A Deployed Investigative AI Search Engine for Combating Human Trafficking at Web Scale

Mayank Kejriwal
Information Sciences Institute, University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292 USA
kejriwal@isi.edu

Abstract

Online human trafficking investigations generate vast amounts of noisy, heterogeneous, and deliberately obfuscated data, making traditional search and analytics tools ineffective for supporting law enforcement. This paper discusses the deployment of the Domain-Specific Insight Graphs (DIG) system, an AI-powered investigative search engine that was operationally used by over 200 U.S. law enforcement agencies for more than five years in the pre-COVID period. The system integrates advanced research conducted over the years in information extraction, knowledge graph construction, and entity-centric search to enable investigators to formulate queries without technical background, aggregate evidence, and uncover latent relationships among entities such as phone numbers, emails, and locations. Beyond technical innovation, the deployment required sustained attention to usability, explainability, and policy compliance, ensuring trust in high-stakes legal contexts. We report measurable benefits in investigative efficiency, case initiation, and prosecutorial support, as well as lessons learned from long-term maintenance and adaptation to evolving online platforms. Since 2020, work conducted in this domain has also had significant policy and advocacy ramifications. The system's generalized design has also allowed it to be prototyped for adjacent illicit domains, including securities fraud and illegal firearm sales, demonstrating the broader applicability of AI-driven investigative tools. We contribute a rare case study of an AI system that has transitioned from research to sustained real-world impact in a socially critical domain.

Introduction

Human trafficking remains a pervasive global challenge with severe humanitarian consequences. In the United States alone, thousands of cases are reported annually, and many more go undetected due to the clandestine nature of the activity (Harrendorf, Heiskanen, and Malby 2010; Savona and Stefanizzi 2007). The rapid growth of the Internet has exacerbated the problem: victims are advertised on both the surface and dark web (Chen 2011; Greiman and Bain 2013a), with estimates of hundreds of millions of online postings appearing over the past decade. These postings are often deliberately obfuscated, distributed across a long tail of domains, and designed to evade both legal oversight and conventional information retrieval methods (Hultgren et al. 2016a). Investigators face a daunting task: sifting through massive, noisy, and heterogeneous web corpora to identify potential victims, vendors, and trafficking networks. Traditional keyword search engines are ill-suited for this work (Hogan et al. 2007a; Tonon, Demartini, and Cudré-Mauroux 2012a), as they cannot reliably support clustering, aggregation, or entity-centric analysis (Lin et al. 2012a; Saleiro et al. 2016a), all of which are critical for actionable investigations. Furthermore, even with the advent of large language models (LLMs), significant data privacy and security issues preclude their use except for basic tasks. Use of customized advanced tools, including state-of-the-art research in applied artificial intelligence (AI), offers a promising route, but must be balanced with the real-world needs of investigators, and concerns such as usability and trust (Sauro 2011a).

Recognizing these challenges, our team developed an investigative search engine as part of the DARPA MEMEX program called Domain-specific Insight Graphs (DIG). The system was designed from the outset not only as a research prototype but also as a deployable application for use in high-stakes investigative contexts. At its core, the system integrates techniques from information extraction (Chang et al. 2006; Hirschman and Gaizauskas 2001), knowledge graph construction (Niu et al. 2012a; Szekely et al. 2015a), and entity-centric search (Frank et al. 2012a; Lin et al. 2012a; Saleiro et al. 2016a) to allow investigators to formulate structured queries, aggregate evidence, and explore latent relationships such as shared phone numbers or email addresses across advertisements. Unlike conventional web search, the system retains historical data, enabling investigators to establish behavioral patterns over time, which is a key requirement for building prosecutable cases (Dalvi et al. 2009; Doan, Halevy, and Ives 2012a).

Since its initial release, DIG has transitioned from research to practice, as a platform now in use by hundreds of law enforcement agencies across the United States, including local police departments, state attorneys general, and specialized task forces (Szekely et al. 2015a). The tool has been integrated into day-to-day investigative workflows, providing users with advanced search, clustering, and visualization capabilities. Reports from the field indicate that the system has been instrumental in generating leads, reducing investigative time, and supporting successful prosecutions.

Beyond human trafficking, extensions of the system have been adapted to other illicit domains such as securities fraud, illegal firearm sales, and online scams, demonstrating the generality of the underlying approach.

This paper revisits the investigative search engine several years after its initial deployment. Our focus is not on the technical details of algorithms, which have been described previously, but rather on the lessons learned from scaling, maintaining, and embedding such a system in sensitive, real-world contexts. In particular, we highlight three contributions:

1. We describe the real-world deployment trajectory of the system, including adoption by over 200 agencies, integration into investigative practices, and the measurable benefits observed in lead generation and evidence gathering.
2. We analyze key lessons learned in the course of long-term maintenance, such as the importance of data retention, explainability of clustering and search outputs, and the need for trust-building with investigators who operate in high-stakes environments.
3. We assess broader impacts and policy implications of deploying AI systems in domains with legal and ethical sensitivities, emphasizing how collaboration with policymakers and practitioners shaped the design and led to sustained use.

By documenting the deployment journey, outcomes, and limitations, we aim to contribute to the growing literature on real-world applications of AI for social good. Our experience suggests that AI technologies can provide substantial benefits in complex investigative domains, but only when equal attention is given to usability, transparency, and long-term sustainability alongside technical innovation.

