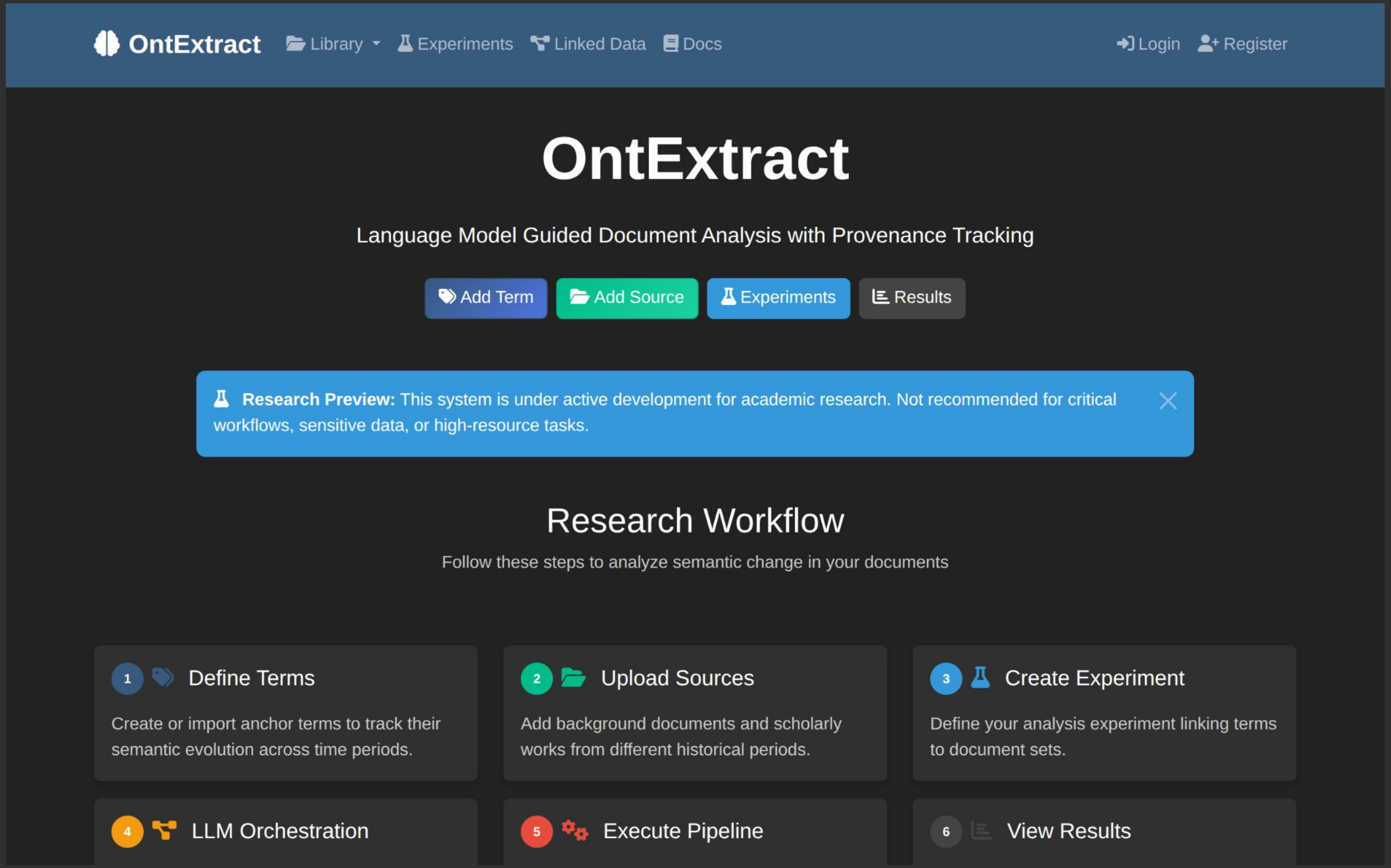
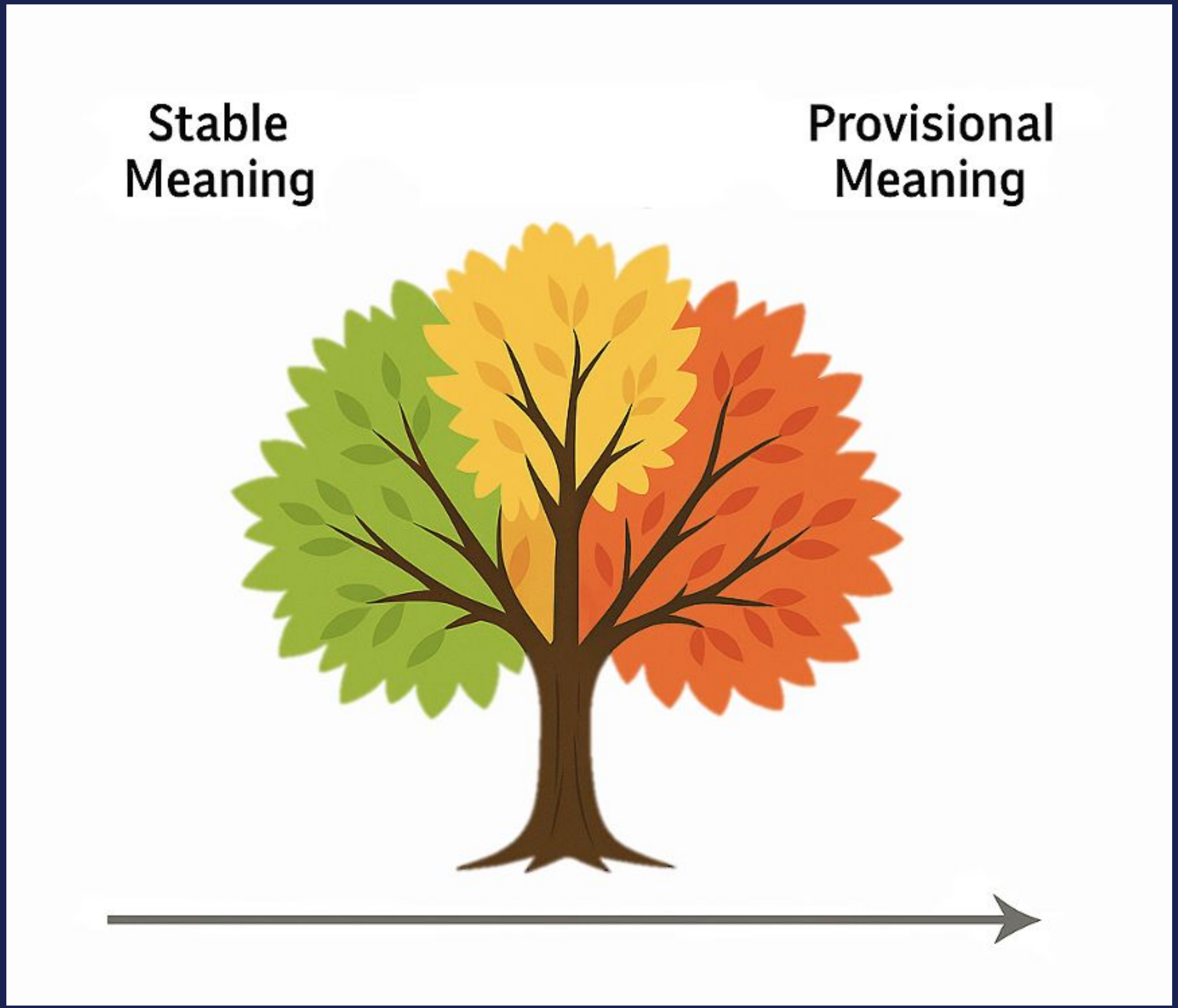
OntExtract: PROV-O Provenance Tracking for Document Analysis Workflows



