

The Digital Presence of Tourist-Facing Businesses in Taranto: a Pre-Registered Open-Data Census Ahead of the XX Mediterranean Games

Preliminary report — v1.0, 2026-07-17. Transmitted for legal review; any update will be published as a new version of this deposit. An Italian companion report is deposited with this record and is the primary public deliverable (protocol §11.2). **Study protocol:** frozen v0.3 (2026-07-07), amendments v0.3.1–v0.3.8 logged with rationale (PROTOCOL.md, deposited with this record, §14). Snapshot date: **2026-07-16**. **Data controller:** Arrow Web Agency (Nicolas Moerynck) — privacy notice published before collection: <https://arrowweb.it/informativa-studio/>. Analysis pipeline authored with an AI coding assistant under a documented no-PII-access architecture (§2.7). **Competing interests:** the author operates a web agency in the market measured here. The protocol was frozen before collection to constrain that interest; the dataset is under a purpose lock forbidding any per-business use; no business is named in any output. The study was self-funded and commissioned by no one. **DOI:** 10.5281/zenodo.21834633 · **Licence:** CC BY 4.0

Abstract

We measure the state of the commercial web for tourist-facing businesses (food service, lodging, retail) in the comune of Taranto (Italy), two months before the city hosts the XX Mediterranean Games. Using a pre-registered protocol, we construct a census frame as the deduplicated union of three open data sources (Overture Maps places 2026-06-17.0, OpenStreetMap, Regione Puglia DMS registries), yielding 3,021 candidate business locations, of which 2,258 meet the pre-registered inclusion criteria. Each unit's digital presence was audited passively (fixed 3-request budget per site, robots.txt honored) plus Google PageSpeed Insights for eligible sites. **Findings:** 63.1% (95% CI 61.1–65.1) of in-frame businesses have no discoverable web presence of any kind; 24.0% (22.2–25.8) operate an owned website; 20.3% (17.7–23.2) of listed websites are dead. Chain businesses are far more likely to have a web presence than independents (71.5% vs 26.2%; OR 7.06, 95% CI 4.93–10.11), but 84.1% of chain "websites" are national corporate sites rather than local pages. Only 33.8% of owned sites offer any English content. Median mobile Lighthouse performance is 0.59 (median LCP 9.5 s). Cross-source analysis shows the three-source union was necessary: each source contributes a large exclusive share, and sources disagree on website presence for 49.5% of overlapping units. All estimates are descriptive; the design supports no causal claims.

1. Introduction and research questions

Taranto hosts the XX Mediterranean Games (21 Aug–3 Sep 2026; municipal planning anticipates 150,000–200,000 visitors). This creates a time-bounded influx of non-local visitors who must find and use local businesses without prior local knowledge. The study characterises the current digital presence of Taranto's tourist-facing businesses as a snapshot, independent of the event but timed to be informative ahead of it.

Pre-registered research questions (protocol §2, frozen 2026-07-07, before any data collection):

- **RQ1 (Distribution).** What is the distribution of digital-presence characteristics — existence, reachability, security, performance, mobile-readiness, findability, internationalisation, maintenance state — across Taranto's tourist-facing businesses?
- **RQ2 (Association).** Is business type (chain vs independent) associated with these characteristics? (Observational, not causal.)
- **RQ3 (Data integrity).** How reliable is the underlying frame data — coverage and agreement across sources — and how much measurement error does it introduce?

No hypothesis was favoured; RQ2 is reported whatever its direction.

2. Methods

2.1 Design

Single point-in-time census snapshot (frame frozen 2026-07-16); unit of analysis is the business **location**. The full protocol — including the analysis plan, classification rules, thresholds, and publication rules — was frozen before data collection; every subsequent amendment is dated and justified in the changelog (§14) and summarised in the deviations register (§5.3 below).

2.2 Frame construction

The frame is the deduplicated union of three lawful, storable open sources: **Overture Maps** places (pinned release 2026-06-17.0; bbox extract mechanically derived from the ISTAT comune polygon + 0.02°, containment asserted), **OpenStreetMap** (Overpass, relation 40452), and the **Regione Puglia DMS** open registries (strutture ricettive + locazioni turistiche, the authoritative lodging source). No Google service was used as a business-data source (contractual analysis in the decision memo of 2026-07-07). Categories follow a frozen three-class taxonomy (food service / lodging / tourist-facing retail) with per-source mappings reconciled against the pinned release's actual vocabulary; the tourist-facing retail scope (documented pre-freeze decision) includes alimentari, wine/beverage and tourist-adjacent shops and excludes chain-dense verticals such as electronics and cosmetics — a scope choice that shapes RQ2's composition and is stated as such.

