future internet

Article

myDIG: Personalized Illicit Domain-Specific Knowledge Discovery with No Programming

Mayank Kejriwal * and Pedro Szekely

Information Sciences Institute, University of Southern California, Marina del Rey, CA 90502, USA; pszekely@isi.edu

* Correspondence: kejriwal@isi.edu; Tel.: +1-217-819-6696

Received: 19 January 2019; Accepted: 22 February 2019; Published: 4 March 2019

Abstract: With advances in machine learning, knowledge discovery systems have become very complicated to set up, requiring extensive tuning and programming effort. Democratizing such technology so that non-technical domain experts can avail themselves of these advances in an interactive and personalized way is an important problem. We describe myDIG, a highly modular, open source pipeline-construction system that is specifically geared towards investigative users (e.g., law enforcement) with no programming abilities. The myDIG system allows users both to build a knowledge graph of entities, relationships, and attributes for illicit domains from a raw HTML corpus and also to set up a personalized search interface for analyzing the structured knowledge. We use qualitative and quantitative data from five case studies involving investigative experts from illicit domains such as securities fraud and illegal firearms sales to illustrate the potential of myDIG.

Keywords: knowledge discovery; domain specific; no programming; knowledge graphs; information extraction; investigative domains; search; personalized analytics

1. Introduction

As research in machine learning and AI continues to progress, front-facing systems have become steadily more complicated, expensive, and limited to specific applications like Business Intelligence (BI) [1,2]. Democratizing technology is a complex issue that involves multiple stakeholders with different sets of abilities [3,4]. A particular user whose needs are very real, but which are seldom addressed except in BI or military-specific situational awareness situations [5], is a domain expert with extremely limited technical abilities. In particular, such users do not know how to program, let alone cope with complex machine learning or deep learning algorithms, and cannot satisfy their information needs through a simple Google search for several subsequently-described reasons.

A real-world example of such a domain expert that we encountered in the securities fraud domain is an employee from the Securities and Exchange Commission (SEC) who is attempting to identify actionable cases of penny stock fraud [6]. Penny stock offerings in Over-The-Counter (OTC) markets are frequently suspected of being fraudulent, but without specific evidence, usually in the form of a false factual claim (that is admissible as evidence), and their trading cannot be halted. With thousands of penny stock offerings, investigators do not have the resources or time to investigate all of them. One (technical) way to address this problem is to first crawl a corpus of relevant pages from the web describing the domain, a process alternately known in the Information Retrieval (IR) literature as relevance modeling or domain discovery [7]. The latter term is more encompassing, as it involves not just relevance modeling, but the actual crawling of the data.

Once such a corpus is obtained, an expert in information extraction and machine learning would elicit opinions from the users on what fields (e.g., location, company, stock ticker symbol) are important to the user for answering domain-specific questions, along with example extractions per field. This sequence of *Knowledge Graph Construction* (KGC) steps results in a graph-theoretic representation of the data (a *Knowledge Graph* or KG) where nodes are entities and field values and the (directed, labeled) edges are relationships between entities or assignments of field values to entities [8,9]. Since the KG is structured, it is amenable to aggregations and to both keyword and structured querying. With a good interface, for example, the domain expert can identify all persons and organizations (usually shell companies) associated with a stock ticker symbol, aggregate prices, or zero in on suspicious activity by searching for hyped-up phrases that indicate fraud.

A Google search is inadequate for the knowledge discovery described above as it does not allow one to specify fields nor to limit the search to a specific corpus. Google also does not offer the specialized aggregation facilities that such domain experts would need. While Google works exceedingly well for document-level information indexing and retrieval, it is not designed to support complex analytical tasks, either with or without training.

In this paper, we present a system called myDIG (my Domain-specific Insight Graphs)that ingests the output of domain discovery, i.e., a raw corpus of web pages, and presents an intuitive multi-step KGC and search interface to a user that requires no programming, can be refined iteratively, and requires an hour or less of example-based training. An implemented prototype of myDIG has already undergone evaluations by domain experts in five different investigative domains, each of which involves significant technological potential. We describe the dataflow of myDIG and the important algorithmic aspects of the system that allow users to construct and search knowledge graphs, and thereby discover new and relevant knowledge by conducting data-driven analysis. We also present qualitative and quantitative data collected from the five case studies to illustrate the real-world potential of myDIG in making an advanced set of knowledge discovery technologies accessible to non-technical users. Using myDIG, practicing, non-technical domain experts from the five case study domains were able to build a personalized domain-specific search engine over corpora containing more than a million raw web pages in 4–6 working hours. The myDIG project is open-source, with a dedicated web page (http://usc-isi-i2.github.io/dig/) and a GitHub (https://github.com/usc-isi-i2/dig-etl-engine) repository associated with it.

2. Related Work

The myDIG system draws on work in several different communities in AI. To the best of our knowledge, it is the first KGC and search system that allows users to set up their domains (and perform knowledge discovery in it) with no programming or tuning of any kind. However, the components underlying myDIG draw on important advances in areas such as knowledge graph construction and information retrieval, described below. We also provide examples of work in the AI community with a similar philosophy to ours: allowing a non-technical user, typically a domain expert, to avail themselves of, and navigate, AI technology with easy or minimal setup.

2.1. Knowledge Graph Construction

Knowledge Graph Construction (KGC) has witnessed much research attention in the last decade, an industry-scale success story being the Google Knowledge Graph [10]. A knowledge graph is usually defined as a set of triples of the form (h, r, t), where h and t are nodes representing entities and r is a relationships between the entities. In communities like the Semantic Web, much finer-grained definitions have been recognized [11,12]. For example, a node can be not just an entity, but also a field (equiv. attribute) value. A relationship in this context can exist between an entity and a literal (e.g., $(John, : age, 29)$), which is an atomic value like a string or integer, or between two entities (e.g., $(John, : brother\_of, Mary)$). In the research literature, KGC is a broad term that primarily includes Information Extraction (IE), but can also include other post-extraction inferential steps [8,9]. Good surveys of IE, particularly web IE, may be found in [13–15].

The myDIG system uses a set of IE technologies but does not require the user to understand such technologies, provide training data, or perform any algorithmic tuning. IE technologies include an interactive wrapper-based [16,17] IE called Inferlink that is subsequently described, glossary extractions (which use a glossary of terms that are extracted from text using exact matching) [18,19], declarative frameworks like regular expressions [20,21], and natural language rules such as implemented by packages like SpaCy or Stanford NER [22,23]. In general, interactive KGC systems, even involving technical expertise, are still quite uncommon. A good example is the DeepDive system [21], which allows customized, interaction-driven information extraction. Another system, Snorkel, which relies on weak supervision, thereby easing the burden of acquiring and manually annotating large amounts of training data, requires users to code functions and like DeepDive and has no facilities for supporting search and analytics [24].

2.2. Information Retrieval

The search in myDIG draws upon several independently-developed techniques in the IR community. Chief among these are query reformulation and constraint relaxation [25–28]. These techniques are designed to be robust to erroneous or missing data, which is an unavoidable problem for semi-automatic KGC systems. Our search system uses advances in NoSQL databases like Elasticsearch, which use optimized inverted index querying for fast retrieval [29]. A survey of NoSQL can be found in [30]. The myDIG system also supports faceting and filtering [31], in addition to basic keyword search.

2.3. AI for Non-Technical Domain Experts

In recent years, there has been the recognition in both the machine learning and database communities of the importance of building end-to-end interactive systems that adapt to users' needs and are relatively robust. We cite some influential examples that are pertinent to myDIG. In the database community, interactive data exploration and novel methods of data visualization are currently active areas of research, especially for specific domains like crisis informatics or social media analysis. Examples of influential systems that are visualization- and human-centric include Ushahidi [32], Twitris [33], Twitcident [34], AIDR [35], CrisisTracker [36], TweetTracker [37], and several others [38,39]. These systems are highly engineered to their domain (describing crisis informatics data like natural disasters and terrorism) and usually facilitate a specific task, such as gleaning situational awareness at scale. Another such highly-tuned application that we noted in the Introduction is Business Intelligence (BI) [1,2]. A commonality with myDIG is that such systems, post-engineering, are designed to be operated without programming or technical ability and involve graphical interfaces. An important difference, however, is that myDIG can be set up without technical effort, which makes it amenable to search and analytics for arbitrary domains and experts.