ACM/IEEE Joint Conference on Digital Libraries (JCDL 2025)
Christopher B. Rauch, Hyung Wook Choi, Mat Kelly
Drexel University, Philadelphia, PA

How to track semantic changes in terminology over time?



- Language is in constant flux
- Words evolve continuously, reflecting shifts in culture, science, and technology
- When terminology shifts, it can disrupt communication, hindering interdisciplinary knowledge integration
- Meanings rarely change overnight, they fade, blend, and overlap
- Old and new meanings often coexist
- Research in semantic drift often requires quantitative analysis as well as qualitative interpretation
- In future work, ontological structure will represent the relations between the different meanings

Tracking "Agent" Across 114 Years

Documents (7) — Document Pipeline

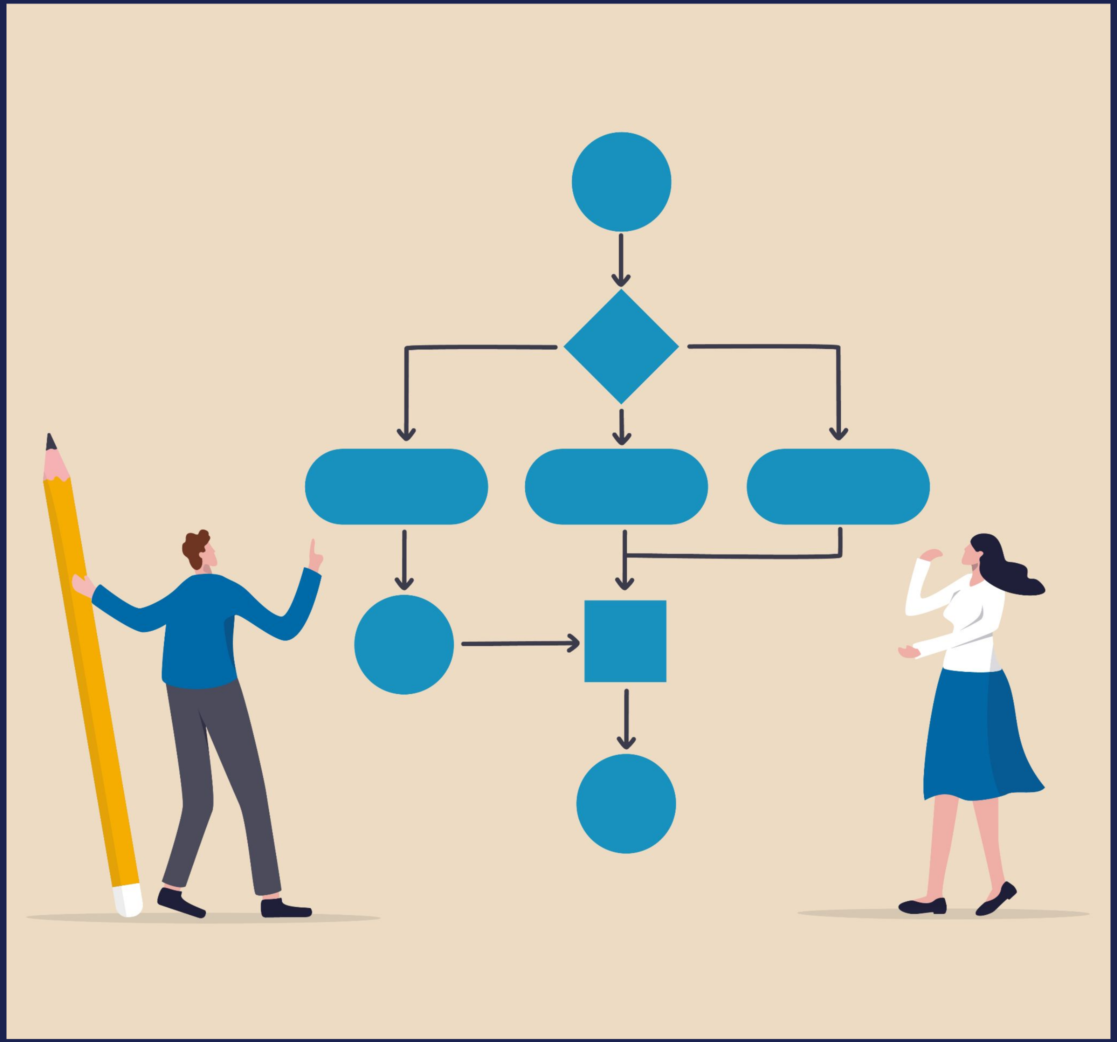
Document	Date	Processing	Actions
agent, n.1 & adj. Oxford English Dictionary v2	2024	embeddings: period_aware entities: spacy_ner temporal: temporal_extraction definitions: definition_extraction	
Agent (Black's Law Dictionary, 12th ed.) Bryan A. Garner v2	2024	embeddings: period_aware entities: spacy_ner definitions: definition_extraction	
Intelligent Agents Stuart Russell, Peter Norvig v2	2022	embeddings: period_aware entities: spacy_ner definitions: definition_extraction	
Agent (Black's Law Dictionary, 11th ed.) Bryan A. Garner v2	2019	embeddings: period_aware entities: spacy_ner definitions: definition_extraction	
Intelligent Agents: Theory and Practice Michael Wooldridge, Nicholas R. Jennings v2	1995	embeddings: period_aware temporal: temporal_extraction definitions: definition_extraction	
Intention G.E.M. Anscombe v2	1956	embeddings: period_aware entities: spacy_ner definitions: definition_extraction	
Agent (Black's Law Dictionary, 1st ed.) Henry Campbell Black v2	1910	segmentation: paragraph embeddings: period_aware entities: spacy_ner definitions: definition_extraction	

