sOS Research Lab 2026-07-01 · Barcelona

WP14 – Barcelona Sandbox: Experimental Protocol

Series	sOS Execution
Status	Revised Draft – Branch A + Branch B (Anthropic Opus 4.8) revisions applied
Date	2026-06-28
Priority	Critical – Phase 1 (Phase 0 verification gate before full commitment)
Review Mode	Full Redundant Review (Branch A: Pi/DeepSeek → Branch B: Anthropic Opus 4.8 → Synthesis → Revision)
Grounded in	sOS doctoral thesis (specifically §14.3 six contract types, §14.4 validation by labels, §14.5 debate platform, §14.6 bureaucratic dynamics), WP6 (Barcelona Sandbox proposal), WP8 (Technical Architecture v2.0), WP9 (Legal compliance), WP10 (Tokenomics), Barcelona Open Data / Sentilo infrastructure, sOS doctoral thesis
H1 (Primary)	Stigmergic coordination through labeled traces and personal AI will achieve ≥15% improvement over the existing municipal mechanism on a composite score of 5 metrics over an 18-month intervention period.
H0 (Null)	No measurable difference exists between stigmergic coordination and the existing municipal mechanism.

Abstract

This document defines the experimental protocol for the Barcelona Sandbox: the controlled environment in which sOS's core hypothesis will be tested with real urban data and real participants. It specifies the research question, falsifiable hypotheses, experimental design (controlled, with baseline comparison), participant recruitment, the minimum-viable sOS system to deploy, data sources, pre-registered metrics, success/failure criteria, ethical safeguards, and a 24-month timeline. This is not an architecture paper – it is a pre-registered experimental design. Results from this experiment determine whether the sOS architecture proceeds to Phase 2 or is revised.

1. Research Question

> **Does stigmergic coordination through labeled traces and personal AI agents produce measurably better urban resource governance outcomes than the existing centralized municipal mechanism?**

"Better" is operationally defined by 5 pre-registered metrics (see §8). "Urban resource governance" is operationally scoped to streetlight maintenance in Barcelona's Eixample district as the initial test domain. The experiment does not test the full sOS architecture – it tests the core coordination claim with the minimum viable system.

The entire sOS thesis rests on one claim: distributed coordination handles complexity better than centralized command. Every working paper, every architectural component, every budget line depends on this. If the sandbox shows no measurable difference between stigmergic coordination and municipal mechanisms, the thesis requires significant revision. If stigmergic coordination performs *worse*, the architecture is wrong. Science advances by being wrong in useful ways – this experiment is designed to be usefully wrong if it fails.

---

2. Core Hypothesis (Falsifiable)

H1 (Primary): Stigmergic coordination through labeled traces and personal AI will achieve ≥15% improvement over the existing municipal mechanism on a composite score of 5 metrics over an 18-month intervention period.

H0 (Null): No measurable difference exists between stigmergic coordination and the existing municipal mechanism.

Falsification condition: If the composite score difference is <5% (within measurement noise) or negative, H1 is rejected and the sOS coordination architecture requires significant revision.

2.1 Auxiliary Hypotheses

H2 (Equity): Stigmergic coordination will produce a more equitable distribution of maintenance resources across neighborhoods (measured by Gini coefficient of repair response time across census tracts) than the existing mechanism.

H3 (Responsiveness): Citizen-reported faults (via traces) will have shorter mean time-to-resolution than sensor-detected faults in the existing system, demonstrating that human-in-the-loop labeling improves prioritization.

H4 (Emergence): ≥30% of maintenance actions in the intervention group will be coordinated stigmergically (traces → agent behavior modification) rather than through explicit central dispatch, demonstrating emergent coordination.

---

3. Experimental Domain: Streetlight Maintenance in Eixample

3.1 Why Streetlights

- **Existing IoT infrastructure:** Barcelona has deployed smart street lighting with sensors since 2012. The Sentilo platform aggregates sensor data. The infrastructure exists – no hardware deployment needed.
- **Well-defined outcomes:** streetlights are binary (working / faulty), maintenance is discrete (repair / replace), outcomes are measurable.
- **Moderate complexity:** enough interdependence (weather, usage patterns, neighborhood priorities, budget constraints) to test coordination, not so complex that results are uninterpretable.
- **Low stakes:** streetlight maintenance failures don't harm people, making this ethically appropriate for an experimental governance system.
- **Municipal data availability:** Barcelona Open Data portal (opendata-ajuntament.barcelona.cat) provides access to urban infrastructure data. Specific streetlight datasets must be verified – see §6.