Pipeline: 13,539 source records -> category mapping and boundary filtering (ISTAT non-generalized 2026 comune polygon) -> cross-source deduplication (token-set name similarity ≥ 0.82 within 75 m; 34 borderline pairs logged) -> **3,021 units**. DMS records without coordinates in the open data (634/861) were retained: 213 name-linked to a geolocated twin (≥ 0.93 token-set match, logged), 421 kept at comune level (zona = not_geolocated), rather than silently dropping most of the authoritative lodging registry. Overture-only units below the pre-registered confidence threshold (0.85) are flagged analytically (frame_included = false), not removed — RQ3 runs on all 3,021 candidates, RQ1/RQ2 on the **2,258 frame-included** units.

Sub-municipal geography (zona) uses a frozen list of 104 ISTAT 2021 census sections for the old-town island (città vecchia), the remainder of the Città Vecchia–Borgo circoscrizione (borgo), and the rest of the comune.

2.3 Classification

chain_flag was assigned by matching unit names against a version-frozen reference list of 101 brands (name/alias token-containment; the operationalisation is amendment v0.3.3); site_scope distinguishes location-specific sites from national corporate sites by registrable-domain suffix match against the brand list's corporate domains, finalised from the audited post-redirect URL.

2.4 Measurement instrument

Passive audit per distinct site within a fixed budget — 1× robots.txt, 1× homepage GET (redirects followed and recorded), 1× sitemap existence check — with identified User-Agent linking to the public study notice, jittered pacing, robots.txt disallow honored (12 sites), and raw HTML parsed and discarded at fetch time. Sites sharing one corporate domain were fetched once. Performance and mobile-readiness metrics come from the Google PageSpeed Insights API (mobile + desktop) for eligible sites only (owned/placeholder); 712/748 eligible units (95.2%) obtained PSI data; the remainder are 24 per-site Lighthouse failures (recorded as observations) and 12 robots-blocked sites. TLS characteristics were read from one handshake; certificate expiry is bucketed (exact dates are quasi-identifiers). Web status classes: owned, social_only (frozen social/aggregator domain list), placeholder (frozen heuristics), unreachable (listed URL, no response), none (no URL in any source — no discovery step exists by design).

2.5 Statistical analysis

Pre-registered plan (§10): descriptive distributions with Wilson 95% CIs; associations tested by Pearson χ^2 (Fisher exact where expected counts < 5), effect sizes as Cramér's V and, for 2×2, odds ratios with log-method 95% CIs; adjusted standardized residuals identify driving cells. Missing data are reported, never imputed. All statistics computed in pure Python with self-checked implementations.

2.6 Data protection

GDPR basis: legitimate interest with documented LIA, DPIA screening and ROPA; public privacy notice live before any pull; contact data stripped at pull; working stores pseudonymised (keyed HMAC unit IDs); names never entered working stores; the URL-bearing Identifier Store was deleted on 2026-07-17 after the agreement indicators were extracted (gated, logged); analysis uses exclusively the pseudonymous Metrics Store. Publication rule: minimum cell size 3 with complementary collapse.

2.7 AI assistance and the no-PII boundary

The pipeline and analysis code were authored by an AI coding assistant under an architectural constraint: all personal-data stores are encrypted to the controller's public key, the private key never left the controller, and every step touching personal data was executed by the controller. The assistant operated only on synthetic fixtures, public geodata, and pseudonymous/aggregate outputs. Code was hardened through multi-perspective adversarial review before each personal-data run (26 confirmed defects fixed across three reviews; all logged).