The demonstration uses seven documents spanning from 1910 to 2024.

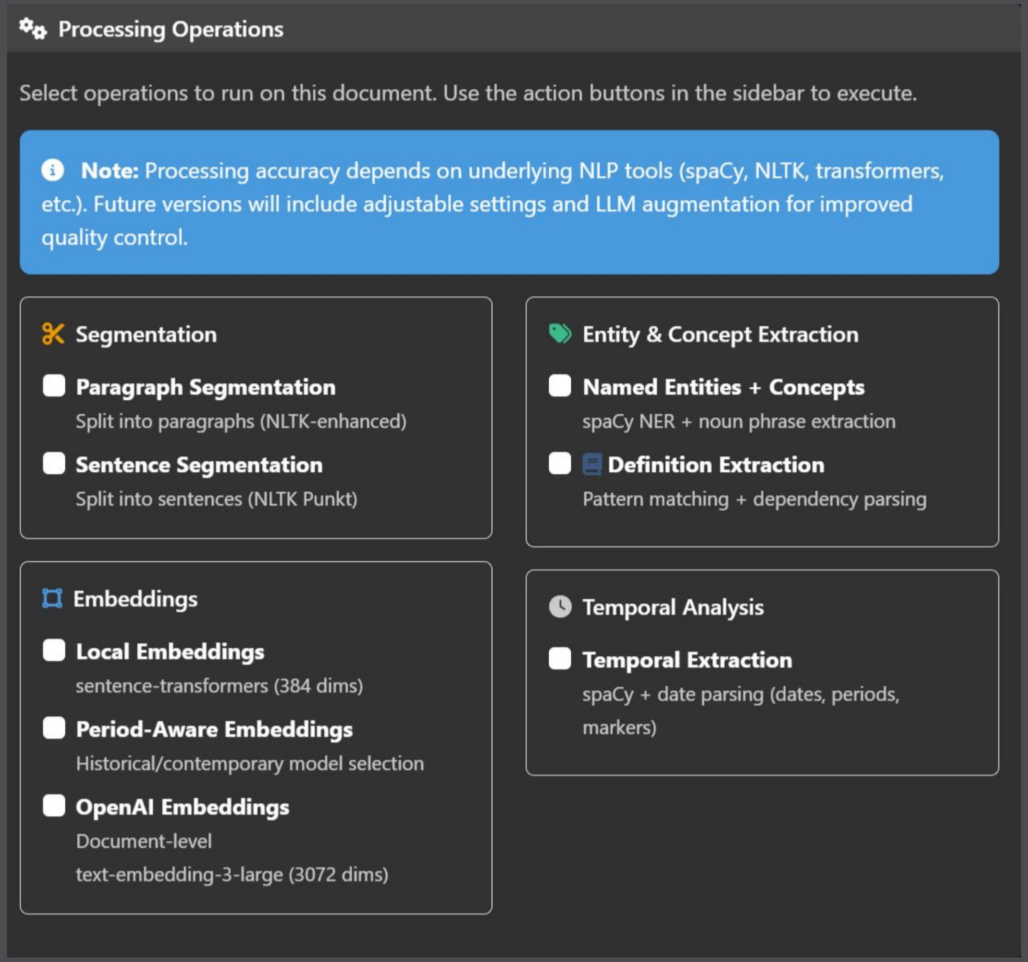
1. **1910**
   Black's Law Dictionary, first edition, representing legal usage
2. **1956**
   Anscombe's "Intention," representing philosophical discourse
3. **1995-2020**
   Wooldridge and Jennings along with Russell and Norvig, representing the computational turn
4. **2019-2024**
   Recent editions of Black's Law Dictionary and the OED, representing contemporary usage

