FORDHAM Gabelli School of Business

pwc

AI Deal & Company Intelligence Agent

Automating M&A Monitoring for Deal Professionals

A fully automated RAG pipeline monitoring 51 public M&A sources, extracting structured deal records via GPT-4o, and surfacing actionable intelligence through a purpose-built dashboard — live in production today.

Team: Nathanael Lara · Lepun Lamia · Alban Cotaj · Hitansh Nagdev

679	339	$4.7T	51
Deal Records in DB	High Confidence	Tracked Deal Value	Live Data Sources

01

The Problem

Deal professionals lose 2–4 hours daily to fragmented, delayed, unstructured intelligence

Fragmented

Signal is scattered across 50+ sources — SEC filings, press wires, law firm tombstones, PE announcements — with no unified view.

Delayed

By the time a deal hits mainstream financial news, competitors have already positioned. EDGAR filings arrive 24–72 hours before press coverage.

Unstructured

Raw 8-K filings and press releases require manual interpretation to extract deal parti terms, values, and stage.

The Opportunity

$1.5T
U.S. M&A (2024)

2–4 hrs
/day per analyst

$37K–$75K
recovered / yr

Who feels it most:

Corp Dev Teams

Tracking acquisition targets and competitive moves across industries

Investment Bankers

Building deal comps and industry trackers for pitch books

Private Equity Analysts

Monitoring sector deal flow and portfolio-adjacent transactions

02

Stakeholders & User Personas

Five distinct user types — each with a distinct intelligence need

01 Corp Dev Executive

Evaluates strategic acquisitions

Pain

Manually tracks 20+ targets across news and filings

Gets

Morning brief + deal stage tracking by company

02 M&A Associate

Builds deal comps and pitch books

Pain

Hours building transaction comparables from scratch

Gets

Deal table with XBRL financials + EV multiples

03 PE Analyst

Monitors sector deal flow

Pain

No unified view of buyout and add-on activity

Gets

Sector analysis + acquirer profile pages

04 Corp Strategy / CX

Monitors competitive landscape

Pain

Reactive — hears about deals via news, not filings

Gets

Configurable source feed + AI one-pagers on demand

05 Finance / Risk

Assesses deal implications and counterparty exposure

Pain

No structured data layer for deal terms or risk

Gets

SQLite records with confidence scores + XBRL financials

03

Solution Overview

End-to-end automated intelligence pipeline with live production dashboard

51 Sources

SEC · Press Wires
News RSS · Law Firms
PE Firms · AI Search

→

Agentic Crawler

LLM-ranked source selection · 6 parsers
URL deduplication

→

RAG Pipeline

Chroma vector store
GPT-4o extraction
15-field DealRecord

→

Financial Enrichment

SEC XBRL · Tavily
EV multiples · DCF
Monte Carlo

→

Live Dashboard

Next.js + FastAPI
6 pages · 10 endpoints
Vercel + Railway

Competitive Positioning

Feature	News Aggregator	Bloomberg Terminal	Our System
51-source M&A coverage	✗	✓	✓
Structured deal extraction	✗	Manual	✓ Automated
Direct SEC EDGAR integration	✗	✓ Paid	✓ Direct API
Financial enrichment (XBRL)	✗	✓ Paid	✓ Free
M&A lifecycle stage tracking	✗	Manual	✓ Automated
Annual cost	$0–$$	$24K+ / seat	API costs only

04

Data Sources

51 sources across 7 categories — from live SEC filings to AI-powered web search

10 SEC EDGAR

8-K, S-4, SC TO-T
10-K, Form 4
DEFM14A, SC 13G

Custom EDGAR API

Highest yield
Minutes post-filing

7 Press Wires

PR Newswire
BusinessWire
GlobeNewswire

RSS + Full Text

First official
announcement

8 News RSS

CNBC · Reuters
Nasdaq · MarketWatch
Seeking Alpha

RSS + Full Text

Broad coverage;
paywall on some

5 Law Firms

Kirkland · Skadden
Latham · Weil
Paul Weiss

HTML Scraping

Tombstones confirm
closure + advisors

7 PE Firms

Blackstone · KKR
Apollo · Carlyle
TPG · Warburg

HTML Scraping

Portfolio deal
announcements

5 IB Advisors

Evercore · Lazard
Moelis · Houlihan
PJT Partners

HTML Scraping

Completed deal
tombstones

9 AI Search

Tavily (4)
Serper (3)
NewsAPI (3)