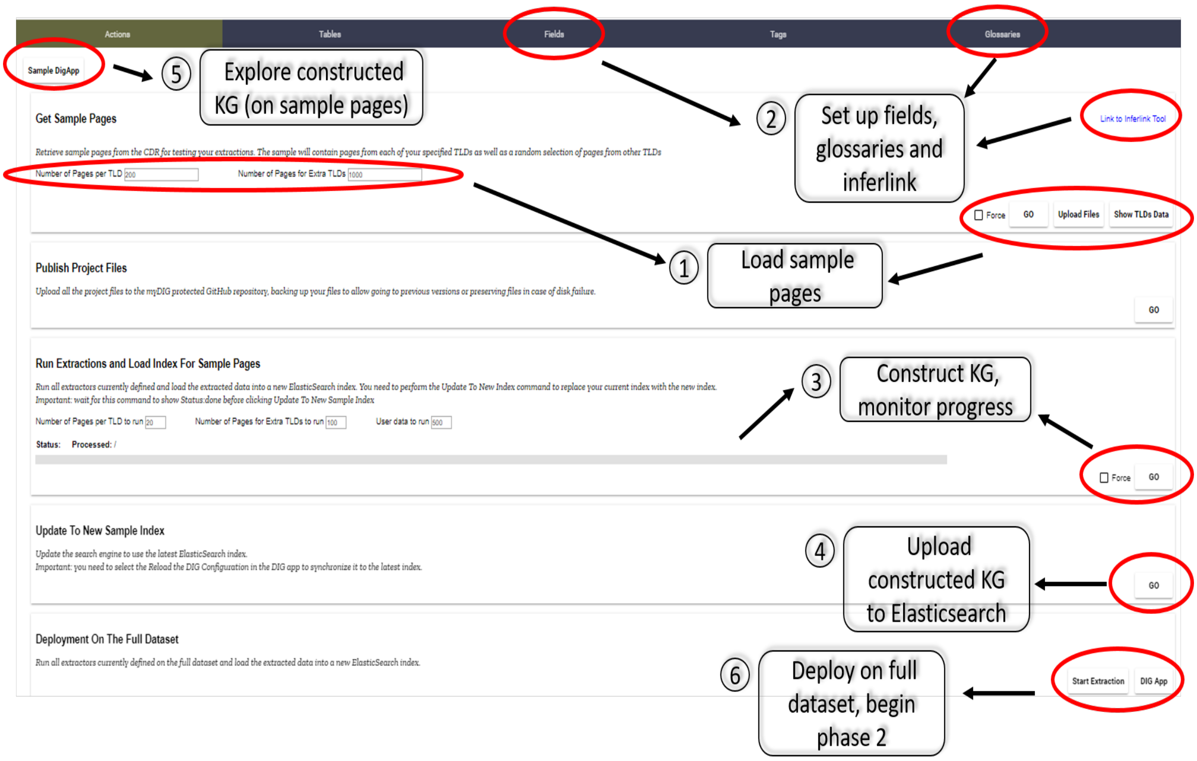
The myDIG architecture is heavily influenced by similar developments in the HCI community that attempt to ease the pain of setting up sophisticated algorithms or visualizations. A good example of such a development is NewsViews [40]. NewsViews is an automated news visualization system that generates interactive, annotated maps without requiring professional designers. Its motivation is similar to myDIG, which allows complex knowledge graph-based search and analytics engines to be set up without requiring programmers or data scientists.

3. System Overview and User Experience

The myDIG action dashboard is illustrated in Figure 1. The input to the system is a corpus of web pages that is assumed to have been crawled by a domain discovery system prior to the user engaging with myDIG. While domain discovery is, by no means, a solved research problem, significant advances have been made in recent years (including the *deep crawls* of specific *Top Level Domains* or TLDs like backpage.com) [41–43]. We make the loose assumption that a significant fraction of this corpus is relevant, but that there may be many irrelevant pages as well. The myDIG architecture is designed to be reasonably robust to such irrelevance, as we explain.

User experience in myDIG can be separated into two phases. The first phase, *domain setup and curation*, involves setting up the domain with the goal of answering a certain set of questions in which the user is interested. This phase does not comprise a strictly linear set of steps, but can involve several interleaved steps (Figure 1). At a high level, the user loads a sample of the corpus to explore, followed by defining wrappers using the Inferlink tool, customized domain-specific fields, and add field-specific glossaries (if desired). Periodically, the user can crystallize a sequence of steps by running extractions (akin to constructing the knowledge graph) and uploading the knowledge graph to an index. Once the upload is complete, the user can "demo" their effort by clicking on the *Sample DigApp* button in the upper left corner in Figure 1 to explore the knowledge graph using a search interface. The process is iterative: the user can always return to the dashboard to define or refine more fields, define more wrappers with Inferlink, or input more glossaries.

The second phase is the "in-use" phase, when the system is actually being used to satisfy information retrieval needs. This phase tends to begin when the user is finished setting up the domain and is ready to deploy all changes on the full dataset (the entire corpus). In subsequent sections, we describe the user experience and dataflow for both phases.


Actions
Tables
Fields
Tags
Glossaries
Sample DigApp
5 Explore constructed KG (on sample pages)
2 Set up fields, glossaries and inferlink
Link to Inferlink Tool
☐ Force
GO
Upload Files
Show TLDs Data
1 Load sample pages
GO
3 Construct KG, monitor progress
☐ Force
GO
4 Upload constructed KG to Elasticsearch
GO
6 Deploy on full dataset, begin phase 2
Start Extraction
DIG App

Get Sample Pages

Retrieve sample pages from the CDR for testing your extractions. The sample will contain pages from each of your specified TLDs as well as a random selection of pages from other TLDs

Number of Pages per TLD 200 Number of Pages for Extra TLDs 1000

Publish Project Files

Upload all the project files to the myDIG protected GitHub repository, backing up your files to allow going to previous versions or preserving files in case of disk failure.

Run Extractions and Load Index For Sample Pages

Run all extractors currently defined and load the extracted data into a new ElasticSearch index. You need to perform the Update To New Index command to replace your current index with the new index.
Important: wait for this command to show Status:done before clicking Update To New Sample Index

Number of Pages per TLD to run 20 Number of Pages for Extra TLDs to run 100 User data to run 500

Status: Processed: /

Update To New Sample Index

Update the search engine to use the latest ElasticSearch index.
Important: you need to select the Reload the DIG Configuration in the DIG app to synchronize it to the latest index.

Deployment On The Full Dataset

Run all extractors currently defined on the full dataset and load the extracted data into a new ElasticSearch index.