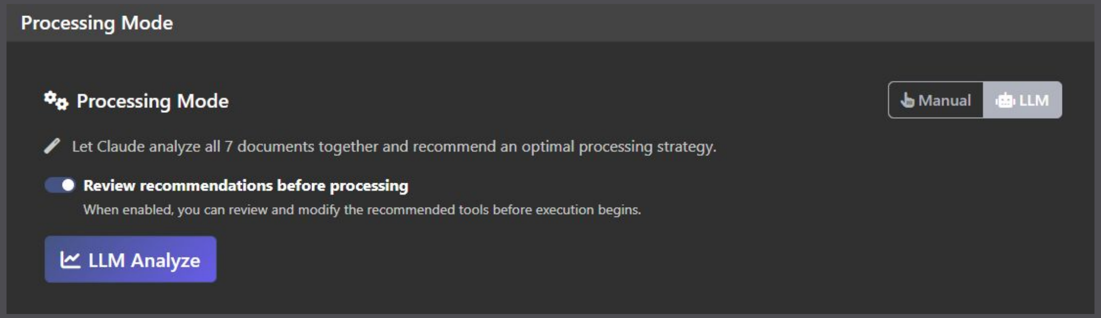
The corpus totals **58,522 words** across legal, philosophical, and computational domains.

The Reproducibility Challenge



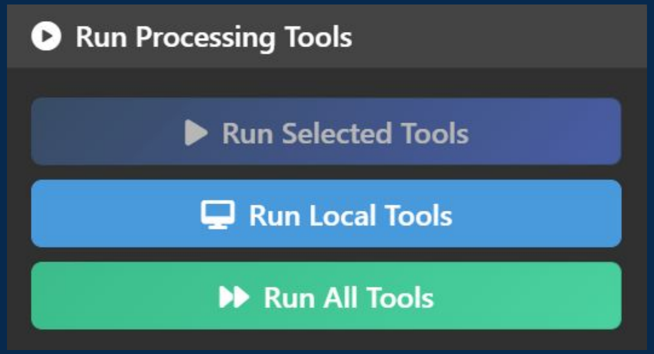
- NLP workflows combine multiple tools with varying parameters
- Results depend on tool versions, processing order, and threshold settings
- Traditional logging captures events but not relationships between artifacts
- Researchers cannot trace analytical conclusions back to source operations
- When results differ, identifying the point of divergence requires manual investigation





OntExtract: A Unified Interface with Embedded Provenance Tracking

- Single web interface for document processing
- Integrates spaCy, NLTK, and sentence-transformers
- PostgreSQL with pgvector for semantic similarity search.



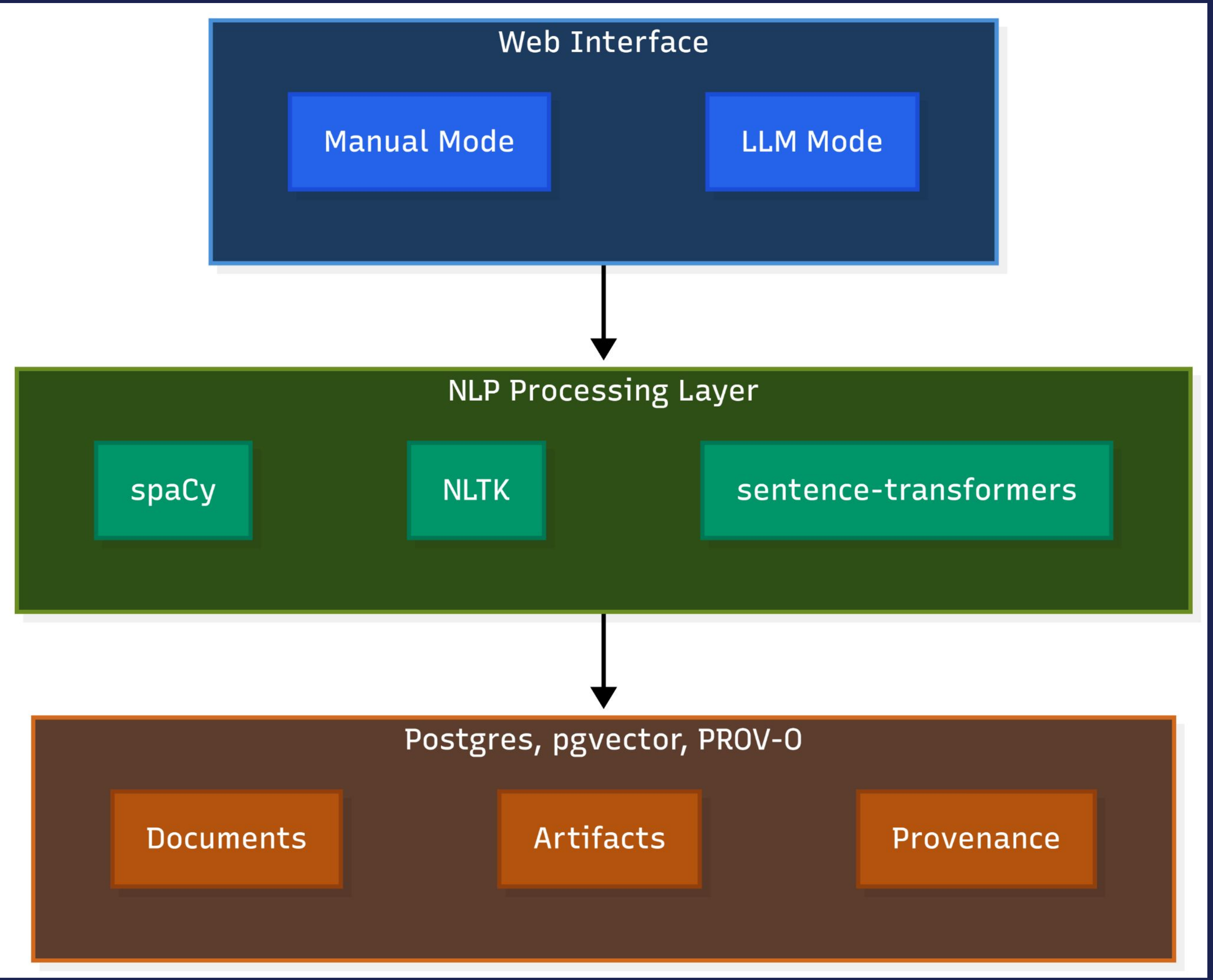
Standalone mode allows direct tool selection without external API dependencies.