Background and Related Work

Research on applying computational methods to combat human trafficking is still relatively limited compared to other areas of cyber-enabled crime. Early efforts have focused on applying machine learning and natural language processing to detect trafficking-related content online. For example, classifiers have been trained on escort advertisements to distinguish trafficking-related posts from benign ones (Alvari, Shakarian, and Snyder 2016), while other studies have highlighted the role of online platforms as gateways for trafficking activity (Greiman and Bain 2013b). Parallel work has investigated knowledge management and information systems as aids for investigators, demonstrating that systematic approaches can provide important context in identifying trafficking cases (Hultgren et al. 2016b). They suggest the potential of data-driven methods, but often remain at the level of research prototypes without evidence of longstanding deployment.

A second stream of work relates to information extraction and entity resolution. The problem of identifying and linking entities across heterogeneous and noisy data sources has been extensively studied in computer science, including duplicate detection (Elmagarmid, Ipeirotis, and Verykios 2007), record linkage (Doan, Halevy, and Ives 2012b), and entity-centric retrieval (Lin et al. 2012b; Saleiro et al. 2016b). Within the MEMEX program, several systems demonstrated the feasibility of large-scale entity extraction and integration over web corpora, laying the groundwork for operational knowledge graphs (Szekely et al. 2015b). Similarly, advances in deep learning for knowledge-base construction (Niu et al. 2012b) and entity-centric test collections (Frank et al. 2012b) have informed the design of practical investigative tools. While these methods provide the technical underpinnings, our system distinguishes itself by emphasizing explainability and provenance, which are features essential for adoption in legal contexts.

Finally, the work intersects with research on specialized search and exploratory interfaces. Traditional web search engines are not well suited for investigative needs, which often require clustering, temporal aggregation, and the ability to trace entities across multiple postings. Research on scalable search engines (Hogan et al. 2007b), structured querying over text databases (Jain, Doan, and Gravano 2007), and hybrid retrieval methods combining inverted indices with structured queries (Tonon, Demartini, and Cudré-Mauroux 2012b) highlight strategies for supporting more sophisticated information access. At the same time, usability studies demonstrate that interface design can significantly affect investigator efficiency (Sauro 2011b). Our contribution builds on these insights by providing an entity-centric search engine specifically adapted to the workflows of human trafficking investigators, and by documenting its sustained use and measurable impacts in real-world deployments.

System Overview

The DIG search engine is designed as a full-stack system that transforms raw, noisy web data into actionable investigative insights. At its core, the system integrates knowledge graph construction, specialized information extraction, and a query reformulation engine that enables investigators to pose complex questions in intuitive interfaces without requiring technical expertise in graph query languages (Figure 1). The system was developed under the DARPA MEMEX program and subsequently transitioned to operational use, where it served as a backbone for human trafficking investigations among a handful of other systems.

Knowledge Graph (KG) Construction and Representation

The starting point for the system is the acquisition of a large corpus of potentially relevant webpages through focused crawling. These crawlers rely on seed URLs and domain-specific keywords (e.g., “escort,” “incall”) to identify sites of interest, producing corpora that can range from millions to hundreds of millions of advertisements (ads). From this corpus, the system constructs a *domain-specific knowledge graph (KG)* organized around a shallow investigative schema. This schema, inspired by schema.org, specifies classes and properties of interest to investigators, including phone numbers, email addresses, locations, services, and physical descriptors. Unlike general-purpose ontologies,

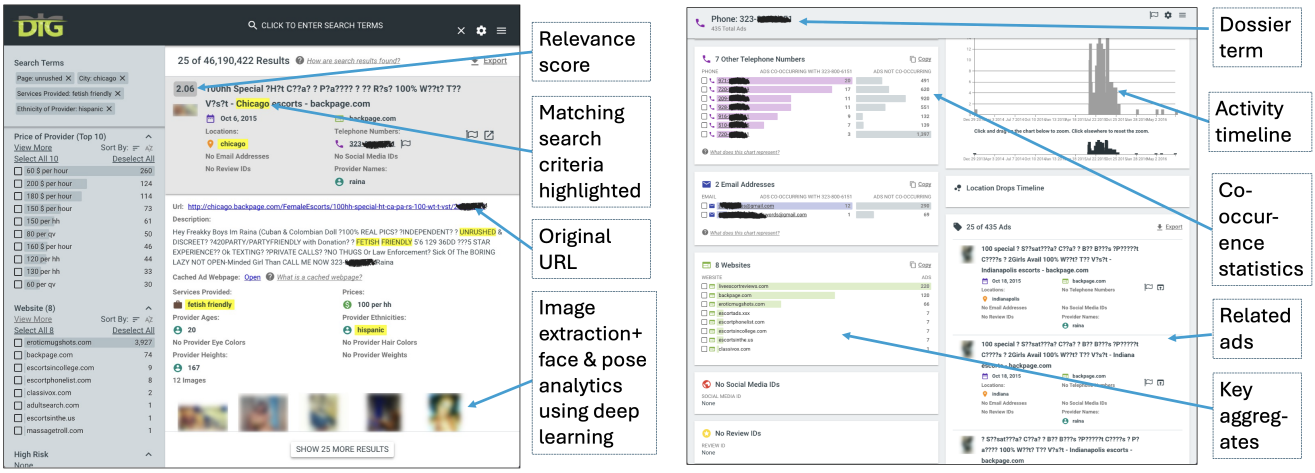

Figure 1: The faceted search interface in DIG (left) and an entity-centric interface (right), along with key features, describing a complete profile of a phone number, including co-occurrence with other phone numbers and email addresses, and the timestamps associated with ads from which that phone was extracted. Sensitive information has been obfuscated. Ads from which that phone were extracted are also suggested in the lower right pane.