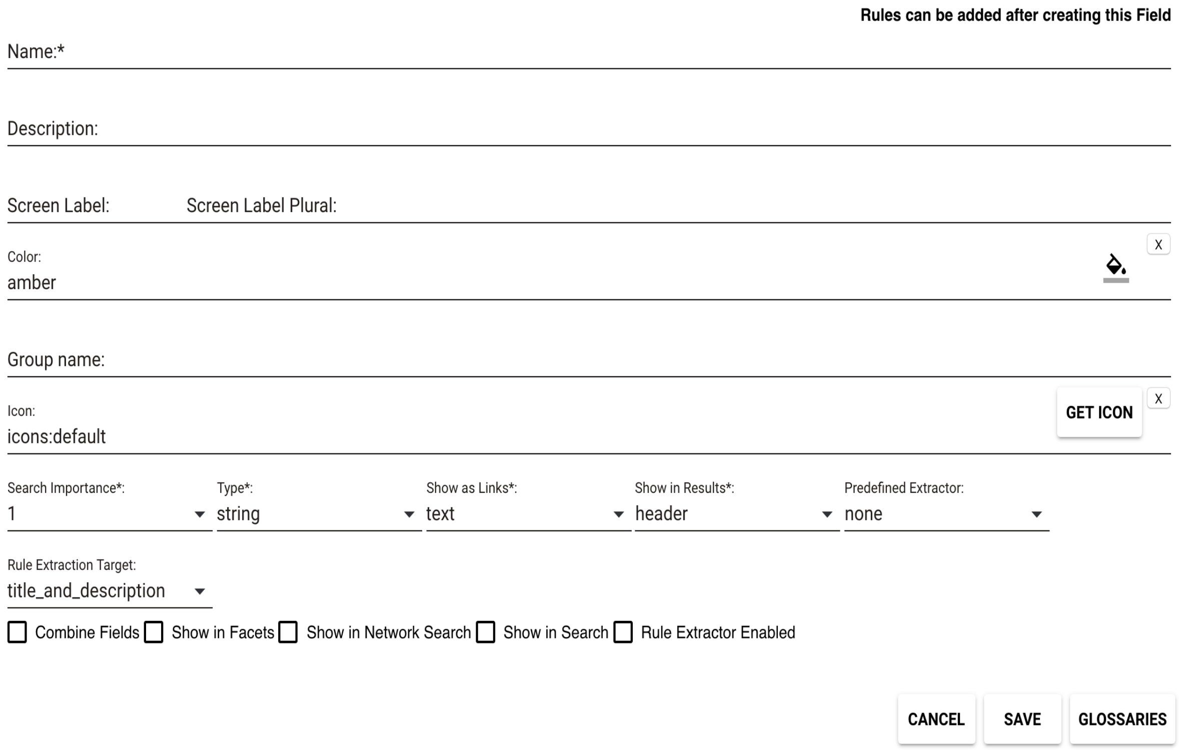
Figure 1. The myDIG dashboard, with corresponding elements and actions.

3.1. Setting Up the Domain

The myDIG system allows personalized setting up of a domain. Users can define and customize their own fields, choose what extractions to retain from web pages (using the Inferlink tool), add glossaries that might be available to them as investigators, and repeat the process till they are satisfied with the domain setup.

Defining Fields and Glossaries

The myDIG system allows users to define their own fields and to customize the fields in several ways that directly influence their use during the second phase (domain exploration). To define a field, users click on the *Fields* tab (Figure 1), which provides them with an overview of fields that are already defined (including pre-defined fields like location, detailed further below) and also allows them to add their own fields. A "field form" is illustrated in Figure 2. In addition to customizing the appearance of a field by assigning it a color and icon, users can set the "importance" of the field for search (on a scale of 1–10), declare the field to represent an entity by selecting the “entity” rather than the “text” option in *Show as Links* (thereby supporting entity-centric search, described in *Exploring the Domain*), and assign it a pre-defined extractor like a glossary.



Rules can be added after creating this Field

Name:*

Description:

Screen Label: Screen Label Plural:

Color: amber

Group name:

Icon: icons:default

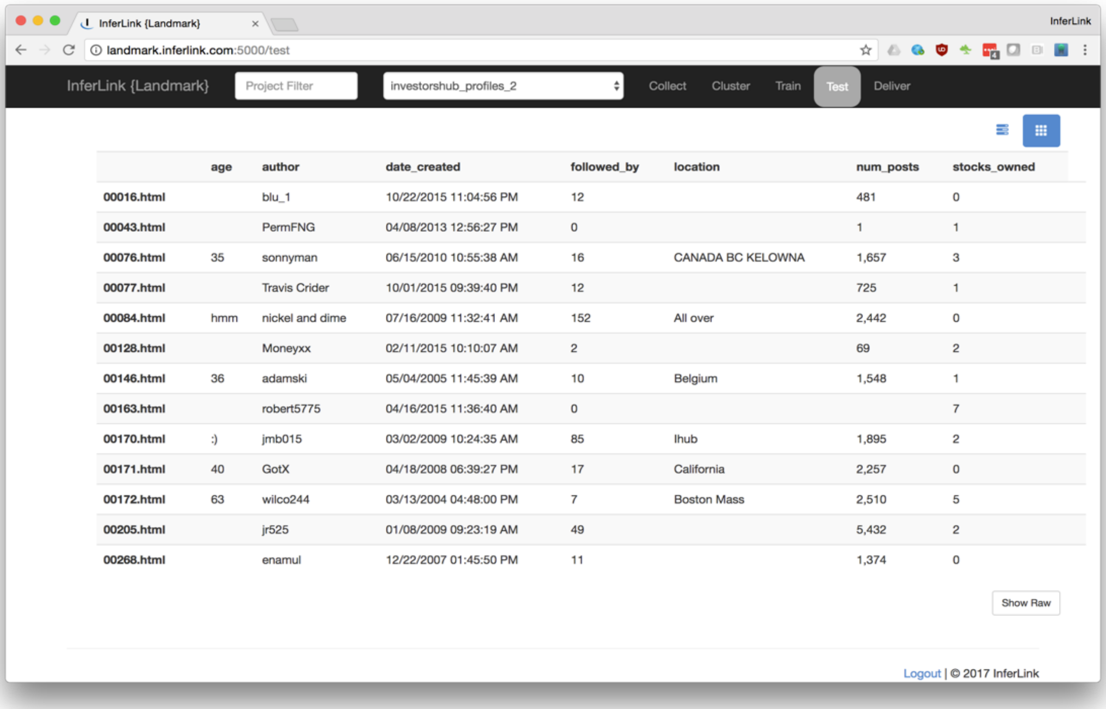
Search Importance*:	Type*:	Show as Links*:	Show in Results*:	Predefined Extractor:
1	string	text	header	none

Rule Extraction Target: title_and_description

□ Combine Fields □ Show in Facets □ Show in Network Search □ Show in Search □ Rule Extractor Enabled

Figure 2. A form that allows the user to define/customize fields.

A powerful, interactive extractor that uses wrapper technology [16,17] to extract structured elements from web pages is the Inferlink tool. Inferlink operates in various steps. For convenient formalism, let us define a TLD (e.g., backpage.com) T as a set $\{w_1, \ldots, w_n\}$ of web pages (e.g., backpage.com/chicago/1234). As a first step, Inferlink uses an unsupervised template clustering algorithm to partition T into clusters, such that web pages in each cluster are structurally similar to each other. To provide some real-world intuition, one cluster could contain taxonomy web pages that contain lists of things; another cluster could contain web pages describing forum posts, while a third cluster could contain web pages containing long text. Inferlink presents the users with samples from these clusters and allows users to select a “relevant” cluster for curation. At this point, users see an illustration like the one in Figure 3 wherein Inferlink has extracted common structural elements from the web pages in the cluster and presents it to the users in a column layout, with one row per web page and one column per structured element that is common to the web pages. Users can open a web page by clicking on a link, delete a column, or assign it to a field that has already been defined, as in the figure. In the literature, this step is also called *semantic typing* of columns [44]. Users must separately type columns for each top level domain (TLD), since different web domains share different structures in the web pages they contain. We also trained users to perform more advanced customizations with Inferlink in less than an hour of example-based demonstrations. For example, users could choose to curate and semantically type more than one cluster from a given TLD, if they felt it gave them added value over another cluster from a less important TLD.



	age	author	date_created	followed_by	location	num_posts	stocks_owned
00016.html		blu_1	10/22/2015 11:04:56 PM	12		481	0
00043.html		PermFNG	04/08/2013 12:56:27 PM	0		1	1
00076.html	35	sonnyman	06/15/2010 10:55:38 AM	16	CANADA BC KELOWNA	1,657	3
00077.html		Travis Crider	10/01/2015 09:39:40 PM	12		725	1
00084.html	hmm	nickel and dime	07/16/2009 11:32:41 AM	152	All over	2,442	0
00128.html		Moneyxx	02/11/2015 10:10:07 AM	2		69	2
00146.html	36	adamski	05/04/2005 11:45:39 AM	10	Belgium	1,548	1
00163.html		robert5775	04/16/2015 11:36:40 AM	0			7
00170.html	:)	jmb015	03/02/2009 10:24:35 AM	85	Ihub	1,895	2
00171.html	40	GotX	04/18/2008 06:39:27 PM	17	California	2,257	0
00172.html	63	wilco244	03/13/2004 04:48:00 PM	7	Boston Mass	2,510	5
00205.html		jr525	01/08/2009 09:23:19 AM	49		5,432	2
00268.html		enamul	12/22/2007 01:45:50 PM	11		1,374	0

Figure 3. An illustration of the semantic typing facility in Inferlink. In this illustration, semantic typing has already been done.