API-enhanced mode adds LLM orchestration for automated strategy recommendations with human review before execution.

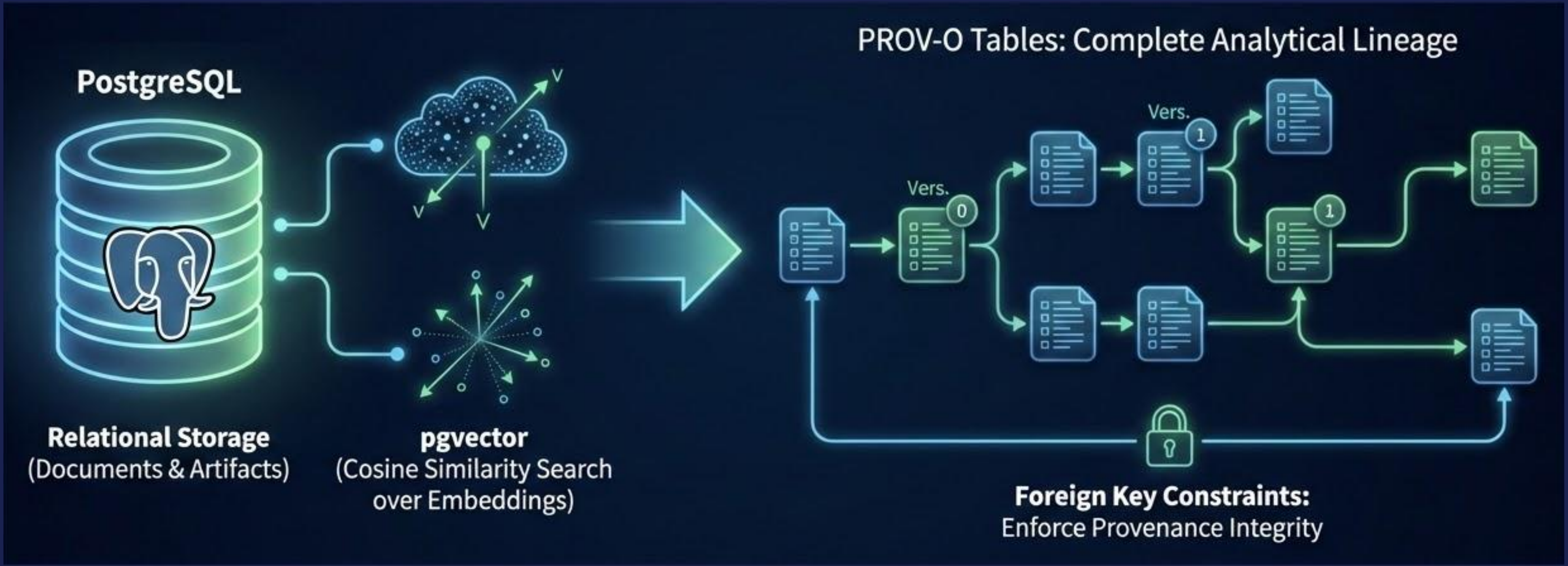
PROV-O provenance embedded database schema. Each processing operation creates a versioned output with corresponding provenance record.

System Architecture: Core Components and Workflow

- User interface provides access to both operating modes
- Mode selection determines strategy generation, not processing
- NLP processing layer handles all document operations
- PostgreSQL backend stores documents, artifacts, and provenance
- pgvector extension enables semantic similarity queries



The Database Backend: PostgreSQL with PROV-O Tables



Same Schema Used Regardless of Operating Mode

Entities Extraction

Activity prov:**wasAssociatedWith** researcher:1

Details:

- model_info: en_core_web_sm + noun phrase extraction
- document_id: 403
- entity_types: ['TIME', 'ORG', 'CONCEPT', 'GPE', 'QUANTITY', 'NORP', 'EVENT', 'F 'WORK_OF_ART', 'CARDINAL']
- service_used: spaCy NLP + Enhanced Extraction
- experiment_id: 57
- avg_confidence: 0.7077461205800051
- entities_found: 3931
- document_version: 2
- extraction_method: spacy
- processing_method: spacy

Entities prov:**wasGeneratedBy** Implementations in Machine Ethics

prov:**wasDerivedFrom** document

Duration: 0.0s

Embeddings

Activity prov:**wasAssociatedWith** researcher:1

Details:

- note: Hierarchical: 1 document + 1343 segment embeddings
- dimensions: 768
- model_used: all-mpnet-base-v2
- document_id: 403
- total_tokens: N/A
- experiment_id: 57
- document_version: 2
- embedding_method: period_aware
- total_embeddings: 1344
- processing_method: period_aware
- segment_embeddings: 1343
- document_embeddings: 1

Entities prov:**wasGeneratedBy** Implementations in Machine Ethics

prov:**wasDerivedFrom** document

Duration: 0.0s

PROV-O Implementation: Modeling Analytical Lineage

The W3C PROV-O standard provides the vocabulary for provenance tracking.

Three core concepts structure the model.