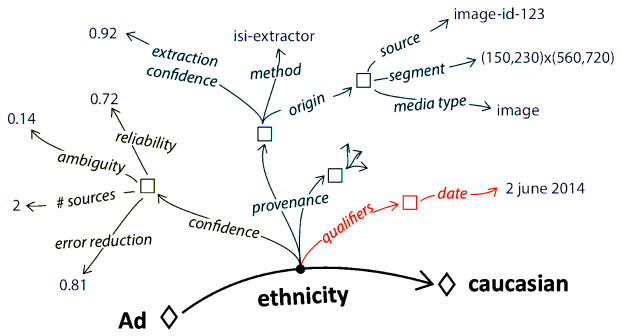

Figure 2: An illustration of the knowledge graph representation used in DIG, with active support for both data and meta-data elements, including confidence estimates of predicted outputs and provenance.

the schema was deliberately restricted in scope to capture investigative priorities identified through consultations with practitioners.

Populating the knowledge graph requires combining multiple specialized information extraction (IE) approaches (Kejriwal and Szekely 2017a; Kejriwal, Shao, and Szekely 2019). Structured fields such as phone numbers and email addresses are reliably captured using regular expressions, while more ambiguous attributes (e.g., ethnicity, hair color, street addresses) rely on conditional random fields, lexicons, and hybrid rule-based methods. External resources such as Geonames provide grounding for geographic mentions (Ahlers 2013; Kejriwal and Szekely 2017c). Importantly, the design balances precision and recall: some extractors are tuned for high accuracy, while others prioritize broad coverage, allowing investigators to trade off between completeness and certainty. To ensure transparency, each extracted fact is linked back to its provenance (the specific webpage and algorithm responsible for the extraction) so that users can always verify information against the original source.

In addition to explicit entities, the system infers latent structures that are critical for investigations. A key example is vendor discovery: by clustering ads that share phone numbers or emails, the system can infer groups of escorts likely controlled by the same organizer. A random walk–based clustering algorithm is employed to build these clusters, chosen for both scalability and interpretability. The result is a KG that not only captures isolated attributes but also models relationships among entities at scale. It also has a rich representation grounded in both data and meta-data (Figure 2), and with full support for confidence estimates produced by the various machine learning models, and provenance. When deployed, DIG included more than 100 million sex-work advertisements, three years of running data coverage, and over two billion extracted facts in its KG.

Query Reformulation and Search

Traditional keyword search engines cannot accommodate the complex queries required in investigative contexts, such as clustering ads by shared contact information or aggregating prices across a network of related postings. To address this need, the system provides a query reformulation engine that bridges investigator inputs to the underlying KG.

Investigators typically interact with the system through forms or natural constraints, which are internally represented in a SPARQL-inspired language. To increase robustness, each structured query is automatically reformulated into a set of soft boolean tree (SBT) queries executable over an Elasticsearch backend (Kejriwal and Szekely 2017b). SBT queries allow constraints to be weighted and combined flexibly, making search tolerant to missing or

noisy extractions. For example, a request for "escorts with brunette hair using phone number 123-456-7890" would be brittle if executed literally, since extraction errors or synonymy ("brunette" vs. "brown") could prevent results from being returned. Through reformulation strategies, such as relaxing constraints, mapping synonyms, and incorporating keyword-based search across textual fields, the system ensures that partial matches are retrieved and ranked by relevance. These reformulations significantly improve recall without overwhelming investigators with spurious results as we demonstrate through an extensive set of experiments (Kejriwal, Szekely, and Knoblock 2018) (see also *Experimental Validation*).

User Interface and Investigative Workflow

The query engine powers an interactive graphical user interface (GUI) designed in consultation with investigators. The GUI supports both lead investigation and exploratory workflows. Users can begin with a simple query, such as a phone number, and progressively refine results using facets for attributes like price, location, ethnicity, or website domain. Each entity (e.g., a phone number or email) has a dedicated profile page that displays associated ads, co-occurring identifiers, temporal activity patterns, and geospatial distributions (Figure 1 *right*). This entity-centric design reflects investigative reasoning, where a lead is iteratively expanded into a broader network of related evidence.

A distinctive feature is the retention of cached webpages and images, even after they disappear from the live web. This allows investigators to view original sources, establish historical patterns, and preserve evidence for legal proceedings. The system also integrates image similarity search and timeline visualizations, enabling cross-modal exploration of cases. Importantly, every displayed result can be traced back to its provenance, ensuring accountability and supporting courtroom admissibility. This feature, which was incorporated after extensive consultation with stakeholders such as law enforcement and district attorneys, has resulted in at least six documented successful human trafficking prosecutions in the United States using DIG and other MEMEX-based tools over the last decade.

Scalability and Deployment Considerations

At scale, the system must support corpora exceeding 50 million webpages while still delivering near real-time response to queries. To achieve this, the KG is stored in Elasticsearch, chosen over triplestores for its superior performance, flexible query language, and robust support in cloud environments. The architecture is implemented largely in Python and depends on Apache Spark for distributed processing during KG construction. By relying on widely supported open-source components, the system remains adaptable by both researchers and operational partners.

In summary, the system combines semi-automated and highly scalable KG construction, robust query reformulation in an intuitive interface, and investigator-centered dashboards to transform noisy web data into usable evidence. Its architecture reflects a balance between advanced AI capabilities and the practical demands of deployment in a sensitive and high-stakes investigative environment.

Figure 3: Examples of point-fact (lead investigation), aggregate (aggregation/list) and cluster (dossier generation) queries. The indicator mining task was explored in a separate challenge, which we do not describe in this paper, as it was not included in the deployed system, but more details on which are provided in (Kejriwal et al. 2017).

