AI-Assisted Recovery of Historical Wildlife Data from Legacy Field Reports

A Case Study Using Robert D. Jones Jr.'s Izembek Refuge Reports, 1948–1974

Kim Bridges
with Claude AI (Anthropic)

Background Paper for the AI Workshop
International Society for Ethnobotany

April 2026

Source Material	AI Extraction	Outputs
38 quarterly/annual reports ~1,300 pages typescript 27 years (1948–1974) Scanned PDF format	72 species identified 875 place names extracted 210 curated locations (deduplicated) 52 sea otter population records	38 annotated digital editions Structured data (CSV) Interactive map (HTML) Conservation narrative (DOCX)

Abstract

Decades of irreplaceable field observations remain locked in legacy documents—typed reports, handwritten notes, and carbon copies—that resist conventional digitization. We demonstrate an AI-assisted methodology for recovering structured scientific data from 38 quarterly and annual wildlife refuge reports written by Robert D. Jones Jr. ("Sea Otter Jones") during his tenure as manager of the Izembek National Wildlife Refuge, Cold Bay, Alaska (1948–1974). Using a vision-language OCR model (olmOCR) to read degraded typescript that defeated commercial OCR tools over two decades of attempts, followed by AI-orchestrated parsing, error correction, species identification, and place-name extraction, we recovered ~1,300 pages of text encompassing 72 identified species and 875 geographic references (deduplicated to 210 curated locations by resolving OCR variant spellings). A focused case study on sea otter (Enhydra lutris) population recovery produced a curated dataset of 52 population records, an interactive geographic map, and an ethnobiological narrative constructed from Jones' own words—documenting the species' recovery from near-extinction through nuclear testing, commercial harvest, and catastrophic ice events. The methodology is replicable and applicable to any corpus of historical field-biology literature.

Keywords: *historical ecology, data recovery, AI-assisted OCR, ethnobiology, sea otter, wildlife management, legacy documents, natural language processing*

---

1. Introduction

The biological sciences hold vast repositories of observational data recorded before the digital era. Field notebooks, typed reports, station logs, and specimen records constitute an irreplaceable archive of ecological conditions that predate contemporary monitoring programs—often by decades. These records document baseline conditions, population trajectories, species distributions, and environmental changes that are essential for understanding long-term ecological dynamics, particularly in the context of climate change. Yet much of this material remains effectively inaccessible: locked in filing cabinets, deteriorating in archives, or existing only as low-quality scans that resist automated text extraction.

The challenge is not merely optical. Legacy documents present a constellation of difficulties: faded ink and carbon copies, inconsistent typefaces, handwritten annotations, tabular data embedded in prose, domain-specific terminology, and structural conventions that vary by era and institution. Commercial optical character recognition (OCR) tools—including Adobe Acrobat, ABBYY FineReader, and their successors—have been tested repeatedly against such materials over two decades, with uniformly poor results. The documents are legible to human readers who bring contextual understanding; they are opaque to algorithms that rely on pattern matching alone.

Recent advances in artificial intelligence—specifically vision-language models and large language models (LLMs)—offer a fundamentally different approach. Rather than matching character shapes, these models *understand* what they are reading. A vision-language OCR model can resolve ambiguous characters by recognizing that the word in context must be "ptarmigan" rather than "ptamigan," because it understands the domain. An LLM can parse a narrative report into structured sections, identify

species names even when misspelled, extract population counts from prose descriptions, and distinguish a geographic place name from a personal name—all tasks that require comprehension, not just recognition.

This paper describes the application of these technologies to a specific and significant corpus: the quarterly and annual refuge narrative reports written by Robert D. Jones Jr. during his 27-year tenure (1948–1974) as manager of the Cold Bay Game Management Area and Izembek National Wildlife Refuge on the Alaska Peninsula. Known to colleagues as "Sea Otter Jones," he produced 38 reports totaling approximately 1,300 pages of typescript—a continuous record of wildlife observations, population counts, management actions, and environmental conditions spanning a period of extraordinary ecological change.

2. The Source Material

2.1 The Reports

Jones' reports follow the standard U.S. Fish and Wildlife Service refuge narrative format: quarterly accounts covering weather, habitat conditions, wildlife populations, management activities, and administrative matters. The earliest reports are relatively brief; by the 1960s, they had grown into detailed annual narratives of 40–60 pages. The reports are typed, sometimes as carbon copies, on standard government-issue paper. Some include hand-drawn maps, tabular data, and appended census forms.

The surviving documents are scanned PDFs of variable quality. Common degradation includes: faded or uneven ink density, carbon-copy artifacts, page-edge shadows, staple marks, handwritten marginalia, and typewriter-specific irregularities (misaligned characters, varying strike pressure). Several reports exist as second or third carbons with significantly reduced contrast.