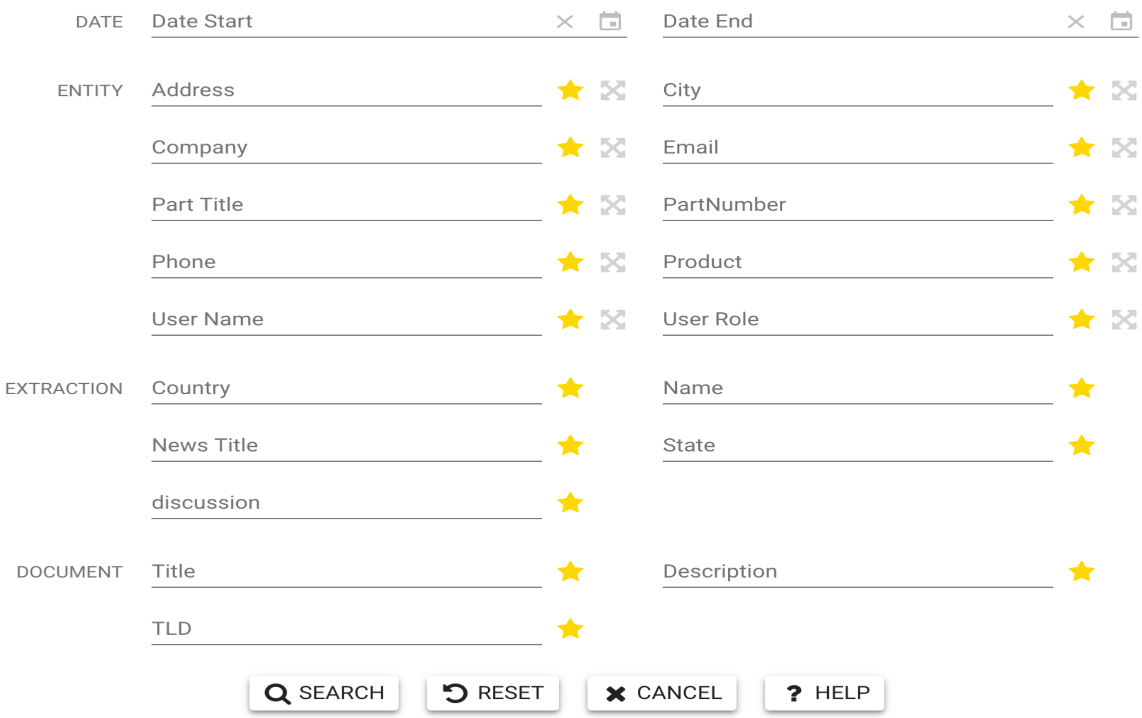
In addition to the Inferlink took, myDIG offers extraction methods suitable for information extraction from blocks of text or other content that is not delimited as structured HTML elements, i.e., delimited using HTML tags. To ease user effort, some generic extractors are pre-trained and cannot be customized, but can be disabled. A good example is the location extractor, which extracts the names of cities, states, and countries, using a machine learning model that was trained offline. However, myDIG offers users the option to input a glossary (with incumbent options, such as whether the glossary terms should be interpreted case sensitively or not) for a given field. This option was popular with domain experts, as we describe in the *Qualitative Feedback* section. Although not evaluated herein, the latest version of myDIG also offers users an intuitive rule editor for expressing and testing simple natural language rules or templates with extraction placeholders. For example, a name extractor can be set up with a pattern recognition rule like “Hi, my name is [NAME]”.

3.2. Exploring the Domain

Once the domain has been set up, users begin the second phase, which involves exploring the domain to get the answers they need or even to generate new leads and questions (an important concern in investigative domains where users are sometimes not sure quite what they are looking for). We describe the main components of a typical user experience.

3.2.1. Specifying Queries: Keywords and More

The myDIG system supports both basic keyword search, as well as *structured search*, wherein the user curating the domain can fill out values in a form containing fields that she/he has declared (during domain setup) as being searchable (Figure 4). As described in *Algorithms and Technologies*, the search engine in myDIG uses an advanced set of techniques from the IR literature to ensure that user intent is captured in a robust and high-recall manner. The myDIG system also supports entity-centric search, which is described shortly. While keyword search is designed to be primarily exploratory, structured search allows a user to quickly hone in on pages containing certain key details that the user has specified in the form. Since myDIG uses ranking and relevance scoring, satisfying more criteria on a form will lead to a page having higher ranking, compared to another page that satisfies fewer criteria.



DATE	Date Start		Date End	
ENTITY	Address	★	City	★
	Company	★	Email	★
	Part Title	★	PartNumber	★
	Phone	★	Product	★
	User Name	★	User Role	★
EXTRACTION	Country	★	Name	★
	News Title	★	State	★
	discussion	★		
DOCUMENT	Title	★	Description	★
	TLD	★		

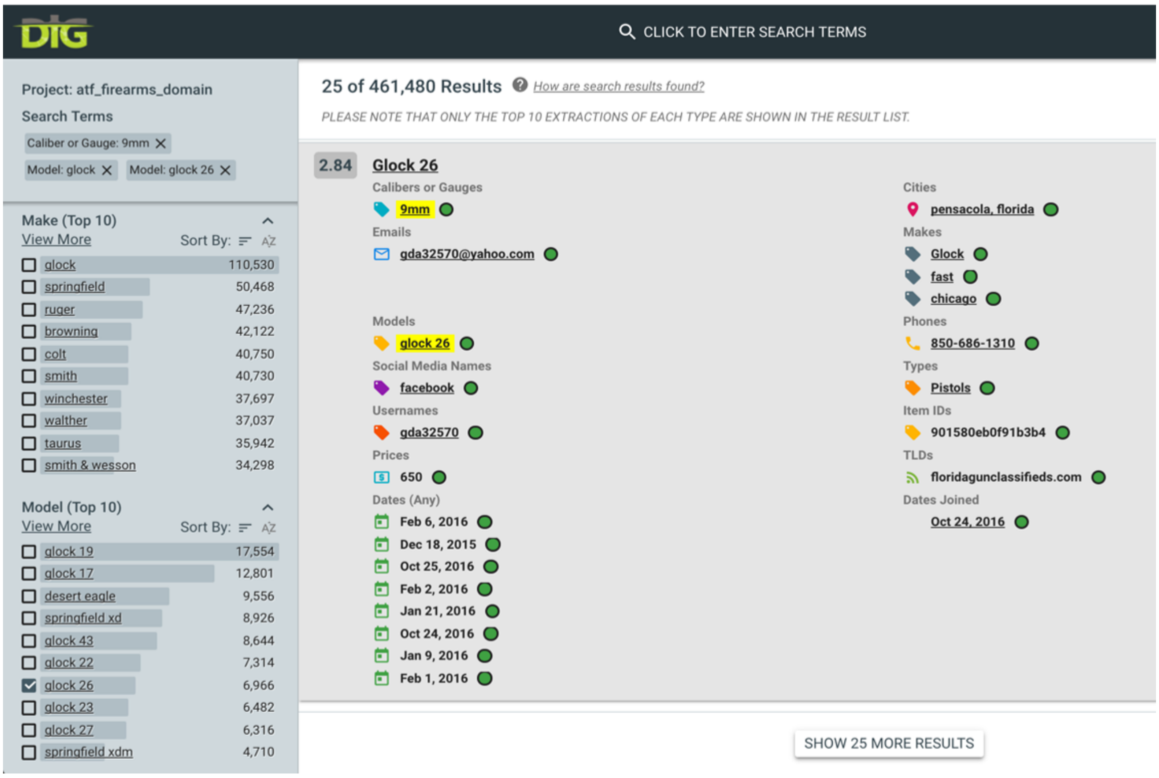
Figure 4. A form that allows the user to pose fine-grained, structured queries (*counterfeit electronics* domain).