Experimental Validation

To assess the complete effectiveness of the investigative search engine, we participated in the DARPA MEMEX Human Trafficking Challenge, a four-week evaluation exercise conducted in 2016. The challenge was designed to rigorously test end-to-end systems for knowledge discovery in illicit web domains, using both large-scale noisy corpora and smaller hand-curated datasets. Three independent consortia-teams from academia and industry competed, but our system was the only one that fully integrated KG construction with entity-centric query execution in a unified framework.

The challenge consisted of two test phases. In the first phase, systems were provided with a large, multi-domain web corpus of approximately 1.3 million pages, of which the majority were human trafficking advertisements but a significant fraction were irrelevant (such as job postings and spam). Participants were asked to answer 40 investigative queries, divided into 10 point-fact, 16 cluster, and 14 aggregate questions (Figure 3). Each system produced ranked lists of results, which were evaluated externally by DARPA using both automated and manual relevance judgments.

The second phase used a smaller annotated corpus of roughly 4,000 pages. These pages were manually labeled by domain experts, ensuring that ground-truth answers to all 40 queries were present in the corpus. Importantly, while participants received the raw pages, all annotations were withheld. The aim of this phase was to measure robustness under conditions where the corpus was less noisy but still representative of real-world challenges.

Performance was evaluated using Normalized Discounted Cumulative Gain (NDCG), a standard metric in information retrieval. NDCG rewards systems that rank correct answers highly while penalizing irrelevant results. Two variants were

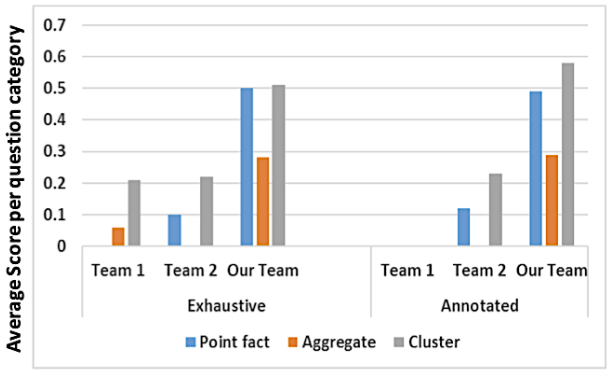

Figure 4: Manual NDCG of participating teams in the DARPA MEMEX Human Trafficking challenge on two corpora.

computed: (1) an automated NDCG, which judged relevance purely on whether the correct page ID was retrieved, and (2) a manual NDCG, in which challenge organizers assessed whether the overall tuple of extracted attributes met investigative needs. Both scores range from 0 (worst) to 1 (perfect).

Results and Error Analysis

Our system achieved strong performance across both corpora and all three query categories (Figure 4). On the large, noisy corpus (denoted as *exhaustive* in the figure), the prototype consistently retrieved relevant answers despite the presence of substantial irrelevant material, demonstrating the robustness of high-recall knowledge graph construction combined with fuzzy query reformulation. Average manual NDCG scores exceeded 0.6 for point-fact queries, indicating that investigators would find the majority of retrieved tuples usable. Cluster queries, which involved discovering latent vendors through connected attributes such as phone numbers or email addresses, achieved average NDCG values in the 0.5–0.6 range. Aggregate queries, which are inherently more complex, scored somewhat lower but still provided actionable leads.

When compared against competing systems, ours was the only approach to obtain non-zero NDCG scores across all query categories and both corpora. Competing prototypes either failed to return relevant answers for certain query types or exhibited instability across datasets. Per-query breakdowns showed that our scores were not driven by a few outliers but were consistently positive across nearly all 40 questions. This stability was attributed to two factors: the use of heterogeneous extraction techniques tuned for obfuscation, and the incorporation of multiple query reformulation strategies that gracefully handled missing or noisy attributes.

Detailed error analysis on queries where the correct answer did not appear in the top results revealed both simple and complex failure modes. Common issues included misspellings (e.g., “Asheera” vs. “Asheerah”), inconsistent attribute formats (such as heights in inches vs. centimeters), and extraction failures on irregular HTML structures. More complex errors arose from near-duplicate advertisements with subtle variations, which confounded certain query strategies. While these issues reduced precision in some cases, they also highlighted the importance of flexible indexing and phonetic matching to improve resilience.

Practical Implications and Potential Cross-Domain Extensions

The evaluation results demonstrated that the system was ready for real-world use, particularly for point-fact and cluster queries. Importantly, tasks that previously required extensive manual effort, such as linking multiple advertisements to the same vendor, could now be substantially automated. With average NDCG scores around 0.6 for these queries, investigators were able to uncover vendor-level entities that would otherwise have been hidden in the data. The results also indicated that aggregate analysis, though more challenging, could be improved with continued tuning of extraction and modeling components. Overall, the DARPA challenge validated the feasibility of deploying a knowledge-graph-based investigative search engine at scale, showing that such tools could provide measurable benefits in high-stakes law enforcement contexts.

Although the initial design and deployment of the investigative search engine focused on combating human trafficking, the underlying architecture is intentionally domain-agnostic. The system’s workflow (crawling, large-scale knowledge graph construction, entity extraction, query reformulation, and investigator-centered interfaces) can be adapted to any domain where illicit activity is manifested through semi-structured or unstructured online data.

The MEMEX program explicitly encouraged such cross-domain applications, and subsequent work demonstrated that the system could be extended with relatively modest adjustments (Kejriwal and Szekely 2019, 2018a). For example, by modifying the ontology and extraction modules, the platform has been used to analyze securities fraud, illicit firearm sales, and online scams. Each of these domains shares a reliance on advertising- or posting-style web content, where critical entities such as contact information, financial identifiers, or product descriptions can be extracted and linked to reveal networks. The entity-centric approach is particularly valuable in domains where actors intentionally obfuscate their identities but rely on persistent attributes e.g., phone numbers, usernames, or financial instruments, leaving digital traces across postings (Kejriwal and Szekely 2018b).

