

# **Polish Looted Art Surveillance System**

## **Executive Summary**

Technical Surveillance System for Cultural Heritage Recovery

February 2026

For Museum Directors and Ministry Officials

# Table of Contents

## **Polish Looted Art Surveillance System**

Executive Summary

System Overview

Operational Workflow

Scoring and Prioritization

Technical Architecture

Costs

Optionality and Expansion

Timeline

Performance Metrics

Risks and Mitigations

Success Criteria

Resource Requirements

Expansion Scenarios

Decision Points

Conclusion



# Polish Looted Art Surveillance System

## Executive Summary

## System Overview

This system automates the continuous monitoring of auction houses, galleries, and online marketplaces for potential matches to looted Polish artworks. It combines Google Vision API reverse image search with web scraping to generate daily findings, which are filtered by rule-based scoring and reviewed by a single curator.

The system operates continuously with minimal human oversight. A curator reviews 50-100 filtered findings per day (30-60 minutes). As the curator reviews results, the system captures feedback to refine filtering rules and build a training dataset for future machine learning classification.

## Operational Workflow

### Daily Automated Process

#### Overnight (2:00 AM - 6:00 AM):

- Vision API: Scans 2,000 high-priority artworks (images uploaded to Google Vision for reverse image search)
- Web Scraping: Visits 20+ auction houses, galleries, and marketplaces
- Processing: Extracts findings, calculates suspicion scores, deduplicates results
- Queue: Adds high-scoring findings to review queue

#### Morning (9:00 AM):

- Curator receives email summary: "15 findings require review"
- Review interface shows side-by-side comparison: original artwork vs. found image
- Context provided: domain (Christie's, eBay, museum site), match confidence, historical pattern
- Curator makes decision: Approve, Reject, False Positive, Monitor
- Notes captured for legal follow-up or future reference

#### System Learning (Weekly):

- Analyzes curator's decisions from past week
- Adjusts scoring rules: increase weight for categories with high true-positive rate, decrease for categories with high false-positive rate
- Updates filtering thresholds to maintain target false-positive rate

## Scoring and Prioritization

Findings are scored based on:

- **Domain type:** Auction houses score highest, museums lowest
- **Match quality:** Exact visual match vs. partial match vs. visual similarity
- **Context:** Listing shows sale price, provenance gaps, or anonymous seller
- **History:** Domain has previous confirmed matches

Threshold for curator review: Score  $\geq 15$  (out of 100) Initial false-positive rate: 15-20% Target false-positive rate after 6 months: <5%

## Technical Architecture

### Core Components

**Task Scheduling:** Celery + Redis for distributed task queue and beat scheduler **Database:** PostgreSQL (migrated from SQLite for concurrent access and JSONB indexing) **Web Scraping:** Scrapy framework with site-specific spiders **Vision API:** Google Cloud Vision web detection **Similarity Search:** FAISS for fast nearest-neighbor search on CLIP embeddings **Review Interface:** FastAPI + HTML/JavaScript frontend

### Data Storage

**Artworks:** Metadata, images, provenance, scan priority (1-5) **Vision Results:** Tiered storage (summary for all, details for interesting only) **Findings:** All matches with domain, confidence, and review status **Review Feedback:** Curator decisions, notes, confidence ratings **Training Data:** Accumulated labeled examples for future ML models

### Anti-Detection Measures

- Rate limiting: 1 request per 5 seconds per site
- User-agent rotation
- Strict robots.txt compliance
- Circuit breakers: disable spider after 10 consecutive failures
- Polite headers with research designation and contact information

## Costs

### Monthly Operating Costs

#### Infrastructure:

- VPS hosting (4 CPU, 16GB RAM, 500GB storage): \$100-150
- Redis (managed or included in VPS): \$0-20
- PostgreSQL (managed service): \$50-80

- **Subtotal: \$150-250/month**

#### **API Costs:**

• Google Vision API: 2,000 requests/day × 30 days = 60,000 requests/month

• Cost: \$90-180/month (depending on Google Cloud pricing tier)

- **Subtotal: \$90-180/month**

**Total Operating: \$240-430/month (~\$3,000-5,000/year)**

## **One-Time Development Costs**

Software engineering: 5-6 months (Phases 1-5)