3.2 Scope

District	Eixample (high density, mixed residential/commercial, existing smart lighting)
Streetlights in scope	~500–1,000 (to be confirmed from municipal inventory)
Fault types	Outage, dimming, flickering, physical damage
Maintenance actions	Bulb replacement, electrical repair, fixture replacement, cleaning
Data granularity	Per-light ID, timestamp, fault type, resolution type, resolution time

---

4. Experimental Design

4.1 Design Type

Controlled three-arm design. The original two-arm design (intervention vs. control) could not isolate the stigmergy effect from the "better fault-reporting app" effect. The three-arm design solves this:

Baseline	Months 0–5	Existing municipal mechanism	Existing municipal	Existing municipal
Intervention	Months 6–23	Existing municipal (010 + central dispatch)	Citizen app with FIFO/severity routing, NO stigmergy	App + labeled trace registry + stigmergic prioritization + reputation weighting
Analysis	Months 24–26	Statistical comparison	Statistical comparison	Statistical comparison

Note on baseline: The 3-month live baseline (months 0–2) is augmented by 12 months of historical municipal maintenance data (months −12 to 0) to capture seasonal variation in fault patterns. Streetlight faults vary by season (more usage hours in winter, heat effects in summer). Historical data is requested from municipal records; if unavailable, baseline is extended to 6 months minimum.

4.2 Control and Intervention Arms

Coordination mechanism	Existing municipal: centralized fault detection → work order → contractor dispatch	Citizen reporting app + simple FIFO routing + standard severity classification (no stigmergy, no reputation, no AI)	App + labeled trace registry + stigmergic prioritization + reputation weighting + personal AI
Fault detection	Sentilo sensors + citizen calls to 010	Sentilo + citizen app submissions (FIFO)	Sentilo + citizen traces via sOS interface (stigmergic)
Prioritization	Municipal schedule + emergency override	Simple FIFO with severity flag	Stigmergic: traces + reputation-weighted urgency signals
Maintenance execution	Existing municipal contractors	Existing municipal contractors	Existing municipal contractors
Participants	N/A (standard service)	20–25 residents with app access	20–25 residents with sOS interface

Why the three-arm design matters: If Arm C outperforms Arm A but Arm B also outperforms Arm A by a similar margin, the effect is attributable to technology adoption (better fault reporting), not stigmergy. For the sOS thesis to be supported, Arm C must outperform Arm B specifically on coordination-sensitive metrics (MTTR, equity, precision@k).

5. Participants

5.1 Recruitment

Sample size	20–50 active participants in intervention group (minimum 15 for statistical power)
Recruitment area	Eixample district residents
Recruitment channels	UB mailing lists, neighbourhood associations (associacions de veïns), Decidim platform, community boards
Inclusion criteria	Resident of intervention sub-district, ≥18 years, smartphone access, basic digital literacy, informed consent
Exclusion criteria	Municipal employees in lighting/urban services departments (conflict of interest)
Incentive	€150 total: €25 enrollment, €25 quarterly engagement bonus (×4 quarters), €25 completion bonus. Structured as research honorarium through UB. Token governance credit in sandbox DAO (testnet, no financial value).

Retention design: Streetlight faults are rare events – participants may find nothing to report for weeks. Retention mechanisms: (a) quarterly surveys with bonus payment, (b) neighborhood-level feedback ("your reports helped fix 3 lights this quarter"), (c) community meetup every 6 months (€500/event from contingency budget), (d) trace submission is not the only activity – validating others' traces and reviewing neighborhood status dashboard counts as participation, (e) app sends monthly summary of area lighting health. Expected retention: 70% at 6 months, 50% at 12 months, 40% at 18 months. If retention drops below 40% (<16 participants across arms B/C), the experiment is underpowered and the inconclusive criterion triggers.

5.2 Participant Role

Participants in the intervention group:

1. Receive a simple mobile interface showing streetlight status in their area.

2. Can submit traces: "light #847 is out," "light #312 is flickering," "this corner is dangerously dark."

3. Can see traces from other participants and aggregated urgency signals.

4. Can validate or challenge others' traces (reputation mechanism).

5. Answer quarterly satisfaction surveys.

6. Do NOT make maintenance decisions – the stigmergic system surfaces priorities; existing contractors execute.

5.3 Ethical Safeguards

- **Informed consent:** Participants understand this is a research experiment, not a production service.
- **Right to withdraw:** Any participant can leave at any time without penalty.
- **No service degradation:** The control group continues to receive normal municipal service.
- **Data minimization:** Participants share only streetlight-related traces, not personal data.
- **DPIA:** A Data Protection Impact Assessment will be filed before data collection (per WP9 §3.4).
- **Ethics approval:** Universitat de Barcelona ethics committee review before participant recruitment.
- **Transparency:** All experimental data, analysis code, and results will be published open-access regardless of outcome.