3.2.2. Facets

As shown in Figure 5, myDIG supports faceted search and filtering on select fields (e.g., Model or Make in the figure (taken from the *Illegal Firearms Sales* (IFS) domain)) that the user can specify during domain setup. In all of our case studies, users made fairly intuitive choices: except for "free-form" fields like text, descriptions or comments, they favored faceting over non-faceting. In addition to allowing more informed search and filtering, facets also help the user to see an overview of the search results. For example, in the figure, one can deduce that (among the *glock* models) Models 19 and 17 occur far more often in the data than other models.

3.2.3. Entity-Centric Search and Summarization

Entity-Centric Search (ECS) and summarization is an important argument that distinguishes the domain exploration facilities of myDIG from the more generic Google search. An example is illustrated in Figure 6 for the entity *Glock 26*, which is presumably a firearm model that an investigator in the Illegal Firearms Sales (IFS) domain is interested in further investigating for suspicious transaction-related activity. The ECS dashboard summarizes the information about this entity by providing a (1) timeline of occurrences, (2) locations extracted from web pages from which the entity was extracted, (3) other entities co-occurring with the entity (along with *non co-occurrence* information to enable intuitive significance comparisons), and (4) relevant pages related to that entity. Users can declare the extractions of any field to be ECS-amenable by toggling the *Show as Links* option in the form in Figure 2.



CLICK TO ENTER SEARCH TERMS

Project: atf_firearms_domain

Search Terms

Caliber or Gauge: 9mm ✕ Model: glock ✕ Model: glock 26 ✕

Make (Top 10)
View More Sort By:

□ glock	110,530
□ springfield	50,468
□ ruger	47,236
□ browning	42,122
□ colt	40,750
□ smith	40,730
□ winchester	37,697
□ walther	37,037
□ taurus	35,942
□ smith & wesson	34,298

Model (Top 10)
View More Sort By:

□ glock 19	17,554
□ glock 17	12,801
□ desert eagle	9,556
□ springfield xd	8,926
□ glock 43	8,644
□ glock 22	7,314
☑ glock 26	6,966
□ glock 23	6,482
□ glock 27	6,316
□ springfield xdm	4,710

25 of 461,480 Results How are search results found?

PLEASE NOTE THAT ONLY THE TOP 10 EXTRACTIONS OF EACH TYPE ARE SHOWN IN THE RESULT LIST.

2.84 Glock 26

Calibers or Gauges
9mm ●

Emails
gda32570@yahoo.com ●

Models
glock 26 ●

Social Media Names
facebook ●

Usernames
gda32570 ●

Prices
650 ●

Dates (Any)
Feb 6, 2016 ●
Dec 18, 2015 ●
Oct 25, 2016 ●
Feb 2, 2016 ●
Jan 21, 2016 ●
Oct 24, 2016 ●
Jan 9, 2016 ●
Feb 1, 2016 ●

Cities
pensacola, florida ●

Makes
Glock ●
fast ●
chicago ●

Phones
850-686-1310 ●

Types
Pistols ●

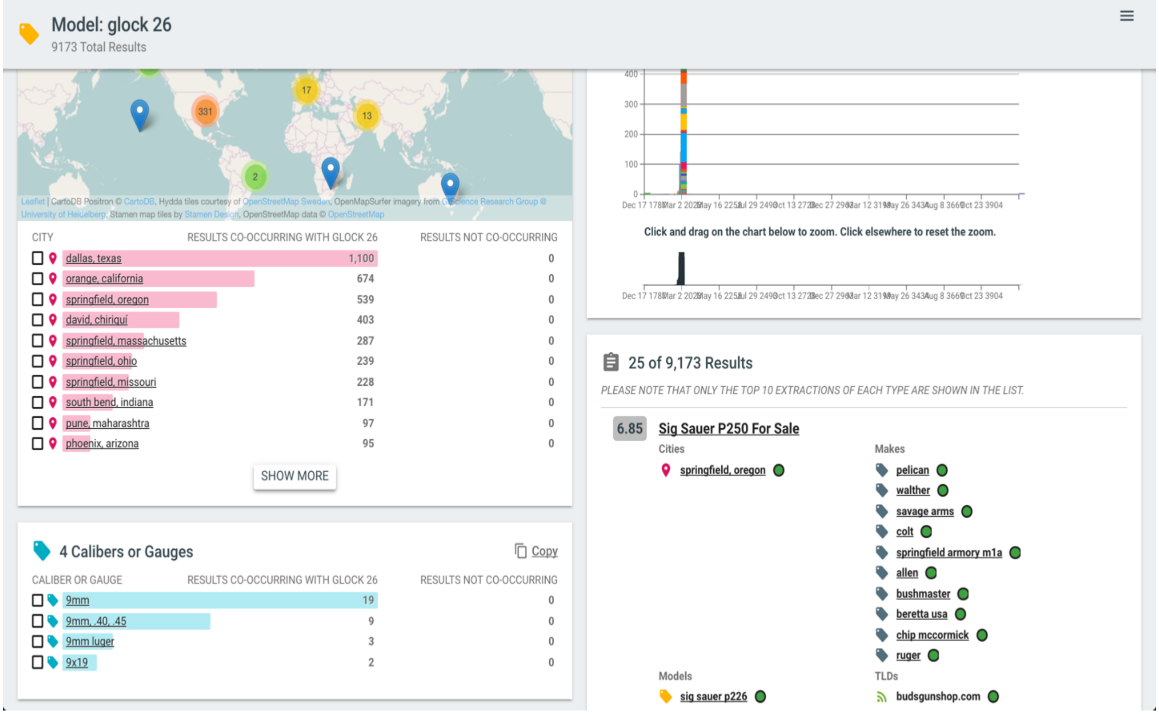
Item IDs
901580eb0f91b3b4 ●

TLDs
floridagunclassifieds.com ●

Dates Joined
Oct 24, 2016 ●

SHOW 25 MORE RESULTS

Figure 5. The search interface offered by myDIG (*illegal firearms sales* domain).



Model: glock 26
9173 Total Results

Leaflet | CartoDB Positron © CartoDB, Hydda tiles courtesy of OpenStreetMap Sweden, OpenMapSurfer imagery from GIScience Research Group @ University of Heidelberg, Stamen map tiles by Stamen Design, OpenStreetMap data © OpenStreetMap

CITY	RESULTS CO-OCCURRING WITH GLOCK 26	RESULTS NOT CO-OCCURRING
□ dallas, texas	1,100	0
□ orange, california	674	0
□ springfield, oregon	539	0
□ david, chiriquí	403	0
□ springfield, massachusetts	287	0
□ springfield, ohio	239	0
□ springfield, missouri	228	0
□ south bend, indiana	171	0
□ pune, maharashtra	97	0
□ phoenix, arizona	95	0

SHOW MORE

4 Calibers or Gauges Copy

CALIBER OR GAUGE	RESULTS CO-OCCURRING WITH GLOCK 26	RESULTS NOT CO-OCCURRING
□ 9mm	19	0
□ 9mm, .40, .45	9	0
□ 9mm luger	3	0
□ 9x19	2	0

Click and drag on the chart below to zoom. Click elsewhere to reset the zoom.

