



Disambiguating Multi-Lingual Global Author Profiles at AI Scale



Case Study

Overview



The world's largest technical professional organization possessed massive, valuable repositories of author metadata. The information existed as fragmented, unstructured records scattered across decades of global scientific publications. The organization wanted to develop an automated solution to process, standardize, and disambiguate author information from various sources to create a unified and reliable author database.

Straive deployed its proprietary AI-powered platform to extract, parse, and normalize this unstructured content, transforming raw text fields into a unified, high-integrity information asset.

Business Challenges



- **Massive Unstructured Content Volumes:** Decades of unstandardized string entries, embedded in PDF headers, footers, and reference lists, created severe pipeline engineering bottlenecks.
- **High Variance & Low Record Quality:** Inconsistent syntax, multi-language character sets, and missing attributes made raw inputs disparate and challenging to query.
- **Complex Entity Resolution:** Overlapping details (e.g., identical author string names linked to different real-world entities) prevented the construction of a reliable relational database schema without labor-intensive manual cleaning.

Straive Solution



Straive utilized its proprietary **AuthEntik™** AI platform to convert raw, unstructured document text into structured, schema-compliant feeds.

Custom machine learning models were trained on domain-specific metadata formats to extract target fields automatically. The platform then ran complex entity-matching algorithms to generate normalized streams directly into the enterprise database.

Key Components



Computer Vision & Auto-Zoning Engine: Automatically mapped spatial document regions to identify and segment raw metadata zones within unstructured files.



Multilingual NLP Extraction: Processed heterogeneous text inputs across multiple languages, isolating distinct attributes (names, affiliations, citations) into standardized key-value pairs.



Algorithmic Disambiguation: Executed probabilistic matching algorithms against master reference datasets to resolve duplicate records and merge fragmented points into clean entity nodes.



Human-in-the-Loop QC Framework: Integrated expert analysts to validate low-confidence algorithmic outputs, ensuring 100% record fidelity before database ingestion.

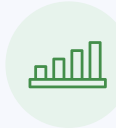
Impact Delivered



12 Million Unstructured Records Processed:
Cleansed and standardized raw inputs in just 3 months.



63,840 Disambiguated Entities Created:
Built a fully structured, multi-dimensional knowledge graph enabling advanced search, analytics, and record mapping.



Enterprise-Scale ROI:
Converted high-maintenance unstructured backfiles into a high-value, queryable enterprise data asset.

Industry-Agnostic Solution

At its core, this project demonstrates how AI can solve the universal enterprise challenge of converting messy, unstructured data into clean, structured assets.

This challenge spans every industry. The same pipeline processes unstructured clinical notes in Healthcare, unstandardized loan applications in Banking, and raw supplier specs in Supply Chain. By combining visual auto-zoning, NLP attribute extraction, and algorithmic resolution, Straive turns dark, unusable content into actionable intelligence.

Is unstructured data holding back your enterprise analytics and system performance?

[Contact Straive today](#)

To discover how our custom AI pipelines can structure, cleanse, and monetize your legacy assets.



About Straive

At Straive, we operationalize Data Analytics and AI for global enterprises and work with several Fortune 500 companies. We don't just build top-tier data analytics and AI solutions— we integrate them seamlessly into your core workflows. This approach drives enhanced efficiency, improves user experience, and boosts client revenue, setting you apart from the competition.

Straive serves clients in various industries, including Banking and Financial Services, GCC, Retail and CPG, Media, Sports and Entertainment, EdTech, Science and Research, Logistics and Supply Chain, Education, and Pharmaceutical and Life Sciences.

Our strategically positioned resource pool operates in ten countries: the Philippines, India, the United States, Nicaragua, Vietnam, the United Kingdom, Singapore, South Africa, UAE, and Canada.