2.2 Previous Digitization Attempts

Multiple attempts were made over the period 2005–2023 to digitize these reports using commercial OCR software, primarily Adobe Acrobat. Each new software version was tested against the same documents. Results were consistently unusable: character recognition rates were low, formatting was destroyed, and the output required more effort to correct than to retype manually. The documents—while clearly legible to a human reader—remained beyond the capabilities of conventional OCR technology.

2.3 Scientific Significance

The reports span a period of exceptional ecological importance. The 1948–1974 window encompasses: the post-war recovery of sea otter populations from near-extinction by the fur trade; three underground nuclear tests at Amchitka Island (1965, 1969, 1971) within the core sea otter habitat; Alaska statehood (1959) and the resulting jurisdictional shifts in wildlife management; the beginning of large-scale commercial sea otter harvest; catastrophic Bering Sea ice events; and the early decades of what would become one of the longest-running waterfowl monitoring programs in Alaska. The reports also document species distributions, weather patterns, and habitat conditions that constitute a

pre-climate-change baseline for the region.

3. Methodology

The recovery pipeline operates in four phases: OCR extraction, structural parsing, knowledge extraction, and output generation. The entire workflow was orchestrated through iterative collaboration between a human researcher and an AI assistant (Claude, Anthropic), with the AI performing automated processing and the human providing domain expertise, quality validation, and editorial judgment.

3.1 Phase 1: AI Vision OCR

Each PDF page was processed through olmOCR-2-7B (Allen AI), a vision-language model designed for document understanding, accessed via the DeepInfra API. Unlike traditional OCR, which segments an image into individual characters and matches shapes, olmOCR processes each page as a visual scene and generates text by *comprehending* the document's content. This contextual understanding proved critical for resolving the ambiguities inherent in degraded typescript.

The model successfully processed pages that had defeated commercial OCR for nearly twenty years. While not perfect—certain character-level errors persisted, particularly in proper nouns and specialized terminology—the output was coherent, structurally sound, and amenable to systematic correction.

3.2 Phase 2: Error Correction and Structural Parsing

Raw OCR output was refined through two parallel processes. First, a curated correction dictionary (corrections.json) applied pattern-based find-and-replace operations targeting known OCR errors—misspelled place names, species names, and recurrent character substitutions. This dictionary was built iteratively through human review of output samples.

Second, a structural parser (report_parser.py, ~550 lines of Python) decomposed each report into machine-readable JSON. The parser detected section boundaries, identified and reconstructed tabular data, filtered garbled text passages (sequences with anomalous punctuation density or HTML tag artifacts from the OCR process), and tagged content by type (narrative, table, heading, metadata).

3.3 Phase 3: Knowledge Extraction

From the 38 parsed report JSONs, three classes of structured knowledge were extracted:

Species identification. Regular expression patterns matched species names across all reports, producing an index of 72 distinct species with per-report mention counts and taxonomic classification (mammal, bird, fish, invertebrate). For species with generic/specific ambiguity (e.g., "otter" could mean sea otter or land otter), a contextual classification system used surrounding keywords—marine, kelp, island for sea otters; trap, pelt, fur for land otters—to assign probabilistic categories.

Geographic references. Place-name extraction identified 875 geographic references across all reports. However, raw extraction is only the first step. Because OCR introduces character-level errors into proper nouns, a single location may appear under multiple variant spellings. "Izembek Bay," for example, was rendered by the OCR model as 21 distinct strings—Iseabek, Iseebek, Isemek, Isenbok,

Issebok, Izenbek, and 15 others—depending on the quality of the source page. Amchitka Island appeared as Anchitka, Anchatka, and Architka. Morzhovoi Bay as Morshovoi, Morshovod, and Marshovoi.

A systematic deduplication process—using contextual knowledge of Alaska geography, feature-type classification from name patterns (words ending in Island, Bay, Point, Lake, etc.), and manual review of high-frequency entries—collapsed the 875 raw extractions into 210 unique locations. Each entry in the curated gazetteer preserves its OCR variant names, providing a transparent record of the deduplication decisions. The variants column itself serves as documentation of how OCR quality varied across the source material. Locations were classified by feature type (69 islands, 45 bays, 25 lakes, 24 points, 15 streams, 11 lagoons, and others), assigned to geographic regions, and geocoded where possible. Year ranges were compressed to show temporal spans of mention (e.g., "1948–1956; 1958–1971; 1974"), revealing both the continuity and the gaps in Jones' geographic coverage.

Population data. For the sea otter case study, all 437 otter mentions were extracted and classified. Numeric values were filtered to exclude dates, page numbers, specimen IDs, and measurements. Each remaining count was categorized by type (aerial survey, census form, observation, estimate, harvest, mortality, qualitative) and validated against the raw OCR source text, producing 52 curated population records.