3. Results

3.1 RQ1 — Existence and ownership (n = 2,258 frame-included units)

web_status	n	%	95% CI
none (no URL in any source)	1,425	63.1	61.1–65.1
owned website	541	24.0	22.2–25.8
unreachable (listed URL dead)	169	7.5	—
social media only	63	2.8	—
placeholder / parked	60	2.7	—

Among the 833 units whose sources list any URL, **20.3% (17.7–23.2) point to a dead site**. Web presence (owned + social + placeholder) differs sharply **by category**: retail 43.7% (40.1–47.4), food service 25.1% (22.1–28.5), lodging 20.9% (18.3–23.8); $\chi^2(2) = 105.3$, $p < 0.001$, $V = 0.216$ — residuals show retail far above expectation (+10.1) and lodging far below (–6.8). **By zona** the differences are real but weak: resto del comune 38.8% (35.8–41.8), città vecchia 35.4% (26.6–45.2), borgo 30.5% (27.2–33.9); $\chi^2(2) = 12.7$, $p = 0.002$, $V = 0.083$. The digital divide in Taranto is categorical, not geographic. (421 non-geolocated units — overwhelmingly DMS short-term rentals — are excluded from zona comparisons and reported in §3.6.)

3.2 RQ1 — Security and reachability (owned sites, n = 541)

HTTPS 97.8%; valid certificate 95.2%; HSTS 33.5%; mixed content 12.8%. Security-header adoption is thin: 42.5% of owned sites send **zero** of the six scored headers; only 8.1% send ≥ 5 .

3.3 RQ1 — Performance, mobile, findability (owned sites)

Median mobile Lighthouse performance **0.59** (24.0% below 0.5); median mobile LCP **9,482 ms**; median desktop performance 0.79. CrUX field data exists for 56.9% of owned sites. Viewport meta is near-universal (94.1%). Findability basics: title 95.0%, meta description 63.2%, structured data 66.2%, sitemap 76.5%, robots.txt 86.0%, canonical 71.7%.

3.4 RQ1 — Internationalisation (owned sites)

33.8% offer any English content (homepage-detected); hreflang 22.4%; visible language switcher 37.9%. By category: lodging 38.7%, food service 32.5%, retail 31.7%. Two months before an international multi-sport event, roughly two of three business websites — and three of five lodging sites — are Italian-only.

3.5 RQ1 — Maintenance and stack (live sites, n = 601)

23.3% of live sites carry ≥ 1 staleness flag (coming-soon/error content 72; copyright year lagging ≥ 3 years 55 — 21.1% of the 261 sites disclosing a footer year; stale Last-Modified 11; clearly outdated CMS major 5; broken asset references 2). Identifiable stacks among owned sites: WordPress 203 (37.5% — 61% of identifiable), Shopify 33, Wix 10, Magento 10, others ≤ 5 each; 38.6% unidentifiable by passive inspection.

3.6 RQ2 — Chain vs independent

158 units (7.0%) classified as chain, 2,100 independent. Web presence: **chain 71.5% (64.0–78.0) vs independent 26.2% (24.4–28.2)**; OR **7.06 (4.93–10.11)**, $\chi^2(1) = 145.1$, $p < 0.001$, $V = 0.254$; full chain $\times$ web_status table $\chi^2(4) = 181.5$, $p < 0.001$, $V = 0.283$. However, site_scope reverses the reading of what that presence is: **84.1% of chain owned sites are the national corporate site** ($n = 107$), against 1.2% among independents ($n = 434$). The locally-authored web of Taranto is almost entirely the work of independent businesses — most of which have none. Caveats: observational; retail scope excludes chain-dense verticals (composition, §2.2); chain classification bounded by the frozen 101-brand list.

3.7 RQ3 — Frame integrity

Of 3,021 candidates: Overture contributes 2,107 units (1,628 exclusively), DMS 754 (567 exclusively), OSM 669 (341 exclusively); only 24 units appear in all three. Any single-source design would have lost most of the frame. On the 485 units carried by ≥ 2 sources: mapped-category disagreement 9.5%; **website-field presence disagreement 49.5%** — one source lists a site, the other does not, for half the overlap. Where two sources both list URLs ($n = 66$), all agree on the registrable domain in 66.7% of cases (Overture–DMS 74.1%, Overture–OSM 65.9%; OSM–DMS 25.0% on $n = 4$, not interpretable). Operating-status disagreement is **not measurable** in this frame (no source can emit "closed": Overture's field is blank in the pinned release for 12,001/12,005 rows; OSM disused elements never match the taxonomy query; DMS presence means "registered") — reported as such rather than as spurious perfect agreement.