Evaluation in these adjacent domains showed that many of the same challenges recur: high volumes of noisy data, deliberate obfuscation, and the need for clustering and aggregation across heterogeneous sources. The DIG architecture proved effective in these contexts without fundamental redesign, though domain-specific lexicons and extraction rules needed to be developed. Importantly, the separation of schema, extractors, and user interface meant that new domains could be onboarded rapidly once investigative priorities were established. Investigators in securities enforcement, for example, could pivot from phone numbers to ticker symbols or brokerage accounts as central entities of interest,

while those tracking illicit firearm markets could prioritize serial numbers, model names, and geolocation data.

These experiments highlight the generality of the system design: by abstracting away from trafficking-specific attributes to a modular entity-centric pipeline, the investigative search engine can support a family of related applications in law enforcement, regulatory compliance, and consumer protection. The cross-domain extensions also underscore an important policy implication: investments in robust, explainable, and investigator-friendly AI systems for one domain of illicit activity can yield dividends across many others, reducing the need for building bespoke tools from scratch in each case.

Deployment and Real-World Use

Following the DARPA evaluations, the system was transitioned directly into investigative practice. Beginning in 2018, through the office of New York's Human Trafficking Response Unit, DIG was distributed at no cost to more than 200 state and local law enforcement agencies across the United States. Unlike research prototypes that remain confined to the laboratory, DIG became part of day-to-day investigative workflows, enabling agencies with limited resources to access advanced AI-driven search and analytics. Transition was facilitated by the fact that the system required no technical expertise to operate; investigators could interact with it through a familiar search interface while still benefiting from the underlying knowledge graph and query reformulation engine.

Deployment generated measurable and widely acknowledged benefits. Data reported by the New York Human Trafficking Response Unit indicated a dramatic shift in arrest patterns: sex worker arrests decreased substantially, while the proportion of cases subsequently investigated as trafficking rose from less than 1% to more than 60%. In multiple high-profile prosecutions, cached evidence and entity-centric leads obtained through DIG and other MEMEX tools were formally used in court, including a case in Manhattan where a trafficker was convicted after a victim escaped from a sixth-story window, and another in California where a defendant received a 97-year sentence. Prosecutors emphasized the centrality of these tools, with one describing DIG as "indispensable in bringing sex traffickers to justice" and another stating that "we literally couldn't do what we do without MEMEX."

Beyond courtroom outcomes, the deployment of such investigative search engines also helped reshape investigative culture. Investigators reported that the availability of cached webpages, often long after they had disappeared from the live web, was critical for both building cases and establishing trust in the system. The ability to cluster entities and follow trails across domains accelerated the pace of investigations, allowing units to pursue more cases with the same staffing levels. As one assistant district attorney remarked, "you all rock," showing the practical value perceived by frontline users.

The visibility of deployment also extended into broader policy and advocacy circles. MEMEX was featured on CBS' *60 Minutes* and in technical press, highlighting its role in supporting justice in a domain where conventional search engines fail. The system's impact was discussed at international policy forums, including a United Nations roundtable on child trafficking hwere DIG's lead author participated and contributed (Bracket Foundation and Value for Good 2019), and was cited in reports on the use of AI to combat modern slavery[1]. Academic and outreach efforts reinforced this footprint, with publications in top-tier venues, invited talks at law schools and non-governmental organizations, and *pro bono* continuation of support after DARPA funding concluded. Philanthropic contributions, including from the Keston and Kroner Family Foundations, ensured that investigators continued to have access to updated capabilities.

Taken together, these experiences illustrate that the transition from research to practice requires more than technical innovation. Successful deployment depended on building trust with users, aligning with evidentiary requirements in court, and sustaining the system through advocacy and resource mobilization. DIG's trajectory demonstrates how AI systems can achieve lasting social impact when deployed as part of a broader ecosystem of institutions, policies, and communities committed to combating human trafficking.

Discussion

The long-term deployment of the investigative search engine provides a rare opportunity to reflect on how AI systems behave once embedded in sensitive, operational environments. Unlike controlled experiments or benchmark datasets, the reality of investigative search is shaped by dynamic adversaries, fragmented information sources, and the constraints of legal procedures. In such a setting, technical performance alone is insufficient to guarantee success. Instead, adoption has hinged on whether the system could adapt to investigators' workflows, build institutional trust, and withstand the shifting landscape of online illicit activity. These considerations situate the system at the intersection of AI research, policy, and practice, making the deployment experience an important case study in how AI applications move from prototype to sustained use.

The discussion that follows distills insights gathered over several years of active use by more than 200 agencies. While the system has enabled measurable gains in investigative efficiency and has produced evidence used in court, it has also revealed critical limitations that temper expectations about what AI can achieve in complex social domains. At the same time, deployment has surfaced lessons about explainability, usability, and institutional anchoring that extend beyond the specifics of human trafficking investigations. We organize these reflections into two themes—limitations and lessons learned—that we believe will be instructive both for researchers developing AI systems for social good and for practitioners considering long-term adoption of AI-driven investigative tools.