Entities

Artifacts such as documents, text segments, and extracted terms.

Activities

Operations such as upload, segmentation, and entity extraction

Agents

Actors including human researchers and NLP tools with their specific version numbers.

PROV-O improves chronological logging with relational provenance.

Upcoming feature: Researchers can query relationships between artifacts, operations, and tools.

Mapping Workflow Steps to Provenance Records

Each workflow step maps to formal PROV-O relationships.

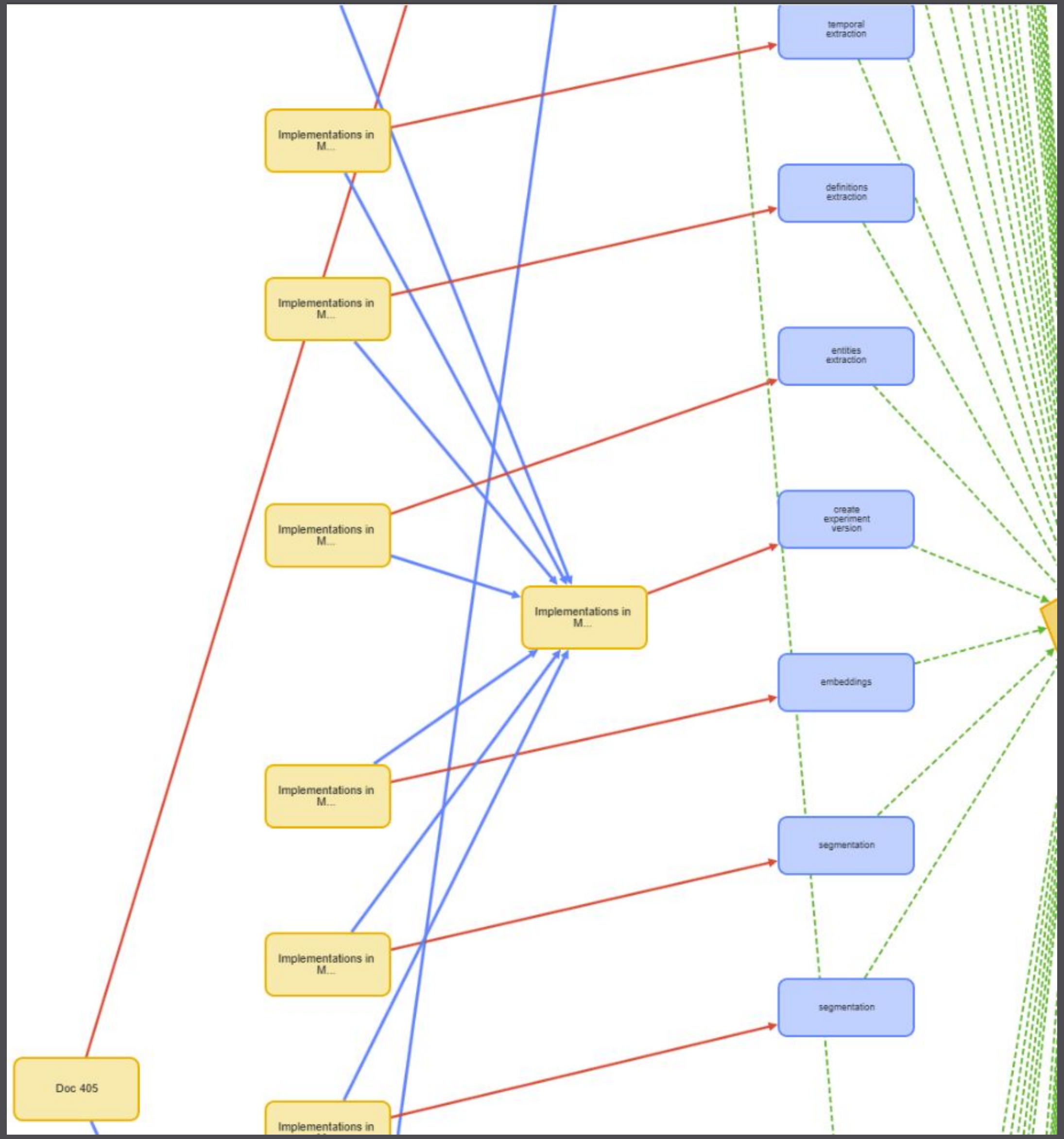
PROV-O Relationships

wasDerivedFrom — Links artifacts to source documents

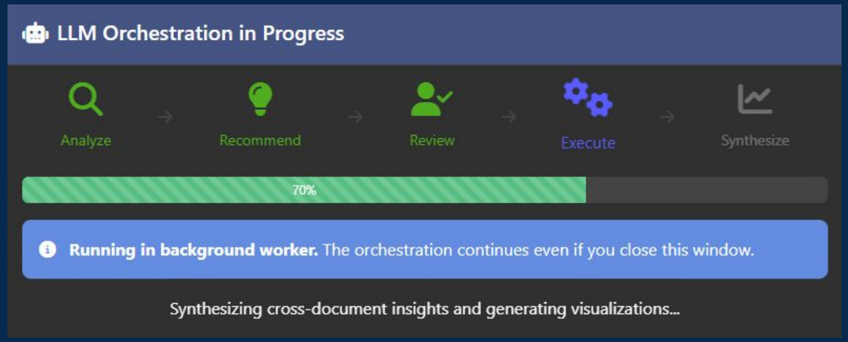
wasGeneratedBy — Links artifacts to creating activities

used — Links activities to consumed entities

wasAssociatedWith — Links activities to responsible agents



LLM Orchestration: A Five-Stage Pipeline



The API-enhanced mode implements a five-stage workflow built on LangGraph for state management.