---

6. Data Sources

6.1 Existing Infrastructure (to be verified before experiment start)

Sentilo IoT platform	Operational since 2012, open source	API access via Barcelona City Council / IMI
Barcelona Open Data BCN	Active, CC BY 4.0 datasets	opendata-ajuntament.barcelona.cat
Streetlight sensor data	Exists (smart lighting deployed)	**TO VERIFY:** specific dataset availability, granularity, update frequency
Maintenance records	Municipal contractor logs	**TO VERIFY:** access for research purposes
010 citizen complaint data	Exists as public service	**TO VERIFY:** historical data for baseline

6.2 Data to Be Generated by the Experiment

Citizen traces	Participant submissions via sOS interface	Labeled trace: {light_id, fault_type, timestamp, submitter_id, urgency_flag}
Trace validations/challenges	Other participants	{trace_id, validator_id, action, timestamp}
Stigmergic signals	System-calculated	{light_id, urgency_score, evidence_count, reputation_weight}
Maintenance outcomes	Contractor reports	{light_id, action, timestamp, resolution_time}
Participant surveys	Quarterly	Satisfaction, usability, trust, perceived fairness
System logs	sOS platform	All coordination events for audit and analysis

---

---

7. The Minimum Viable sOS System

This is NOT the full WP8 architecture. It is the absolute minimum needed to test the core coordination hypothesis.

7.1 What Gets Built (Phase 0: Months 1–3)

Trace submission interface	Simple mobile web app: report a faulty light, see nearby lights	PWA (Progressive Web App), no app store needed
Trace registry	Append-only database of all submitted traces	PostgreSQL + public read-only API (no DLT needed for experiment)
Label ontology	Minimal: {light_id, fault_type, urgency, neighborhood}	JSON schema, 5–10 labels
Personal AI	Rule-based filtering + basic LLM prompt: "Here are lights near you. Rank by urgency."	Lightweight LLM integration (phase 1) or rules-only (phase 0)
Stigmergic dashboard	Shows aggregated urgency signals per light/neighborhood	Simple web dashboard
Work order bridge	Converts high-urgency stigmergic signals into municipal work order format	API integration or manual bridge

7.2 What Does NOT Get Built

- ❌ DAO (no token, no voting, no governance infrastructure)
- ❌ UILL (no personal data vaults, no ZK proofs, no smart contract consent)
- ❌ Full Social Machine (debate platform, C2/C3 escalation)
- ❌ DLT / blockchain (unnecessary for single-domain experiment)
- ❌ dApp marketplace
- ❌ Multi-model AI diversity enforcement

The experiment tests whether stigmergic coordination works. If it doesn't, none of the above matters.

7.3 Design Rationale

Building the minimum system forces honest prioritization. If the core mechanism doesn't work, the architecture was wrong. If it does work, the infrastructure can be expanded incrementally with evidence, not faith.

---

8. Metrics (Pre-Registered)

All metrics are registered here before data collection begins. Post-hoc metric selection or adjustment invalidates the experimental design.

Primary outcome: Mean time-to-resolution (MTTR) – selected because it directly measures coordination speed, is objectively measurable from sensor/contractor timestamps, has well-understood statistical properties, and is the operational metric that matters most to residents.

Secondary outcomes: The remaining metrics provide context and test auxiliary hypotheses but do not gate the primary success criterion.

8.1 Primary Outcome

M1	**Mean time-to-resolution (MTTR)**	Time from fault detection to repair completion, in hours	Sensor timestamps + contractor logs	Lower = better

Primary success criterion: M1(Arm C) < M1(Arm A) AND M1(Arm C) < M1(Arm B), with the Arm C vs. Arm B difference statistically significant at $\alpha=0.05$ using a difference-in-differences estimator that controls for baseline fault rates per arm.