3.4 Phase 4: Output Generation

Four categories of output were generated from the extracted knowledge:

Annotated digital editions (38 DOCX + 38 PDF files). Each original report was reconstituted as a formatted document with corrected text, species names italicized, section headings restored, and metadata headers. These preserve the full content of the originals in a searchable, archival format.

Structured data (CSV). The curated sea otter dataset provides analysis-ready records with standardized fields: year, location, count, count type, context, source report, and species notes.

Interactive map (HTML/Leaflet). Forty-six mapped locations display population data as proportionally scaled markers, color-coded by era, with multiple base map layers and toggleable overlays.

Conservation narrative (DOCX). An ethnobiological document organized in eight thematic sections, constructed from Jones' own words—his observations, concerns, frustrations, and reflections—telling the story of sea otter recovery through the eyes of the person who witnessed it.

4. The Sea Otter Case Study

The sea otter (*Enhydra lutris*) was selected as the focal species for the initial case study for several reasons. It is a keystone species whose population trajectory across 1948–1974 tells a story of continental significance: recovery from fur-trade decimation, nuclear-age disruption, and environmental catastrophe. Jones was personally invested in these animals—colleagues called him "Sea Otter Jones"—and his reports contain some of his most vivid and emotionally engaged writing about them. The species also presented a clean test case for the data extraction methodology: population counts are relatively unambiguous, locations are specific, and the recovery trajectory provides a built-in

validation framework.

4.1 Population Recovery Arc

The extracted data reveal a clear geographic progression. In 1949, the core population of approximately 3,420 animals was concentrated at Amchitka Island in the central Aleutians—a remnant of the once-vast population that had been hunted to the brink of extinction during the Russian and American fur trades. By 1951, regional surveys documented 550+ animals at sites beyond Amchitka, including the Shumagin Islands. In 1953, 633 were counted at the Shumagins and 321 at Prince William Sound. The recovery frontier moved steadily eastward.

The colonization of Izembek Bay provides the most detailed trajectory in the dataset. A single sea otter appeared at Cold Bay in 1955. The first animal was documented in Izembek Bay itself on October 13, 1960. Census records then track the growing colony: 9 (1962), 13 (1963), 23 (1964), 25 (1965). By 1967, Jones reported sea otters "now sufficiently common in Izembek Bay" to be encountered routinely, with whole family groups present.

4.2 The Bomb and the Otter

The most dramatic episode in the dataset concerns the conflict between nuclear testing and wildlife conservation at Amchitka Island. The Atomic Energy Commission conducted three underground nuclear tests on the island: Long Shot (1965), Milrow (1969), and Cannikin (1971). Cannikin, at five megatons, was the largest underground nuclear test in United States history. Jones was directly involved in attempts to protect the sea otter population, but found himself powerless against the AEC:

> *"We were in effect without portfolio... Our repeated recommendations and pleas fell on deaf ears."*
>
> — Jones, Annual Report, 1971

The Cannikin test killed "perhaps several hundred" sea otters. Jones documented the aftermath with the precision of a scientist and the anguish of a conservationist who had watched this population for over two decades.

4.3 Ice and Endings

Jones' final report (1974) documents a catastrophic Bering Sea freeze—the third in four years—and its devastating impact on sea otters:

> *"These animals cannot maintain holes in the ice as seals do, and when their world freezes they have little recourse but to die... except for the 25 we intercepted and released in Cold Bay on the Pacific Ocean, none survived."*
>
> — Jones, Final Report, 1974

These passages illustrate what structured data alone cannot capture: the lived experience of a field biologist witnessing both the recovery and the destruction of a species he had devoted his career to studying. They are precisely the kind of material that ethnobiological approaches are designed to preserve.

5. Results and Products

The complete processing of 38 reports produced the following deliverables:

Product	Format	Scale	Purpose
Annotated Editions	DOCX + PDF	38 reports	Archival preservation; searchable text
Species Index	JSON	72 species	Cross-report species tracking
Gazetteer	CSV	210 places	Deduplicated; typed; with variants
Sea Otter Data	CSV	52 records	Population analysis; time series
Interactive Map	HTML/Leaflet	46 locations	Spatial visualization of recovery
Conservation Narrative	DOCX	8 sections	Ethnobiological story; Jones' voice
Timeline	HTML	27 years	Chronological context; events
Methodology Flowchart	SVG	4 phases	Pipeline documentation

Table 1. Summary of products generated from the Jones report corpus.

6. Discussion

6.1 The Ethnobiological Dimension

This project began as a data recovery exercise and evolved into something more: an ethnobiological investigation. Jones' reports are not merely containers of numeric data. They are the written record of a person who spent 27 years observing, managing, and caring about a landscape and its wildlife. His language shifts across the decades—from the measured optimism of the early surveys to the frustration of the nuclear testing era to the grief of the final freeze. Extracting his words alongside his numbers preserves both dimensions of the scientific record.