1

Analyze

The LLM examines document characteristics and research goals

2

Recommend

The LLM suggests processing tools with confidence scores and explicit rationale

3

Human Review

The researcher examines recommendations and decides whether to approve, modify, or reject the proposed strategy.

4

Execute

Approved operations run using local NLP tools. No LLM involvement occurs during actual data processing (deterministic results) except for post OCR-Text Cleanup if selected. Future integrations for LLM for quality review planned.

5

Synthesize

The LLM organizes extraction results and identifies patterns. Processing products are available for export. Interpretation remains with the researcher

All decisions, including rejections, are captured in provenance records.

Live Demonstration

Demo available at https://ontextract.ontorealm.net

Credentials: Create account for instant access or

`demo` / `demo123`

1

Experiment Overview

Displays seven documents with metadata and temporal period assignments

2

Document Pipeline

Shows processing status for each document with available tools and completed operations

3

Results Review

Displays extracted entities, definitions, temporal expressions, segments, and embeddings

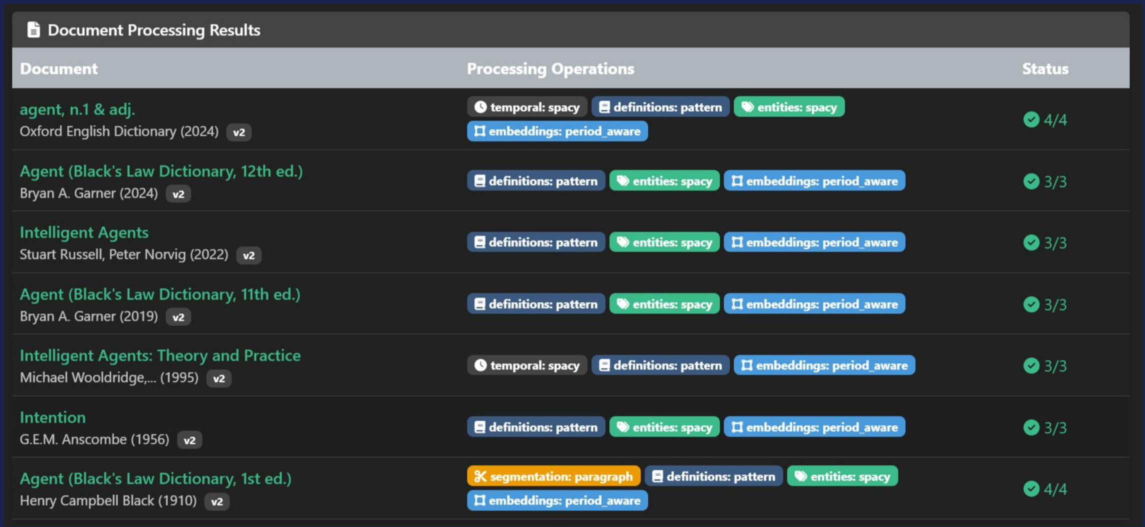
4

Provenance Graph

Visualizes PROV-O relationships between documents, activities, and derived artifacts

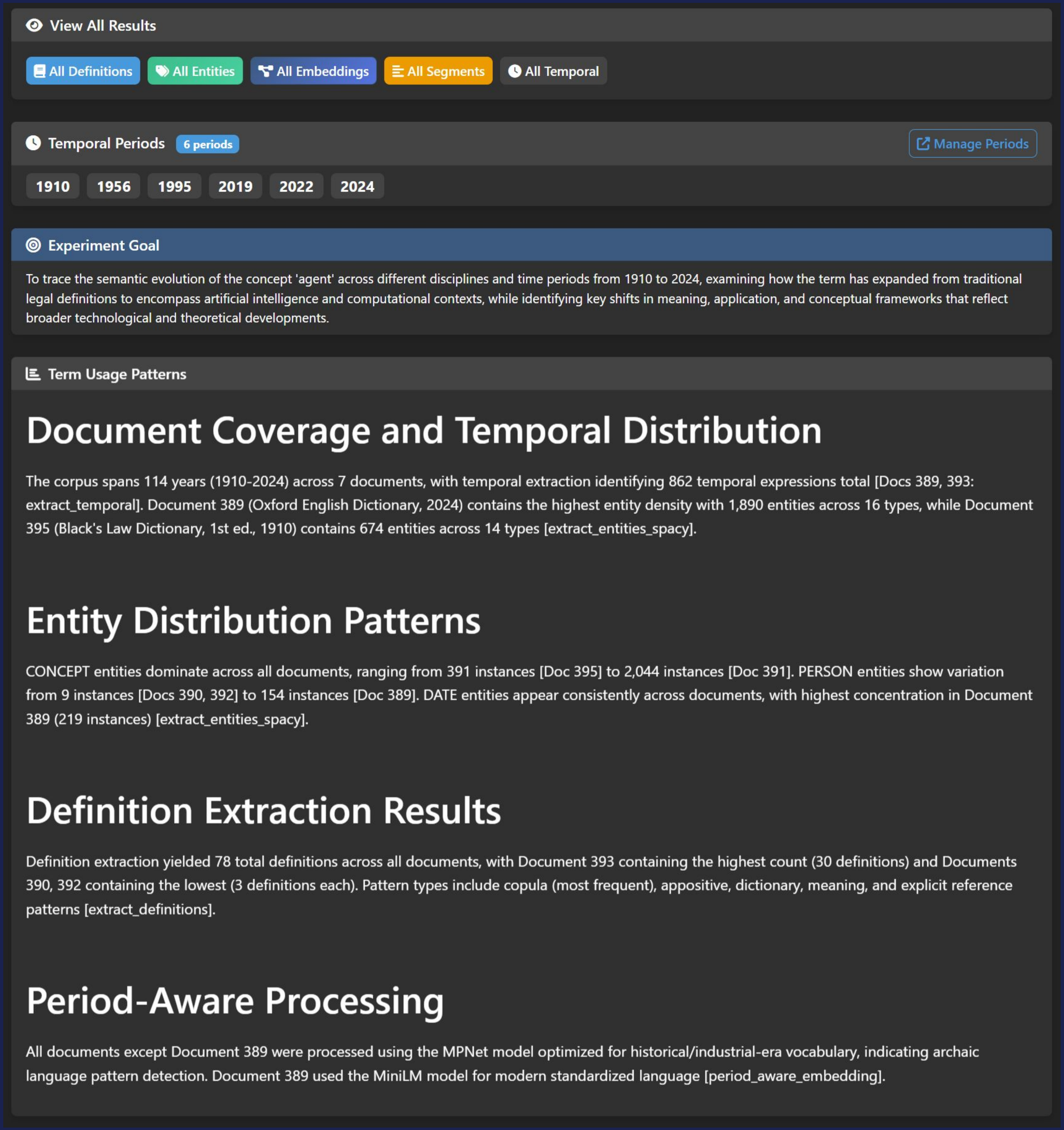
Synthesis Output: Patterns for Researcher Evaluation