8.2 Secondary Outcomes

M2	**Resource efficiency**	Cost per repair (€)	Contractor invoices	Lower = better
M3	**Citizen satisfaction**	Mean score on quarterly survey (1–5 Likert), surveyed in ALL arms (control via 010 users, arms B/C via app users)	Surveys	Higher = better
M4	**Coverage equity**	Gini coefficient of MTTR across neighborhoods within each arm	M1 + geographic aggregation	Lower = more equitable
M5	**Precision@k**	Proportion of top-k prioritized faults that corresponded to highest-impact faults, as rated by independent municipal assessor blind to arm	Prioritization log + assessor ratings	Higher = better coordination

Why M5 replaced emergent coordination ratio: The original M5 ("% of actions driven by stigmergic signals") was largely tautological – in Arm C, the system routes through traces by construction. Precision@k measures whether stigmergic prioritization actually surfaces the *right* faults, which is the outcome that matters. An independent municipal assessor (blind to which arm a fault came from) rates fault impact after resolution; precision@k measures agreement between stigmergic ranking and assessor ranking.

8.3 Statistical Power

A formal power analysis must be completed before the Go/No-Go gate (Month 2). This section defines the framework; the analysis itself depends on historical fault-rate data that is pending municipal access (§13 Q1).

Key parameters for the power analysis:

- **Unit of analysis:** fault events (not participants), since M1–M5 are properties of maintenance actions.
- **Minimum detectable effect:** 15% difference in MTTR between Arm C and Arm B.
- **Required fault count:** To be estimated via simulation using historical Barcelona Eixample streetlight fault data. With ~500–1,000 lights and typical urban fault rates (15–30% annually), expected fault count per arm is 75–300 over 18 months.
- **MTTR distribution:** Typically right-skewed; power analysis must use a log-normal simulation rather than normal assumptions.
- **If the simulation shows power < 0.8 at expected fault counts:** expand the experimental domain (add districts, extend duration, or select a higher-fault domain like waste management).

- **For M3 (satisfaction):** n=40–50 participants across arms B/C with quarterly surveys gives moderate power (0.6–0.7) to detect large effects (Cohen's d > 0.8). Satisfaction is secondary and not gated on statistical significance.

M6	Trace accuracy (verified faults / total traces)	Measures signal quality of stigmergic traces
M7	Participant retention rate	Measures engagement sustainability
M8	Response time variance (standard deviation of MTTR)	Measures consistency of coordination
M9	False-positive rate (traces reporting faults where no fault exists)	Measures trace ecosystem health
M10	Escalation rate (% of traces escalated to manual dispatch)	Measures how often stigmergy fails → centralization

---

---

9. Success and Failure Criteria

9.1 Success

M1(Arm C) < M1(Arm A) AND M1(Arm C) < M1(Arm B), with Arm C vs. Arm B difference statistically significant at $\alpha=0.05$ (difference-in-differences). Precision@k (M5) shows Arm C > Arm B. No negative trend in M3 satisfaction.

→ **Action:** Proceed to Phase 2. sOS stigmergic coordination demonstrated measurable improvement over both the municipal mechanism and a comparable technology intervention. Build Phase 2 infrastructure based on evidence.

9.2 Partial Success / Technology Effect

M1(Arm C) < M1(Arm A) BUT M1(Arm C) ≈ M1(Arm B) (no significant difference between sOS and app-only). Improvement is attributable to the fault-reporting app, not stigmergy.

→ **Action:** Publish finding. The app improved fault reporting – that's valuable. The stigmergy thesis is not supported. Redesign the stigmergic mechanism and re-test, or pivot to investigating what *does* produce coordination improvement.

9.3 Failure

M1(Arm C) ≈ M1(Arm A) (no improvement over existing mechanism).

→ **Action:** Publish negative result. The sOS coordination thesis is not supported by evidence. Revise the architecture. Consider whether the domain (simple binary faults, well-functioning existing system) left insufficient room for any coordination mechanism to differentiate.

9.4 Inconclusive

Insufficient data: participant dropout >50%, data access blocked, municipal cooperation withdrawn, or power analysis simulation shows <0.6 power for the primary outcome at achieved fault counts.

→ **Action:** Document obstacles transparently. Redesign with alternative domain, location, or retention strategy. An inconclusive first experiment is expected; learning from it is the scientific process.

Note on contractor cannibalization: If Arm C improvements come at the expense of Arm A/B (shared contractor pool creates zero-sum resource allocation), the effect is not a coordination improvement but a prioritization bias. This will be checked by monitoring total contractor utilization across arms and ensuring Arm A receives at least its historical service level.