25 of 9,173 Results

PLEASE NOTE THAT ONLY THE TOP 10 EXTRACTIONS OF EACH TYPE ARE SHOWN IN THE LIST.

6.85 Sig Sauer P250 For Sale

Cities
springfield, oregon ●

Models
sig sauer p226 ●

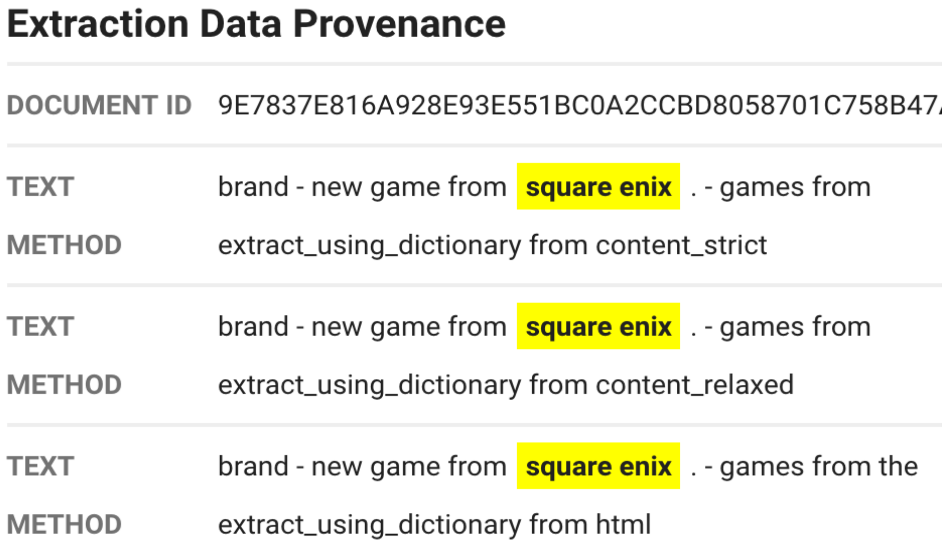
Makes
pelican ●
walther ●
savage arms ●
colt ●
springfield armory m1a ●
allen ●
bushmaster ●
beretta usa ●
chip mccormick ●
ruger ●

TLDs
budsgunshop.com ●

Figure 6. Entity-Centric Search (ECS) for the *illegal firearms sales* domain in myDIG.

3.2.4. Provenance

The interface supports provenance both at the (coarse-grained) level of web pages and the (fine-grained) level of extractions. Concerning the latter, provenance information is obtained by clicking on the green circle next to an extraction (see Figure 5), which brings up the specific extraction method such as Inferlink, SpaCy, glossary, etc. (Figure 7), and in the case of context-based extractors that use methods like word embeddings, the text surrounding the extraction. Multiple provenances are illustrated, as shown in the figure, if applicable (e.g., glossary extractions from text that originated in different structures in a web page). We also support web page-level provenance by allowing the user to open the cached (it is important to show the cached, rather than the "live" web page (which can also be shown by clicking on the pre-defined URL extraction that exists for every web page in the corpus) since the web page may have changed, or even been removed, since domain discovery) web page in a new tab.



Extraction Data Provenance

DOCUMENT ID	9E7837E816A928E93E551BC0A2CCBD8058701C758B47
TEXT	brand - new game from **square enix** . - games from
METHOD	extract_using_dictionary from content_strict
TEXT	brand - new game from **square enix** . - games from
METHOD	extract_using_dictionary from content_relaxed
TEXT	brand - new game from **square enix** . - games from the
METHOD	extract_using_dictionary from html

Figure 7. Provenance of glossary-based *organization name* extraction "square enix" (*securities fraud* domain).

3.2.5. Summary

User experience in myDIG is geared toward supporting personalized, domain-specific search, but without being constrained to a specific domain (or all domains simultaneously, like Google) a priori. The system operates in two phases, namely domain setup and curation and search, to allow a user to build a search engine over a domain-specific corpus, whether involving an investigative domain like firearms sales or an "ordinary" domain like soccer. Furthermore, the system does not make *ontological commitments*, but allows users to set up and customize their own "meta-data" (fields and importance). No coding or parameter tuning of any sort is required. Users were able to use myDIG on their own with less than an hour of example-based demonstrations, much of which was spent on familiarizing users with the facilities of the Inferlink tool and the search interface.

4. Algorithms and Technologies

From a systems-level perspective, myDIG involves a complex combination of technologies. We focus on the core elements that are critical for supporting an interactive user experience. While we aim to be specific about the actual technologies used and their rationale, we omit a full technical account. We also note that all myDIG technologies are available under GitHub under a permissive license.

4.1. Inferlink Wrapper

Inferlink was briefly described earlier as a key step in setting up the domain. Inferlink is based on years of research on wrapper-based web information extraction systems [16,17,45]. As explained earlier, it also involves web page clustering, using a variety of syntactic and structural features. The clustering is designed to be both robust and unsupervised; this is why users already see the clusters (with sample pages) when they first open a TLD in Inferlink. Once a cluster is picked by the user as being relevant, the wrapper algorithms in Inferlink extract structured elements from each HTML page in the cluster and show them to a user in a column layout, as illustrated in Figure 3. Other elements of user experience, including assigning field names to columns (denoted as the semantic typing step), were described earlier. The final output of Inferlink processing (per TLD) is a set of machine-processable rules that include automatically-generated regular expressions that identify the beginning and end of each extraction. These rules are executed on a server on every web page (from that TLD), including web pages not belonging to the initial training cluster that the user selected and curated. Heuristics are used to identify incorrect extractions from "out of cluster" pages. An important advantage of the rule-based execution is that, once the rules are stored on a server, they can be re-used across user sessions or even shared by different sets of users. Rule execution is also computationally efficient: while the clustering and rule inference steps are expensive (5 min/TLD), even complex Inferlink rule-sets are able to execute in milliseconds per web page. Inferlink continues to be actively maintained by a private company.

4.2. Cloud-Based Spark Workflow

Because of the iterative dataflow that users follow and their expectations of a speedy execution, serial processing was found to be too slow even for small samples of documents. This, coupled with the observation that each web page was processed independently of other web pages, led to the adoption of *Apache Spark* as our backend processing infrastructure [46]. The backend of myDIG is set up on a server in Amazon Web Services (AWS), which is configured both to run a workflow of Spark jobs and to report progress (Figure 1), so that users have near real-time feedback (at least one user expressed appreciation for this feature in qualitative feedback) on how long the workflow will take. We note that users are free to both continue setting up the domain using Inferlink and to explore the knowledge base populated by the previous iteration using the search interface, while waiting for the current iteration to finish.

In a similar vein, as new data are crawled and ingested by myDIG (after the domain has been set up), the Spark workflow can be executed at periodic intervals, similar to how a search engine's spider operates. In the human trafficking domain, for example, myDIG was refreshed with new data once every 24 h. For other domains, this period could be shortened or lengthened. Typically, "old" data (generally taken to mean a web page that has changed its content, or since been taken offline) are not thrown away by myDIG since our users have reported that such data can be useful for spotting trends or even producing evidence for prosecutions. An important agenda for future work is to build an integrated, real-time version of myDIG that is able to continuously crawl, analyze, index, and load domain-specific web pages.

4.3. NoSQL Knowledge Graph Representation

The output of the extraction workflow is stored as a *text-enriched knowledge graph* in an Elasticsearch NoSQL database instance [29]. Every document in an Elasticsearch index is a set of key-value pairs, formally serialized using the JSON file format [47]. We assign a permanent ID to every web page in the corpus (e.g., see the document ID in Figure 7), which serves as an internal means of identifying and storing each web page as a separate document. Keys in the document represent fields: by the nature of NoSQL, a key only exists if an extraction for that field (for that web page) exists. Depending on the field, the value is either atomic (like a string) or a list of atoms. With some minor exceptions (such as the document ID field), each field is indexed by Elasticsearch using the inverted index facilities of the Lucene information retrieval system that is used to support Elasticsearch queries and indices [29,48]. Finally, to support fast ECS (Figure 6), we also populate the NoSQL database with a set of *entity documents*, with the entity summaries contained within as key-value pairs.