- Organizes results by document date and source type
- Summarizes how entity counts and definition patterns vary across documents
- Notes processing decisions such as embedding model selection for historical vs. modern texts
- Presents structured observations, not validated findings
- Researcher evaluates whether patterns reflect genuine semantic differences or corpus artifacts
- Every observation traces back through provenance to specific extraction operations



Document Processing Results

Document	Processing Operations	Status
agent, n.1 & adj. Oxford English Dictionary (2024) v2	temporal: spacy; definitions: pattern; entities: spacy; embeddings: period_aware	4/4
Agent (Black's Law Dictionary, 12th ed.) Bryan A. Garner (2024) v2	definitions: pattern; entities: spacy; embeddings: period_aware	3/3
Intelligent Agents Stuart Russell, Peter Norvig (2022) v2	definitions: pattern; entities: spacy; embeddings: period_aware	3/3
Agent (Black's Law Dictionary, 11th ed.) Bryan A. Garner (2019) v2	definitions: pattern; entities: spacy; embeddings: period_aware	3/3
Intelligent Agents: Theory and Practice Michael Wooldridge,... (1995) v2	temporal: spacy; definitions: pattern; embeddings: period_aware	3/3
Intention G.E.M. Anscombe (1956) v2	definitions: pattern; entities: spacy; embeddings: period_aware	3/3
Agent (Black's Law Dictionary, 1st ed.) Henry Campbell Black (1910) v2	segmentation: paragraph; definitions: pattern; entities: spacy; embeddings: period_aware	4/4



View All Results

All Definitions | All Entities | All Embeddings | All Segments | All Temporal

Temporal Periods 6 periods — Manage Periods

1910 | 1956 | 1995 | 2019 | 2022 | 2024

Experiment Goal

To trace the semantic evolution of the concept 'agent' across different disciplines and time periods from 1910 to 2024, examining how the term has expanded from traditional legal definitions to encompass artificial intelligence and computational contexts, while identifying key shifts in meaning, application, and conceptual frameworks that reflect broader technological and theoretical developments.

Term Usage Patterns

Document Coverage and Temporal Distribution

The corpus spans 114 years (1910-2024) across 7 documents, with temporal extraction identifying 862 temporal expressions total [Docs 389, 393: extract_temporal]. Document 389 (Oxford English Dictionary, 2024) contains the highest entity density with 1,890 entities across 16 types, while Document 395 (Black's Law Dictionary, 1st ed., 1910) contains 674 entities across 14 types [extract_entities_spacy].

Entity Distribution Patterns

CONCEPT entities dominate across all documents, ranging from 391 instances [Doc 395] to 2,044 instances [Doc 391]. PERSON entities show variation from 9 instances [Docs 390, 392] to 154 instances [Doc 389]. DATE entities appear consistently across documents, with highest concentration in Document 389 (219 instances) [extract_entities_spacy].

Definition Extraction Results

Definition extraction yielded 78 total definitions across all documents, with Document 393 containing the highest count (30 definitions) and Documents 390, 392 containing the lowest (3 definitions each). Pattern types include copula (most frequent), appositive, dictionary, meaning, and explicit reference patterns [extract_definitions].

Period-Aware Processing

All documents except Document 389 were processed using the MPNet model optimized for historical/industrial-era vocabulary, indicating archaic language pattern detection. Document 389 used the MiniLM model for modern standardized language [period_aware_embedding].

Limitations and Future Work

Current Limitations

- Orchestration mode requires commercial LLM API (Claude or GPT)
- Standalone mode operates without external dependencies
- Prototype focuses on English-language text
- Architecture accommodates other languages with appropriate spaCy models
- Entity extraction accuracy depends on spaCy pretrained models
- Domain-specific terminology may require fine-tuned models

Future Directions

- Integration of local open-source LLMs (Llama, Mistral)
- Extension to phrase-level semantic analysis
- Validation studies comparing output against expert annotations
- User studies on provenance query utility
- LLM Quality Control
- Fine grained settings for NLP Tools

Summary

OntExtract provides a web based interface for NLP document processing operations that includes integrated provenance tracking. PROV-O concepts are embedded in the database schema for effective reproducibility. Researchers can trace any result back through the chain of operations, tools, and source materials. The dual-mode architecture supports both manual tool selection and LLM-assisted orchestration.

Live Demo

ontextract.ontorealm.net

Credentials: `demo` / `demo123`

Source Code

github.com/MatLab-Research/OntExtract

Contact

hc685@drexel.edu (wook)

cr625@drexel.edu (chris)

mrk335@drexel.edu (mat)