---

10. Timeline

Pre-experiment	M0–M2	Verify data access (Sentilo/Open Data BCN), secure municipal agreement, ethics approval, UB collaboration agreement, build minimum viable system, recruit participants	Go/No-Go: data access confirmed + ≥15 participants recruited
Baseline	M3–M5	Collect 3 months of baseline data (both groups, existing mechanism)	Data quality check
Intervention	M6–M23	18 months of sOS coordination in intervention group	Quarterly data reviews; mid-point analysis at M14
Analysis	M24–M26	Statistical analysis, report writing, pre-registered comparison	
Publication	M27	Publish results open-access regardless of outcome	

---

11. Budget

System development (PWA, trace registry, dashboard, app-only arm B)	€25K	1 developer, 3 months (was 2 due to 3-arm complexity)
LLM integration (basic personal AI for Arm C)	€5K	API costs for experiment duration
Participant incentives	€12K	50 participants × €150 × arms B/C (40–50 participants total) + control group survey incentives (€25 × 20 survey respondents × 4 quarters)
Municipal data access	€0–€5K	Depends on agreement; may be free for research
Ethics review	€0	UB internal process
Analysis & publication	€5K	Statistical consulting, open-access publication
Community events	€2K	3 meetups over 18 months
Contingency	€5K	Unexpected costs, interface iterations
Total	**€54K–€59K**	

Phase 0 verification budget (before full commitment):

Data access verification + municipal agreement	€2K
Power analysis (simulation with historical data)	€2K
Ethics review	€0
Participant recruitment pilot (10 signups)	€1K
Phase 0 total	**€5K**

This is approximately 10% of the Phase 1 budget allocation. The experiment is designed to be cheap enough to run, fail, learn, and iterate.

12. Risks and Mitigations

Data access denied by municipality	Medium	Critical (experiment blocked)	Alternative domain (waste management, park maintenance); alternative city (L'Hospitalet, Sant Cugat)
Insufficient participant recruitment	Medium	High	Extended recruitment period; increase incentive; partner with neighborhood associations
Participant dropout >50%	Medium	High	Quarterly engagement touches; community-building events; analyze dropout reasons
Contractor behavior differs between groups	Low-Medium	Medium	Blinding; contractor briefing emphasizes business-as-usual
Municipal election cycle disrupts cooperation	Medium	Medium	Municipal elections May 2027; secure multi-year agreement pre-election
Technology failure (app downtime, data loss)	Low	Low	Simple architecture; manual fallback for trace submission
Negative press ("City experiments on residents")	Low	Medium	Proactive communication; emphasize consent, ethics approval, right to withdraw
COVID-like external shock disrupts urban patterns	Low	High	Extend baseline or intervention period; document external confounds

13. Open Questions for Phase 0 Resolution

These must be answered before the Go/No-Go gate (Month 2):

1. **Data access:** Does Barcelona Open Data BCN or Sentilo provide per-light streetlight status data at sufficient granularity and update frequency? If not, what alternative urban dataset is available?

2. **Municipal agreement:** Will Barcelona City Council (via IMI or Urban Services) authorize a research experiment using their infrastructure data and contractor network?

3. **UB collaboration:** Does Universitat de Barcelona provide ethics review, recruitment infrastructure, and institutional backing?

4. **Legal:** Does the experiment require a data-sharing agreement? Are there GDPR implications for trace data linked to geographic locations?

5. **Domain suitability:** Is streetlight maintenance in Eixample complex enough to show stigmergic benefit, or is it too simple (few faults, fast resolution already) for any coordination mechanism to differentiate?

---

14. Conclusion

This protocol transforms sOS from a paper architecture into a falsifiable scientific experiment. The Barcelona Sandbox is not a pilot project, a demo, or a proof-of-concept – it is a controlled experiment designed to test whether the core claim of sOS (distributed stigmergic coordination outperforms centralized governance for complex resource allocation) is true.

If the experiment succeeds: the architecture has empirical foundation. Build Phase 2.

If the experiment fails: the architecture is wrong. Revise it. That is science.

If the experiment is inconclusive: learn, adapt, try again. That is also science.

What is not acceptable: building the full system without testing the core hypothesis. That is engineering without science – and sOS is a scientific research program, not a startup.

Revised draft – Branch A (Pi/DeepSeek) + Branch B (Anthropic Opus 4.8) + Synthesis. Revisions applied 2026-06-28: three-arm design (control/app-only/full sOS), MTTR as primary outcome (composite demoted to secondary), M5 replaced with precision@k, power analysis framework defined, seasonality addressed (12-month historical baseline), retention redesigned (€150 structured honorarium + community events), three-arm budget (€54–59K), Phase 0 verification gate (€5K), contractor cannibalization check, geo-trace re-identification noted for DPIA. Pending Steward review.