We note that using a key-value document store such as Elasticsearch (or other rival NoSQL document stores such as MongoDB) is a design choice and not completely necessary, since the front-end queries (posed using the form) are initially formulated to SPARQL, before being reformulated as a forest of Elasticsearch Boolean tree queries. One could potentially represent and store the KGs in a triplestore such as Apache Jena (https://jena.apache.org/). At the time that myDIG was being implemented, such triplestores had two problems that precluded us from using them. First, they did not support an AI-centric query reformulation strategy such as what we intended for the system to have high recall and correctly deal with issues of noise in real-world KG construction. Second, and more fundamentally, there were issues with speed. Technologies like Elasticsearch are inspired by indexing research in the Information Retrieval (IR) community and are consequently able to retrieve and rank results very quickly. In contrast, triplestores can handle more sophisticated reasoning, but are slower. Since speed, rather than reasoning, was an important desideratum for our system to be adopted by real-world users, we chose to represent our KG as key-value documents. Similar arguments applied to graph databases like Neo4j (https://neo4j.com/). In future work, a re-implementation of the system could potentially include cloud-based graph databases with significantly improved speeds (an excellent recent example is Blazegraph, which is supported by Amazon Web Services: https://www.blazegraph.com/).

4.4. Search Engine

The myDIG search engine uses a robust set of algorithms to increase both recall and precision. Recall is important because user intent can sometimes be different from user input. Even if input closely represents intent, there are multiple such representations possible, not all of which will be directly observed in a web page. As a simple example, a user might search for *Microsoft Corporation*, whereas only *Microsoft* is present in the web page (or only *Microsoft* got extracted for a field like *Company*). Clearly, an *exact match* will not work in such situations, mainly because of data heterogeneity. However, handling recall should not come at the cost of decreasing the precision of responses. The search should be reasonably immune to wrong and missing extractions.

Based on our early experimentation conducted over more than a year, we found two techniques to be particularly useful for addressing these challenges in myDIG, namely *loosely-constrained search* and *selective boosting*. The latter is controlled by users, since users can assign importance values to fields when defining them or when modifying pre-defined fields (Figure 2). We use the importance values in our Elasticsearch-based query engine to boost matches on certain fields so that the score is *influenced more* by hits on those fields. Loosely-constrained search is based on *constraint relaxation* in the query reformulation literature [27,28]. The key idea is to "relax" the query (e.g., by also searching on text fields, in addition to the fields specified by the user on the search form) to maximize recall, but to lower the relevance scores of hits obtained by the relaxed queries (as opposed to hits obtained by the original, strict-semantics query). This ensures that retrieval is "smooth" in that the system is still able to retrieve information even when an extraction is missing or misplaced, as long as there is some signal in the text.

5. Evaluations and Case Studies

In the last six months, myDIG was evaluated by the DARPA MEMEXprogram in five different investigative domains, namely securities (specifically, penny stock) fraud, illegal firearms sales, illicit shipments via USPS mail, narcotics, and counterfeit electronics. Each of these domains is extremely specialized, but has common characteristics that make it particularly amenable to analysis in myDIG. When describing each domain, we focus on the types of investigative questions domain experts are looking to answer.

Concerning the evaluation protocol, to keep the user study as unbiased as possible, DARPA, NIST, and a contracted private firm conducted all evaluations (the one exception is a post-hoc evaluation of extraction quality that was conducted by researchers in our group) described herein and released a select set of results both to inform future work and to provide guidance on the current state and usability of the system. Based on a number of factors, the most important of which was the preference expressed by domain experts, not every result was released to our research group or can be published in the open research literature. By necessity, therefore, some evaluations are more detailed than others.

5.1. Users and Domains

The myDIG system was evaluated by five pairs of users, each of which was affiliated with either a federal agency or a national organization. The users are practicing experts in their respective domains. Because they involve highly illicit activities, the domains described below may also have a significant Dark Web presence, but in this paper, all data discussed, released, analyzed, or presented were crawled over the Open Web.

Concerning protocol, each pair of domain experts arrived on a separate day in the Washington D.C. facility reserved by DARPA for these evaluations. A member from our research team trained the users for 1–2 h in navigating the system, setting up the domain, and conducting search as we have described in this paper. Users were then left to set up their domains (Phase 1, as described in *System Overview and User Experience*) in myDIG. On a separate day (usually the next, but always within the week), users conducted actual knowledge discovery and search (Phase 2) by using the search interface for 2 h. All feedback was collected by NIST and DARPA after Phase 2 had concluded.

5.1.1. Securities Fraud

Securities, particularly *penny stock*, fraud is a complex domain that falls under the direct authority of the Securities and Exchange Commission (SEC) in the United States. Penny stock fraud is unusual because much of the activity that accompanies fraudulent behavior, including hype and promotional activity, is legally permitted. Many of the actual actors involved may not be physically present in the U.S., but for regulatory reasons, "shell" companies fronting such activity for promotional and legal purposes have to be registered in the U.S. to trade stocks legitimately in Over-The-Counter (OTC) exchanges. In addition to the longer term goal of investigating, and gathering information on, such shell companies and the people involved in them, investigators are also interested in taking preventive activity. This can happen when a penny stock company is caught actively engaging in factually fraudulent hype (for example, a false claim that a contract was just signed with a well-known customer firm), in which case, trading can be halted or even shut down (this is why the step is preventive; trading is shut down before unwary investors "buy in" and subsequently end up losing their savings). The myDIG system supports these goals by allowing users to aggregate information (in the crawled corpus, which contains many web domains) about suspicious penny stocks using the ECS facilities, and also to zero in on burgeoning promotional activity.

5.1.2. Illicit Shipments via USPS Mail

Unlike the securities fraud domain, illicit shipments via USPS happen in the physical world, but communications about the illicit shipment, particularly tracking numbers, happen digitally in specific forums and typologies that investigators in this domain understand well. Like the *Narcotics* domain, and unlike the other three domains, the USPS domain is an "under the cloak" domain about which investigators have not shared much information. After training, however, these investigators were able to set up the domain and use it on their own despite having no technical or programming abilities. While the actual search methodology and the questions that investigators were seeking to answer are strictly confidential and cannot be revealed, we were allowed to assess (and report on) user effort in setting up the domain and subsequent KG quality.

5.1.3. Illegal Firearms Sales

In the U.S., firearms sales are regulated in the sense that transactions cannot be conducted with arbitrary persons or over arbitrary channels like the Internet. Investigators in the IFS domain are interested in pinpointing activity that, either directly or indirectly, provides evidence for illicit sales that leave some digital trace. The domain is similar to the SF domain (and dissimilar to the CE domain, described below) for the important reason that investigators limit their focus to domestic activities.

5.1.4. Counterfeit Electronics

Despite what the name suggests, investigators in the counterfeit electronics domain are interested, not in consumer electronics, but in microchips and FPGAs that form the computational backbones of more complex "application" devices. The FPGAs may resemble an FPGA from a genuine contractor, but are fakes, and may have malicious modifications at the hardware level. Certain countries, companies, and devices are more relevant to this kind of activity than others. We note that there is an obvious national security component to these investigations, and just like with the other described domains, domain expertise plays a crucial role both in setting up the domain and in the knowledge discovery itself.

5.1.5. Narcotics

Due to confidentiality constraints, there is little that we can reveal about the narcotics domain, except the self-explanatory implication in the name itself. Like the other domains, it is important to note the caveat that investigators are specifically interested in activity that has some digital component that can guide further investigation. That is, it does not purport to cover all narcotics activity. Among all five domains, this domain is featured most prominently in the Dark Web data.