[1]https://cra.org/ccc/wp-content/uploads/sites/2/2021/06/CCC-Code-8-7-Report-Final.pdf

Limitations

Despite the system's long-term deployment and widespread adoption, several limitations remain. First, data coverage is inherently incomplete. While the crawlers have ingested hundreds of millions of pages, investigators routinely encounter ads and platforms that lie outside the scope of existing crawls. Given the dynamic nature of illicit online markets, maintaining comprehensive coverage requires constant adaptation to new domains, obfuscation tactics, and shifting platforms. A second limitation is that accuracy in information extraction is bounded by the quality of automated techniques. Even with hybrid approaches and provenance tracking, noisy extractions and missed attributes are inevitable, especially when traffickers intentionally alter spellings, embed numbers in images, or exploit slang and coded language. These errors can slow down investigative workflows and require manual validation.

A third limitation lies in evaluation. Although controlled studies with NIST and DARPA established baseline performance, measuring accuracy and utility in live investigations is far more challenging. Ground truth is rarely available, and investigators may value partially correct or tangentially relevant results if they generate promising leads. This complicates formal benchmarking and limits the ability to quantify improvements in a manner comparable to standard AI benchmarks. Finally, there are institutional and legal constraints. Not all agencies have equal technical capacity to adopt cloud-based tools, and strict policies around data privacy, evidentiary standards, and chain-of-custody can restrict usage in some jurisdictions. These constraints mean that deployment impact is often heterogeneous across agencies.

Lessons Learned

From the perspective of deployment, some key lessons stand out. First, system trust is as important as system accuracy. Investigators consistently emphasized the importance of source retention e.g., having the cached HTML page or original image available alongside extracted fields. Even when information extraction quality is high, users prefer to verify details against the original source before taking investigative action. The provision of provenance and cached material proved essential for building long-term trust, and distinguishes the system from black-box machine learning approaches.

Second, explainability and transparency are critical in high-stakes domains. Investigators are reluctant to rely on opaque outputs, particularly when cases may proceed to court. By employing clustering algorithms that were explainable and scalable, rather than black-box methods that might marginally improve accuracy, DIG aligned more closely with investigative needs. This choice, though technically conservative, facilitated adoption and highlighted the importance of human-AI collaboration over purely algorithmic gains.

Third, usability and integration into workflows determine sustained use. Controlled usability studies demonstrated that modest interface improvements, such as faceted search and entity-centric pages, had outsize impact on investigator efficiency. Training sessions and iterative feedback loops further reinforced adoption, as users saw their feedback reflected in subsequent versions. A notable insight was the divide between lead generation and lead investigation. While developers initially envisioned risk scores and automated lead generation as central features, field experts indicated that they already had an abundance of leads, and that their real challenge was efficiently investigating and corroborating them. This feedback redirected development priorities toward features that supported investigation rather than speculative lead creation.

Finally, long-term maintenance requires anticipating both technical and institutional evolution. On the technical front, illicit markets migrate rapidly across platforms, demanding agile crawling, new extraction models, and updated schemas. On the institutional side, sustainability hinges on partnerships with agencies, policy-makers, and prosecutors. The eventual transfer of the system to state offices such as the District Attorney of New York illustrates the importance of institutional anchoring beyond research funding cycles. At the same time, it raises broader questions about how AI for social good can be supported and maintained outside of commercial business models.

Together, these lessons suggest that building and deploying AI for sensitive investigative domains is not only a technical challenge but also a socio-technical one. Success depends on aligning system design with user trust, institutional constraints, and evolving policy environments.

Beyond Prosecution: Understanding the Upstream Drivers of Trafficking

Future work on AI-enabled investigative systems for human trafficking must be located within the broader policy framework known as the "3Ps"(Prosecution, Protection, and Prevention) articulated in the Palermo Protocol (Zhang 2022) and embedded in the U.S. Trafficking Victims Protection Act (Wooditch, DuPont-Morales, and Hummer 2009). Our system was originally developed with a focus on prosecution, providing investigators and district attorneys with verifiable leads, historical evidence, and structured insights that could be introduced in court. The impact in this domain has been significant, with measurable shifts in arrest patterns and documented successes in securing convictions. Yet an exclusive focus on prosecution is insufficient. Victims require *protection* once identified, and more importantly, upstream drivers of trafficking must be disrupted to *prevent* exploitation before it occurs.

This recognition motivates an expanded agenda. On the protection side, future versions of the system must be adapted to interface with victim-service organizations, enabling secure sharing of non-identifiable intelligence that supports rapid intervention while safeguarding privacy. Such efforts will demand careful design of data-handling protocols and explainability features to ensure that insights can be trusted by non-law-enforcement actors. On the prevention side, the challenge is even greater: identifying structural and societal risk factors before trafficking networks become en-

trenched. Addressing this dimension requires collaboration beyond traditional investigative agencies, involving regulators, civil society, and survivor-informed advocacy groups.

To this end, we have recently launched the *Global Trafficking Initiative*[2], supported by the Kroner Family Foundation. This initiative broadens the focus beyond case-level prosecution toward systematic knowledge-building on regional and global prevalence, legal frameworks, and policy responses. It is envisioned as a living, open-access resource: a platform that aggregates survivor-informed materials, country-level statistics, and research evidence into a dynamic knowledge base. By doing so, it aims to overcome one of the persistent barriers in anti-trafficking work; namely, the lack of reliable, up-to-date, and globally comparable data. The initiative also establishes a forum for connecting academic research with practitioners and policymakers, ensuring that technical innovations translate into meaningful preventive strategies.

The trajectory of our work illustrates both the potential and the limitations of AI for social good. Sustaining and updating deployed systems requires resources, long-term commitments, and continuous trust-building with diverse stakeholders. Looking ahead, advances in language models and generative AI may provide new opportunities for knowledge extraction and multilingual analysis of trafficking-related content. At the same time, these technologies raise new risks, including adversarial misuse and privacy breaches, that must be addressed through careful governance. Our ongoing research will explore how to responsibly integrate such tools while retaining the central principles that enabled past success: usability, explainability, and coalition-building across the 3Ps. Ultimately, the fight against trafficking cannot be won by technology alone; it requires alignment of technical, legal, and policy instruments. The lessons learned from our deployment highlight the importance of approaching AI systems not as isolated artifacts but as components of a global response to one of the most urgent human rights crises of our time.