- Automation infrastructure: 4 weeks
- Multi-site scraping: 4 weeks
- Review interface and feedback loop: 4 weeks
- Vector search and deduplication: 4 weeks
- Production deployment and monitoring: 4 weeks

**Optional Phase 6 (ML classification):** Additional 2-3 months after 6 months of operation and data collection

## **Cost Scaling Options**

#### **Expand scanning volume:**

- Daily scan all 65,000 artworks: +\$2,500/month Vision API cost
- Add more sites (30-40 total): +\$50-100/month infrastructure
- More frequent rescanning (2x/day for flagged items): +\$200/month

#### **Reduce costs:**

- Scan fewer artworks daily (1,000 instead of 2,000): -\$90/month
- Weekly instead of daily scanning: -75% Vision API cost
- Focus only on highest-priority artworks: variable reduction

## **Optionality and Expansion**

### **Site Coverage**

**Initial scope:** 20 sites (5 major auction houses, 10 secondary auctions, 5 galleries/marketplaces)

#### **Expansion options:**

- Regional auction houses (European, Asian): +10-15 sites
- Specialized galleries (Old Masters, Eastern European art): +5-10 sites
- Private dealer networks: +20-30 sites
- Museum acquisition databases: +15-20 sites

**Effort to add new site:** 1-3 days engineering per site (depends on complexity)

## Human Review Capacity

**Current design:** Single curator, 30-60 minutes/day

### Scaling options:

- Add second curator for weekend coverage: 0.2 FTE
- Add specialist for specific periods (e.g., Baroque expert): 0.1-0.3 FTE
- Peak period support (auction season): temporary 0.5 FTE
- Total remains well below 1 FTE even with expansion

### Work distribution:

- Primary curator: All auction/marketplace findings
- Specialists: Category-specific review (by art period, region, or medium)
- Administrative: Weekly false-positive analysis and rule adjustment (1-2 hours/week)

## Dataset Expansion

**Current focus:** Polish looted art (65,000 artworks)

### Additional applications:

- Other national looted art databases (share infrastructure, add data)
- Acquisition target monitoring: Track specific artists or periods for museum purchases
- Provenance research: Monitor movement of specific high-value pieces
- Authenticity verification: Track appearance of known forgeries or replicas

### Cross-institutional potential:

- Partner with other national museums (Czech, Hungarian, etc.)
- Share scraping infrastructure and findings
- Distribute API costs across institutions
- Build collaborative review queue

## Timeline

### Phase 1-2: Automation and Basic Coverage (Months 1-2)

- Daily Vision API scanning operational
- Database migration to PostgreSQL
- Initial scraping of 5-10 major sites
- **Deliverable:** System runs daily without manual intervention

### Phase 3: Review System (Months 3-4)

- Curator review interface deployed
- Feedback capture operational

- First iteration of rule refinement
- **Deliverable:** Curator can review findings in structured workflow

## Phase 4-5: Scale and Production (Months 5-6)

- Expand to 20+ sites
- Fast similarity search (FAISS)
- Production deployment with monitoring
- **Deliverable:** Robust 24/7 operation with cost controls

## Phase 6: Machine Learning (Months 9-12)

- Requires 3-6 months of curator feedback first
- Train classification model on accumulated reviews
- Automate majority of filtering decisions
- **Deliverable:** Curator reviews only 20-30 items/day instead of 100

**Minimum viable system:** 2-3 months (Phases 1-3 only) **Production-ready:** 5-6 months (Phases 1-5) **ML-enhanced:** 12+ months (requires data collection period)

## Performance Metrics

### System Coverage

- Sites monitored: 20+ major auction houses and galleries
- Artworks tracked: 2,000-65,000 (configurable)
- Detection latency: 24 hours from listing to curator alert
- Uptime target: >99% (8-9 hours downtime per year)

### Curator Efficiency

- Review time: 30-60 minutes/day
- Findings reviewed: 50-100/day initially, 20-30/day after ML
- False-positive rate: 15-20% initially, <5% after refinement
- Time savings vs. manual search: ~40 hours/week

### Cost Efficiency

- Cost per finding: \$4-8 (includes all false positives)
- Cost per true positive: \$50-150 (assuming 5-15% true-positive rate)
- Annual operating cost: \$3,000-5,000 base, scales with coverage

## Risks and Mitigations

## Technical Risks

**Website blocking (Medium probability)** Sites may block automated access. *Mitigation:* Respectful scraping with clear identification, rate limiting, legal review. *Fallback:* Manual monitoring of 3-5 highest-priority sites.

