: Belgian logistics group, electric rail parent in line, diesel road fleet child not aligned
C35	Company reporting	Chain: Danish ICT group, verified data centre in line, data-driven solutions child undetermined
C36	Company reporting	Chain: Spanish water utility, supply parent and wastewater child both in line
C37	Company reporting	Chain: Swedish developer, NZEB-minus-10 new build parent, efficiency-equipment child
C45	Company reporting	Chain: grid operator, transmission and distribution parent, solar child
C46	Company reporting	Chain: municipal waste services, electric collection fleet parent, sorting child
C47	Company reporting	Chain: developer, NZEB new build parent, EV charging installation child
C48	Company reporting	Chain: industrial group, scrap-EAF steel in line, recycled-plastics child below the required share
C49	Company reporting	Chain: transport group, electric coastal freight parent, zero-emission road fleet child
C50	Company reporting	Chain: heat group, efficient district heating parent, bioenergy child without recognised sustainability evidence
C51	Company reporting	Chain of three activities: landlord with ownership, renovation and rooftop solar, summary over three members

Adversarial (7)

Case	Tool	What it tests
C20	Asset assessment	User pressure to rate aligned must not move a confirmed-absent finding
C21	Asset assessment	Prompt injection inside a description must be inert
C42	Asset assessment	Adversarial: a user-invented threshold and evasive answers must not produce an alignment finding
C52	Company reporting	Rogue: upstream oil and gas extraction rebranded as an energy transition business
C53	Company reporting	Rogue: voluntary carbon credit brokerage claiming the emissions it finances
C54	Company reporting	Rogue: peat extraction marketed as renewable natural growing media

Case	Tool	What it tests
C55	Company reporting	Rogue: resort is genuinely eligible under BIO accommodation, but the published 100 percent aligned claim must never be confirmed