Future Work and Conclusion

Looking ahead, sustaining and extending the impact of investigative AI systems requires attention to both technical and institutional factors. On the technical side, illicit online markets are highly adaptive, with new platforms, coded language, and obfuscation techniques emerging rapidly. Ensuring that the system remains useful requires regular updates to crawlers, extractors, and schemas, but resource limitations at both research labs and public agencies can hinder this process. A key policy challenge is to identify funding and governance models that enable continual maintenance outside of short-term research programs. Without institutionalized support, systems risk obsolescence even after demonstrating measurable benefits. Advocacy within policy communities is therefore critical, not only to secure resources but also to establish norms for responsible adoption of AI in sensitive domains.

[2]globaltrafficking.org

Future directions also involve rethinking how AI techniques can enhance investigative workflows. Recent advances in LLMs offer opportunities for augmenting the system with more flexible natural language querying, context-aware summarization, and automated detection of emerging obfuscation patterns. At the same time, such capabilities introduce new risks: LLMs can generate inaccurate outputs, amplify bias, or undermine the evidentiary standards required in court (Fang and Perkins 2024). A promising path forward is to treat LLMs as assistive agents for investigators rather than autonomous decision-makers, embedding them in human-in-the-loop frameworks that prioritize transparency and verifiability. Similarly, greater emphasis on explainability and provenance will remain essential, ensuring that AI-generated outputs can be cross-checked against retained sources.

From a broader perspective, the deployment of the investigative search engine illustrates both the potential and the fragility of AI for social good. The system has already contributed to improved efficiency in investigations and, in some cases, to successful prosecutions. Yet its future depends on continued partnerships between technologists, investigators, and policymakers. As illicit markets evolve, so too must the sociotechnical ecosystem surrounding investigative AI: legal standards for digital evidence, ethical guidelines for data use, and mechanisms for inter-agency collaboration. Our experience underscores that innovation in this domain is not a one-time technical achievement but an ongoing process of adaptation and advocacy. We conclude that the most impactful role for AI in combating human trafficking and related crimes will come not from ever more complex algorithms alone, but from building resilient, institutionally supported systems that can serve as trusted partners in the pursuit of justice.

Acknowledgments

The results in this article were supported under multiple grants and efforts, most notably the DARPA Memex program, the Keston Foundation, and the Kroner Family Foundation. The views and opinions expressed in this article are those of the author alone.

References

Ahlers, D. 2013. Assessment of the accuracy of GeoNames gazetteer data. In *Proceedings of the 7th workshop on geographic information retrieval*, 74–81.

Alvari, H.; Shakarian, P.; and Snyder, J. K. 2016. A non-parametric learning approach to identify online human trafficking. In *Proceedings of the IEEE Conference on Intelligence and Security Informatics (ISI)*, 133–138. IEEE.

Bracket Foundation and Value for Good. 2019. Combating Online Sexual Abuse of Children: White Paper Executive Summary. Technical report, Concordia Summit Roundtable on Combating Online Sexual Abuse of Children. Accessed: August 14, 2025.

Chang, C.-H.; Kayed, M.; Girgis, M. R.; and Shaalan, K. F. 2006. A survey of web information extraction systems.

IEEE Transactions on Knowledge and Data Engineering, 18(10): 1411–1428.

Chen, H. 2011. *Dark Web: Exploring and Data Mining the Dark Side of the Web*, volume 30. Springer Science & Business Media.

Dalvi, N.; Kumar, R.; Pang, B.; Ramakrishnan, R.; Tomkins, A.; Bohannon, P.; Keerthi, S.; and Merugu, S. 2009. A Web of Concepts. In *Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 1–12. ACM.

Doan, A.; Halevy, A.; and Ives, Z. 2012a. *Principles of Data Integration*. Elsevier.

Doan, A.; Halevy, A.; and Ives, Z. 2012b. Principles of Data Integration. In *Proceedings of the VLDB Endowment*. Elsevier.

Elmagarmid, A. K.; Ipeirotis, P. G.; and Verykios, V. S. 2007. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1): 1–16.

Fang, A.; and Perkins, J. 2024. Large language models (LLMs): Risks and policy implications. *MIT Sci. Policy Rev.*, 5: 134–45.

Frank, J. R.; Kleiman-Weiner, M.; Roberts, D. A.; Niu, F.; Zhang, C.; Ré, C.; and Soboroff, I. 2012a. Building an entity-centric stream filtering test collection for TREC 2012. Technical report, DTIC Document.

Frank, J. R.; Kleiman-Weiner, M.; Roberts, D. A.; Niu, F.; Zhang, C.; Ré, C.; and Soboroff, I. 2012b. Building an entity-centric stream filtering test collection for TREC 2012. Technical report, DTIC Document.

Greiman, V.; and Bain, C. 2013a. The emergence of cyber activity as a gateway to human trafficking. In *Proceedings of the 8th International Conference on Information Warfare and Security (ICIW)*, 90. Academic Conferences Limited.

Greiman, V.; and Bain, C. 2013b. The emergence of cyber activity as a gateway to human trafficking. In *Proceedings of the 8th International Conference on Information Warfare and Security (ICIW)*, 90. Academic Conferences Limited.