REST API

Best for private
deal coverage

Data types:

- Structured — SEC filings — CIK, form type, filing date, XBRL financial facts
- Semi-structured — RSS feeds — title, URL, description, publication date per item
- Unstructured — HTML tombstones, full-text press releases, AI search article text

05

MVP — Pipeline Architecture

Four-stage RAG pipeline with LLM-orchestrated agentic source layer

01 Ingestion

- GenericSpider dispatches 51 sources across 6 parser types
- EDGAR: targeted Item 1.01/2.01 parsing for deal terms
- Per-item doc output: URL, content, CIK, form type
- URL-level deduplication prevents reprocessing
- Agentic layer ranks sources by historical yield score

02 Embedding

- RecursiveCharacterTextSplitter: 1,000 chars / 200 overlap
- all-MiniLM-L6-v2 local HuggingFace embeddings
- Chroma vector store with MD5-hash chunk deduplication
- pub_date metadata for temporal filtering
- Incremental updates — persists across runs

03 Extraction

- GPT-4o extracts 15-field DealRecord per document
- 5 source-type-specific prompt templates
- Deterministic confidence scoring formula
- _normalize_entity() strips legal suffixes for dedup
- Keyword pre-filter + exponential backoff retry

04 Enrichment

- SEC XBRL: revenue, EBITDA, assets, debt for both parties
- EV/Revenue + EV/Net Income multiples computed
- Tavily web search fills values for private deals
- S-4/SC TO-T CIK routing → acquirer (not target)
- APScheduler daily automation at 7am UTC

06

MVP — What the Data Shows

Key findings from 679 deals indexed live across 51 sources

FORDHAM Gabelli School of Business

pwc

EDGAR Leads by 24–72 Hours

8-K Item 1.01/2.01 parsing captures deal terms minutes after filing — before mainstream financial news picks them up.

Tavily Wins on Private Deals

Full article text from Tavily AI search dramatically outperforms Serper and NewsAPI for non-public company coverage.

Confidence Formula Works

EDGAR-sourced deals score +10 pts over news articles; both parties named adds +15 pts. 50% of all 679 deals hit high-confidence threshold.

Law Firm Tombstones = Closure Signal

Law firm pages post completed deals weeks before press wires — useful for confirming deal closure and identifying advisor networks.

Live Database — Spring 2026

679	Total Deal Records
339	High Confidence (50%)
438	Announced
140	Completed
94	Pending Regulatory
7	Terminated
293	With USD Deal Value
$4.7T	Total Tracked Value

Top Sectors

07

MVP — Demo Walkthrough

FORDHAM Gabelli School of Business

pwc

1 Home Dashboard

Morning brief · 4-metric stats bar · Deal table with stage badges · Sidebar pipeline status

2 Deal Table & Filters

Sort by USD value · Filter by stage / sector / type · Completeness indicators · Acquirer links

3 Deal Detail Page

Deal terms · XBRL financials · EV multiples · 6-stage M&A progress bar · AI one-pager

4 Market Analysis

3-yr indexed stock chart vs S&P 500 · Historical XBRL quarterly financials · Bear/base/bull pro-forma

5 DCF + Monte Carlo

WACC-based DCF valuation · 1,000-trial EBITDA distribution · P10–P90 percentiles · Formula panels

6 Analytics & Sectors

Monthly deal volume · Sector card grid · Acquirer profiles · Deal comparison · URL-shareable

08

MVP — Performance & Validation

Production metrics and accuracy benchmarks

System Performance

Sources crawled per run	51
Avg crawl + extraction runtime	8–12 min
Total records in production DB	679
High-confidence deals (≥ 0.7)	339 (50%)
With USD deal value populated	293 (43%)
XBRL enrichment success rate	~70% of public US deals
Duplicate suppression	MD5 + entity normalisation
Pipeline cadence	Daily 7am UTC · APScheduler

Validation Approach

96% accuracy

25 EDGAR 8-K extractions spot-checked vs. source filings — acquirer/target name accuracy

Within 5%

10 deal values cross-referenced against original press release announcements

Stage confirmed

Stage inference validated against known M&A lifecycle (e.g. SC TO-T → pending regulatory)

4 analyst queries

RAG retrieval tested with diverse questions; cited sources matched expected filings

Known Gaps

· No labelled eval set — confidence scores are heuristic, not trained
· Accuracy drops for multi-party / consortium deal structures

09

Business Impact & Value Proposition

Quantified time savings, speed advantages, and analytical depth gains

$37K–$75K

per analyst / year

2–4 hrs/day recovered at $150K all-in salary. One seat pays back deployment cost in weeks.