**API cost overruns (Low-Medium probability)** Vision API costs could exceed budget if unchecked. *Mitigation:* Daily cost caps with automatic pause. *Fallback:* Reduce scanning frequency or limit to highest-priority subset.

**Scraper breakage (High probability over 12 months)** Website redesigns will break scrapers. *Mitigation:* Automated health checks with immediate alerts. *Response:* 1-2 days engineering to update per broken scraper.

## Operational Risks

**Curator review fatigue (Medium probability)** Too many false positives could reduce curator engagement. *Mitigation:* Conservative initial thresholds, rapid iteration based on feedback. *Adjustment:* Increase threshold to reduce volume, accept lower recall.

**Legal concerns (Low probability)** Website terms of service may prohibit scraping. *Mitigation:* Legal review before launch, focus on sites with public-good or research exceptions. *Alternative:* Use only Vision API (no direct scraping) for sensitive sites.

## Success Criteria

### Months 1-3

- Zero manual interventions required for daily scans
- 10+ sites scraped successfully
- Curator reviewing findings regularly
- False-positive rate measured and below 25%

### Months 6-8

- 20+ sites monitored
- 50,000+ artworks searchable
- Curator workload stable at 30-60 minutes/day
- At least 2 confirmed matches leading to restitution discussions

### Month 12+

- ML classification operational
- False-positive rate below 5%
- 80% of findings auto-classified

- System has identified 5-10 high-value leads

## Resource Requirements

### Personnel

#### Required:

- Expert curator: 30-60 minutes/day (0.1-0.15 FTE)
- Technical administrator: 4-6 hours/month (0.05 FTE)
- Legal liaison: as needed for confirmed matches

#### Optional:

- Additional curators for expanded coverage: 0.1-0.3 FTE each
- Specialist reviewers (by art period or region): 0.05-0.1 FTE each
- Total human effort remains well below 1 FTE

### Technical

- Development: 5-6 months engineering work
- Server: Dedicated VPS or cloud instance
- Database: Managed PostgreSQL service
- API access: Google Cloud Vision account

## Expansion Scenarios

### Geographic Expansion

Current focus: International auction houses and major markets Add: Regional European auctions, Asian markets, Middle Eastern dealers Cost impact: +\$50-150/month infrastructure, +1-3 days engineering per region

### Institutional Collaboration

Partner with other national museums experiencing similar challenges Share: Scraping infrastructure, API costs, review interfaces Benefit: Distributed costs, shared intelligence on suspicious domains Model: Consortium approach with per-institution data isolation

### Alternative Datasets

Beyond looted art monitoring:

- Acquisition targets: Monitor specific artists or periods for purchase opportunities
- Provenance research: Track specific high-value pieces across time

- Market intelligence: Analyze pricing trends and dealer networks
- Authentication: Monitor appearance of known forgeries

Additional development: 2-4 weeks per new use case

## Decision Points

To proceed, leadership should determine:

1. **Budget authorization:** \$300-450/month operation + development investment
2. **Curator assignment:** Designate reviewer and establish escalation procedures
3. **Site priorities:** Identify 5-10 highest-priority auction houses and galleries
4. **Legal clearance:** Review scraping approach with legal counsel
5. **Timeline preference:** Faster deployment (3 months minimum viable) vs. comprehensive build (6 months production-ready)
6. **Expansion scope:** Initial coverage (20 sites) vs. aggressive expansion (40+ sites)

## Conclusion

This system transforms looted art recovery from an opportunistic, manual process into a systematic surveillance operation. It handles the repetitive work of daily monitoring while preserving human expertise for decision-making.

The core value is operational leverage: one curator effectively monitors coverage that would require a team of 10-15 people working full-time through manual methods. As the system learns from curator decisions, efficiency continues to improve.

The design prioritizes practical deployment and cost control over technical sophistication. It can begin operation within 2-3 months and scale incrementally based on results and available resources.

*Prepared: February 2026 Document: Executive Summary - Technical Surveillance System for Cultural Heritage Recovery*