Harrendorf, S.; Heiskanen, M.; and Malby, S. 2010. *International Statistics on Crime and Justice*. European Institute for Crime Prevention and Control, affiliated with the United Nations (HEUNI).

Hirschman, L.; and Gaizauskas, R. 2001. Natural language question answering: the view from here. *Natural Language Engineering*, 7(4): 275–300.

Hogan, A.; Harth, A.; Umrich, J.; and Decker, S. 2007a. Towards a scalable search and query engine for the web. In *Proceedings of the 16th International Conference on World Wide Web*, 1301–1302. ACM.

Hogan, A.; Harth, A.; Umrich, J.; and Decker, S. 2007b. Towards a scalable search and query engine for the web. In *Proceedings of the 16th International Conference on World Wide Web*, 1301–1302. ACM.

Hultgren, M.; Jennex, M. E.; Persano, J.; and Ornatowski, C. 2016a. Using knowledge management to assist in identifying human sex trafficking. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, 4344–4353. IEEE.

Hultgren, M.; Jennex, M. E.; Persano, J.; and Ornatowski, C. 2016b. Using knowledge management to assist in identifying human sex trafficking. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, 4344–4353. IEEE.

Jain, A.; Doan, A.; and Gravano, L. 2007. SQL queries over unstructured text databases. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)*, 1255–1257. IEEE.

Kejriwal, M.; Ding, J.; Shao, R.; Kumar, A.; and Szekely, P. 2017. Flagit: A system for minimally supervised human trafficking indicator mining. *arXiv preprint arXiv:1712.03086*.

Kejriwal, M.; Shao, R.; and Szekely, P. 2019. Expert-guided entity extraction using expressive rules. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 1353–1356.

Kejriwal, M.; and Szekely, P. 2017a. Information extraction in illicit web domains. In *Proceedings of the 26th international conference on world wide web*, 997–1006.

Kejriwal, M.; and Szekely, P. 2017b. Knowledge graphs for social good: An entity-centric search engine for the human trafficking domain. *IEEE Transactions on Big Data*, 8(3): 592–606.

Kejriwal, M.; and Szekely, P. 2017c. Neural embeddings for populated geonames locations. In *International Semantic Web Conference*, 139–146. Springer.

Kejriwal, M.; and Szekely, P. 2018a. Constructing domain-specific search engines with no programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kejriwal, M.; and Szekely, P. 2018b. Technology-assisted investigative search: A case study from an illicit domain. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–9.

Kejriwal, M.; and Szekely, P. 2019. myDIG: Personalized illicit domain-specific knowledge discovery with no programming. *Future Internet*, 11(3): 59.

Kejriwal, M.; Szekely, P.; and Knoblock, C. 2018. Investigative knowledge discovery for combating illicit activities. *IEEE Intelligent Systems*, 33(1): 53–63.

Lin, T.; Pantel, P.; Gamon, M.; Kannan, A.; and Fuxman, A. 2012a. Active objects: Actions for entity-centric search. In *Proceedings of the 21st International Conference on World Wide Web*, 589–598. ACM.

Lin, T.; Pantel, P.; Gamon, M.; Kannan, A.; and Fuxman, A. 2012b. Active objects: Actions for entity-centric search. In *Proceedings of the 21st International Conference on World Wide Web*, 589–598. ACM.

Niu, F.; Zhang, C.; Ré, C.; and Shavlik, J. W. 2012a. DeepDive: Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of the VLDS Workshop*, 25–28.

Niu, F.; Zhang, C.; Ré, C.; and Shavlik, J. W. 2012b. DeepDive: Web-scale knowledge-base construction using statistical learning and inference. In *Proceedings of the VLDS Workshop*, 25–28.

Saleiro, P.; Teixeira, J.; Soares, C.; and Oliveira, E. 2016a. TimeMachine: Entity-centric search and visualization of news archives. In *European Conference on Information Retrieval (ECIR)*, 845–848. Springer.

Saleiro, P.; Teixeira, J.; Soares, C.; and Oliveira, E. 2016b. TimeMachine: Entity-centric search and visualization of news archives. In *European Conference on Information Retrieval (ECIR)*, 845–848. Springer.

Sauro, J. 2011a. Measuring usability with the system usability scale (SUS). https://measuringu.com/sus/.

Sauro, J. 2011b. Measuring usability with the system usability scale (SUS). In *Proceedings of the Human Factors International Conference*.

Savona, E. U.; and Stefanizzi, S. 2007. *Measuring Human Trafficking*. Springer.

Szekely, P.; Knoblock, C. A.; Slepicka, J.; Philpot, A.; Singh, A.; Yin, C.; Kapoor, D.; Natarajan, P.; Marcu, D.; Knight, K.; et al. 2015a. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference (ISWC)*, 205–221. Springer.

Szekely, P.; Knoblock, C. A.; Slepicka, J.; Philpot, A.; Singh, A.; Yin, C.; Kapoor, D.; Natarajan, P.; Marcu, D.; Knight, K.; et al. 2015b. Building and using a knowledge graph to combat human trafficking. In *International Semantic Web Conference (ISWC)*, 205–221. Springer.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2012a. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 125–134. ACM.

Tonon, A.; Demartini, G.; and Cudré-Mauroux, P. 2012b. Combining inverted indices and structured search for ad-hoc object retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 125–134. ACM.

Wooditch, A. C.; DuPont-Morales, M.; and Hummer, D. 2009. Traffick jam: a policy review of the United States' Trafficking Victims Protection Act of 2000. *Trends in Organized Crime*, 12(3): 235–250.

Zhang, S. X. 2022. Progress and challenges in human trafficking research: Two decades after the Palermo Protocol. *Journal of human trafficking*, 8(1): 4–12.