24–72 hrs

ahead of mainstream news

EDGAR 8-K parsing surfaces deal terms minutes after filing vs. next-day financial press.

Zero

Bloomberg subscription needed

SEC XBRL provides target + acquirer financials, EV multiples, and quarterly history — free.

~10 seconds

to generate morning brief

LLM-generated 7-day market summary ready before team arrives. No manual synthesis required.

Scalability Advantages

- Add a new source = one row in target_sources.csv — no code changes required
- Agentic orchestration auto-deprioritizes low-yield sources; increases fetch frequency for high-signal ones
- Railway + Vercel deployment — zero infrastructure management; live 24/7 with daily auto-refresh
- Searchable, auditable record of market activity — shared intelligence baseline across the entire team

10

Limitations & Risks

Known gaps, model risks, adoption barriers, and compliance considerations

Data Gaps

· Private co deals: ~57% null deal values (Tavily partially mitigates)

· Non-US + pre-IPO companies: no XBRL financial enrichment

· MarketWatch / TheStreet: paywalled — RSS snippets only

Model & Extraction Risk

· GPT-4o may infer incorrect parties from ambiguous text (hallucination)

· Multi-party / consortium deals degrade extraction accuracy

· _normalize_entity() edge case: two 'Summit' companies → same hash

Adoption Challenges

· Requires API keys: OpenAI, Tavily, Serper, NewsAPI — not zero-cost

· Non-technical users need CSV editing to change source configuration

· Brief quality degrades during OpenAI service outages

Ethical & Compliance

· Scraping PE / law firm sites may violate ToS — legal review pre-commercial

· LLM projections (DCF, Monte Carlo) must be labelled as estimates only

· GPT-4o training skews US/English — non-US deal accuracy is lower

Mitigations in place:

Confidence scoring downgrades on missing parties · Tavily web-value enrichment for private deals · Lightweight DCF fallback for no-XBRL companies · Disclaimer banners on all projection panels · MD5 + entity-normalisation deduplication across all runs

11

Future Enhancements & Roadmap

Three horizons — from immediate wins to strategic platform expansion

Near-Term · 0–3 months

- Watchlist & Alerts — POST /api/watchlist with SMTP / Slack notifications
- Natural language query UI — chat interface over the RAG chain
- Better CIK resolution — SEC company search API replaces fuzzy matching
- Ticker resolution fix — subsidiaries, renamed, non-US companies

Medium-Term · 3–6 months

- CRM integration — push deal records to Salesforce / HubSpot via REST
- Document upload — ingest user NDAs, CIMs, pitch books alongside public sources
- International expansion — LSE RNS, Euronext, HKEX + non-English RSS parsers
- Real-time streaming — RSS polling every 15 min, EDGAR event-driven ingestion

Longer-Term · 6–12 months

- Predictive deal scoring — train classifier on historical M&A outcomes
- Advisor graph — map law firm / IB networks across deals over time
- Multi-modal analysis — parse charts and tables from investor PDFs
- White-label SaaS — multi-tenant deployment with per-client source config

Conclusions & Takeaway

The Problem

Deal professionals waste 2–4 hrs/day scanning 50+ fragmented sources for M&A intelligence that arrives too late and requires manual interpretation.

The Solution

An automated pipeline monitoring 51 sources, extracting 15-field deal records via GPT-4o, enriching with XBRL financials, and surfacing insights through a live dashboard.

Why It Matters

EDGAR-speed intelligence (minutes after filing) vs. news-cycle speed (24–72 hours). Bloomberg-grade financials at API cost. 679 deals indexed and growing daily.

How It's Different

Not a news aggregator — a structured intelligence layer. Not static — an agentic system that learns source yield. Not a prototype — live on Railway + Vercel today.

The Ask

Production pilot with a corp dev or PE team — replace one analyst's morning research routine with the AI morning brief for 30 days and measure time saved. Adding a new source requires only a single row in a CSV file; no engineering changes needed.