3.8 Missing data (reported, not imputed)

Non-geolocated units 421 (18.6% of frame-included; zona analyses exclude them); PSI absent for 10 robots-blocked and 17 Lighthouse-failing frame-included sites; footer copyright year undisclosed by 340/601 live sites; CMS version disclosed by only 45 owned sites (version-based staleness is a floor, not a rate).

4. Discussion

Three findings carry the report. First, the **absence baseline**: 63% of tourist-facing businesses in the frame have no discoverable web presence at all, and a fifth of the URLs that do exist are dead — the modal Taranto business is invisible to a visitor's browser. Second, the **chain/independent asymmetry with the site_scope reversal**: chains look digitally present but their presence is corporate and non-local; independents own nearly all the locally-authored web, yet three quarters of them have nothing. Third, the **readiness gap on what exists**: slow mobile performance (median LCP 9.5 s on the median connection a visitor will use), two-thirds Italian-only, and a quarter of live sites showing abandonment signals. The RQ3 result stands on its own methodologically: open-data sources disagree about business websites half the time, so any single-source census of the "local web" — the industry default — measures its source as much as the territory.

5. Limitations

5.1 Design limitations (pre-registered, §12)

Frame = businesses discoverable in open geodata and official registries; **recall is not identifiable** from this design; web_status = none means "no site listed in any source", not "no site exists". Single snapshot; no longitudinal claims. Single city; no comparison baseline measured with the same instrument — cross-city statements are not supported. Indexation not measured. Internationalisation and staleness are homepage-only. Software versions from disclosure only. Observational associations throughout.

5.2 Frame-quality limitations

Frame precision was not independently validated: the pre-registered 150-unit manual ground-truth sample was waived (amendment v0.3.4); RQ3's cross-source agreement partially substitutes but measures consistency, not truth. The planned ISTAT ASIA coverage-ratio benchmark has not yet been computed. Some Overture-only units are likely stale or misplaced; the 0.85 confidence threshold is a pragmatic, pre-registered choice.

5.3 Deviations register (all logged in **PROTOCOL.md §14**)

1. Ground-truth validation waived (v0.3.4) — workload decision; stated throughout.
2. Language detection implemented as a stopword-set heuristic rather than a validated library (lingua/CLD3 were named as examples in the frozen spec); loanword false-positives mitigated (v0.3.5) but the instrument is unvalidated — i18n rates carry unquantified error.
3. Audit instrument defects found and repaired post-run with targeted re-runs (v0.3.6): a PSI back-off bug initially suppressed performance data (~480 sites; repaired to 95.2% coverage) and scheme-less URLs were initially misclassified unreachable (39 units re-audited).
4. Per-source attribute aggregation artifact affecting 6/3,021 units (v0.3.7; audit-derived `web_status` is authoritative, impact ≤ 6 units on one RQ3 indicator).
5. `zona = not_geolocated` class added for coordinate-less DMS records (v0.3.3 decision; alternative — geocoding — was rejected to avoid a new third-party disclosure).
6. Two chain classifications remain in the `fuzzy_reviewed` band pending hand review (logged; both currently counted independent).
7. Public output set extended (v0.3.8): this English companion report and the frozen protocol are deposited alongside the Italian public report under a single DOI; content rules unchanged (aggregate only, min cell size 3, complementary collapse), no figure changed.
8. §4A.7 review transmitted, not completed (v0.3.8): the draft report and the published notice went to external counsel on 2026-08-07 with a circumscribed question on the aggregate output's consistency with the published notice and legal basis; publication proceeded without waiting for a reply, rationale recorded in full in §14; any later response becomes a new deposited version, not a silent correction.

6. Reproducibility and data availability

Deterministic pipeline: ingest is a pure function of (sealed source archive, frozen config, study secret, adjudication log, exclusion list); the sealed archive (encrypted) retains the exact pulled snapshots; every artifact carries provenance (source, release, timestamp, SHA-256). The Metrics Store is pseudonymised (Recital-26 posture documented); the Identifier Store was deleted 2026-07-17 per the retention schedule. Aggregate outputs obey minimum cell size 3 with complementary collapse. Attributions: ISTAT (CC BY 4.0), © OpenStreetMap contributors (ODbL), Regione Puglia (CC BY 4.0), Overture Maps Foundation (per-source licenses).

An Italian-language companion report covering the same aggregate results is deposited with this record. The frozen protocol, including the amendment changelog (§14) and the deviations register, is deposited alongside so that every figure above can be traced to the pre-registered rule that produced it.