The ethnobiological framing proves especially powerful for the sea otter narrative. The species' recovery arc—from fur-trade remnant to growing population to nuclear casualty to ice-bound death—is a conservation story that resonates far beyond the Aleutian Islands. But the story's emotional and moral weight comes from Jones' voice: his awe at watching an otter sleep so deeply it could be touched, his careful documentation of parental care, his rage at bureaucratic indifference to nuclear testing impacts, and his sorrow at the ice that killed the animals he had spent a career protecting.

6.2 AI as Collaborative Tool

The role of AI in this project warrants careful characterization. The AI did not independently "discover" the sea otter story or decide which passages were significant. Rather, it served as a powerful tool that dramatically accelerated tasks that would otherwise have been prohibitively time-consuming: reading 1,300 pages of degraded text, identifying all mentions of a species across 38 documents, classifying ambiguous references by context, extracting numeric data from narrative prose, and generating formatted outputs.

The human researcher provided what AI cannot: deep domain knowledge (decades of ecological expertise), historical context (knowledge of the region, the era, and the people involved), editorial judgment (which passages matter, which numbers are reliable, which themes resonate), and validation through personal networks (Dr. C. Peter McRoy, who worked with Jones at the refuge, confirmed details and provided the "Sea Otter Jones" nickname). The collaboration was genuinely synergistic: neither partner could have produced these results alone.

6.3 Replicability

The methodology is deliberately designed for replication. Any corpus of historical field-biology reports can be processed through the same pipeline: vision OCR, structural parsing, knowledge extraction, multi-modal output. The specific tools may evolve—newer OCR models, better parsers, more capable LLMs—but the architecture is stable. The critical insight is that AI vision models can read documents that defeated two decades of conventional OCR, and that LLMs can perform the kind of contextual classification (species disambiguation, count extraction, thematic organization) that previously required expert human readers.

For the ethnobiology community specifically, this methodology offers a path to recovering historical observations, traditional ecological knowledge, and field-biology narratives that are at risk of being lost as the paper records deteriorate and the people who created them pass away. The urgency is real: Jones died in 1997, and without colleagues like McRoy, the institutional memory surrounding these reports would be significantly diminished.

6.4 Limitations and Future Directions

The AI OCR is not perfect. Character-level errors persist, particularly in proper nouns, and the correction dictionary requires ongoing human curation. The gazetteer deduplication process illustrates both the challenge and the opportunity: "Izembek Bay" appeared as 21 distinct OCR variants, requiring contextual knowledge to resolve. The curated gazetteer preserves these variant forms as a transparent audit trail. Numeric data extraction required careful filtering to separate population counts from dates, page numbers, and measurements.

The sea otter case study demonstrates the methodology for a single focal species. Extending the approach to the full set of 72 identified species—each with its own disambiguation challenges, count types, and narrative richness—is a natural next step. Weather data, habitat observations, and management action records remain to be extracted as structured datasets. Each extension applies the same pipeline architecture to a different facet of the Jones corpus.

7. Conclusion

The recovery of Robert D. Jones Jr.’s wildlife reports demonstrates that AI-assisted methods can unlock historical scientific data that was previously inaccessible. More importantly, it shows that the value recovered extends beyond numbers. The combination of structured data extraction with ethnobiological narrative construction produces a richer, more complete, and more human record than either approach alone.

Reinvigorating "old" data gives respect to the people who devoted their lives to its collection and reporting. With these new tools, we honor those individuals while enhancing our own understanding of the ecological systems they documented. For the ethnobiology community—and for any discipline that values the continuity between past observation and present knowledge—this is both a practical methodology and an ethical imperative.

Companion Materials

The following materials accompany this paper and are available as digital files:

• **Sea_Otter_Narrative.docx** — The conservation narrative, constructed from Jones' own words across eight thematic sections.

• **sea_otter_data.csv** — Curated dataset of 52 sea otter population records (1949–1974).

• **sea_otter_map.html** — Interactive Leaflet map with 46 locations, era overlays, and multiple base maps.

• **jones_timeline.html** — Interactive timeline of key events across Jones' 27-year tenure.

• **jones_gazetteer.csv** — Curated gazetteer of 210 unique locations deduplicated from 875 raw extractions, with feature types, regions, coordinates, and OCR variant names preserved.

• **methodology_flowchart.svg** — Visual diagram of the four-phase processing pipeline.

• **sea_otter_analysis.md** — Detailed analytical notes on the sea otter data extraction methodology.

This document was produced through human–AI collaboration using Claude (Anthropic). The AI assisted with OCR processing, data extraction, document generation, and text composition. All scientific judgments, editorial decisions, and domain validation were performed by the human researcher.