5.2. Domain Discovery Corpora

Figure 8 characterizes the corpus for each domain. From the roughly power-law distributions seen for most of the domains (near-linearity on log-log scale), despite their diversity, we note that some TLDs in each distribution are much more heavily represented than others. In turn, this immediately suggests why a tool such as Inferlink (that requires users to curate on a per-TLD basis) could be useful. By focusing on the *largest TLDs*, users can end up curating a large fraction of web pages. On the other hand, there is a significant *long tail*, which means that Inferlink (which is best suited for the short tail or largest TLDs) cannot be the only solution for recall-friendly knowledge discovery. Glossaries, as well as pre-defined extractions, are required to take up the slack. There is a clear tradeoff between user effort and knowledge graph quality. In the evaluations described subsequently, we measure both.

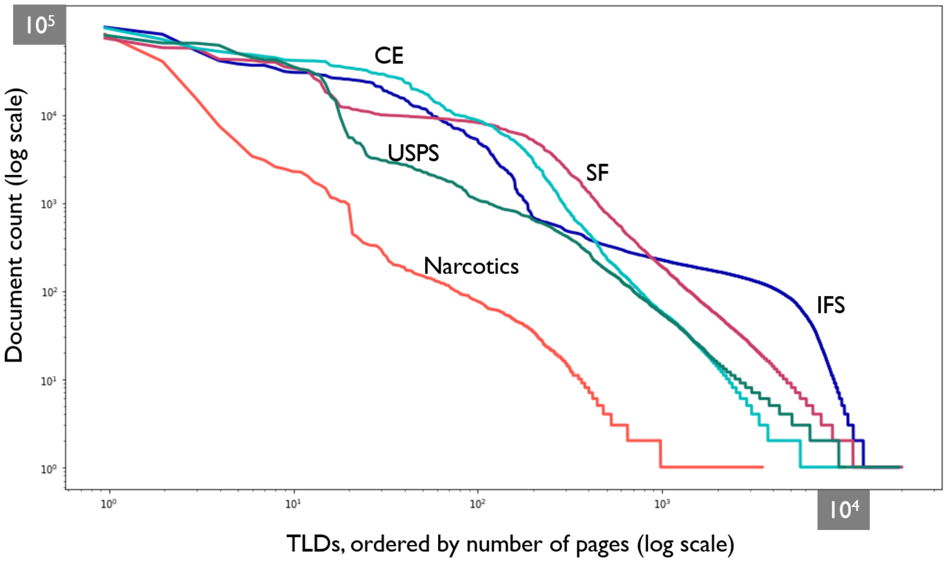


Figure 8. Document counts per TLD (Y-axis) against the ranked list of TLDs (ordered by document count on the X-axis).

5.3. Evaluation of User Effort

5.3.1. Tasks and Materials

We evaluated user effort in setting up the domain in myDIG in two different ways. First, we evaluated the level of effort that was expended in each domain on using the *Inferlink* tool to curate one or more relevant (as judged by the users) clusters of web pages. Since Inferlink operates on a per-TLD (more specifically, a per-*cluster per TLD*), we refer to its extractions as *short tail* extractions. We plot the short tail against the long tail, by reporting the numbers of web pages for both cases per field such that the web page had at least one extraction (from Inferlink for the short tail and overall, i.e., any extraction at all, for the long tail) for that field. Second, we evaluated the level of user effort in defining the domain by reporting the total number of fields for each domain, with some representative examples of fields that users defined from scratch. We also quantified the number of glossaries used per domain, since qualitative feedback by users indicated that glossaries were extremely popular in incentivizing users to set up custom fields.

5.3.2. Results

Figures 9–13 illustrates the distribution of the numbers of web pages with at least one extraction (per field, as noted on the y-axis) both for the short and long tails. The figures illustrate two interesting aspects quantifying the importance of expending effort in the Inferlink tool vs. more all-encompassing methods like glossaries. In the IFS and SF domains, we find that Inferlink has made significant fractional contributions to some fields (at the bottom of the y-axis) compared to domains like narcotics, where the long tail overwhelms Inferlink. On the other hand, we find that, for some fields, only Inferlink has yielded any extractions at all. For example, in the CE domain, fields like *UserRole* and *Product* had non-zero extraction numbers only because of Inferlink. In post-hoc qualitative feedback, users expressed that using Inferlink gave them a means of expending effort with the confidence that the effort yields usable, interpretable results.

Results in Table 1 show that between 20 and 40 fields were retained (from the generic set of pre-defined fields, like city or state) or defined by users. Some domains are much more fine-grained than others, an example being N. Users also made good use of the glossary facility, with the number of glossaries (eight) being the highest for the IFS domain.

Table 1. Quantifying user effort in terms of total numbers of fields and glossaries per domain, after the users had finished setting up their respective domains.

Domain	Number of Fields	Number of Glossaries	Examples (User-Defined Fields)
IFS	27	8	caliber_gauge, ffl_status
CE	21	6	PartNumber, Company
USPS	21	5	confirmations, Tracking_Number
N	43	5	used_bitcoin, vendor
SF	32	6	counsel, disclaimer

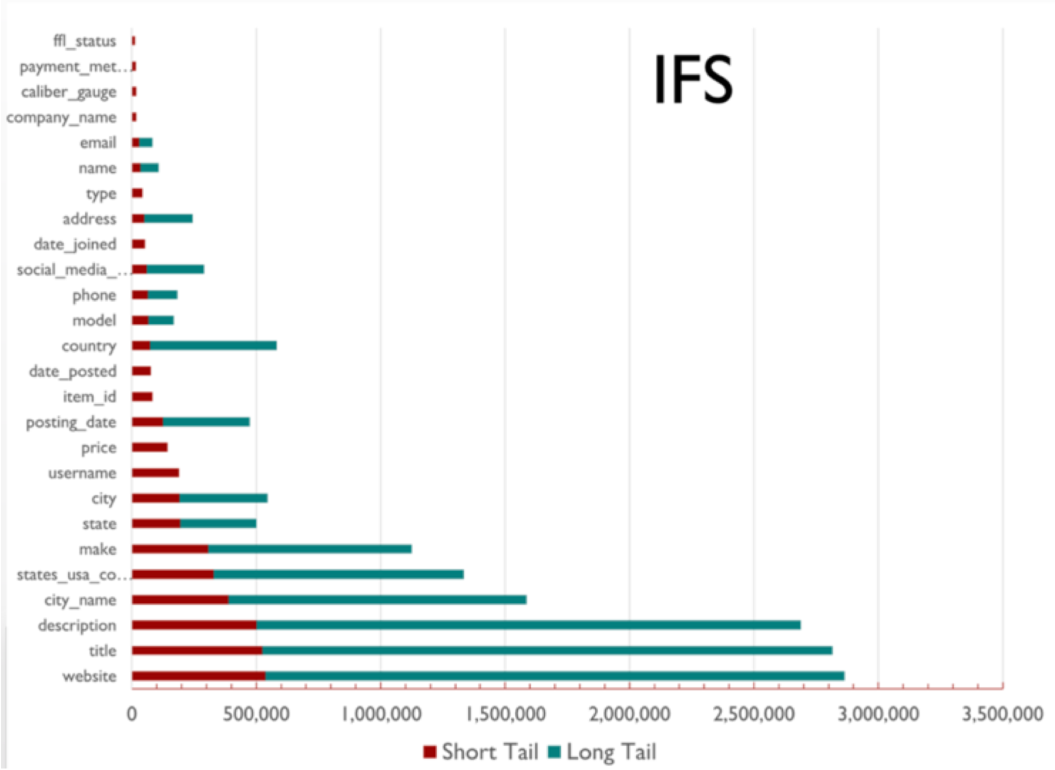


Figure 9. For the IFS domain, a comparison of web page count (x-axis) such that at least one value per field (y-axis) was extracted from that website (green bar/long tail), against the short tail/red bar wherein only Inferlink extractions are considered.

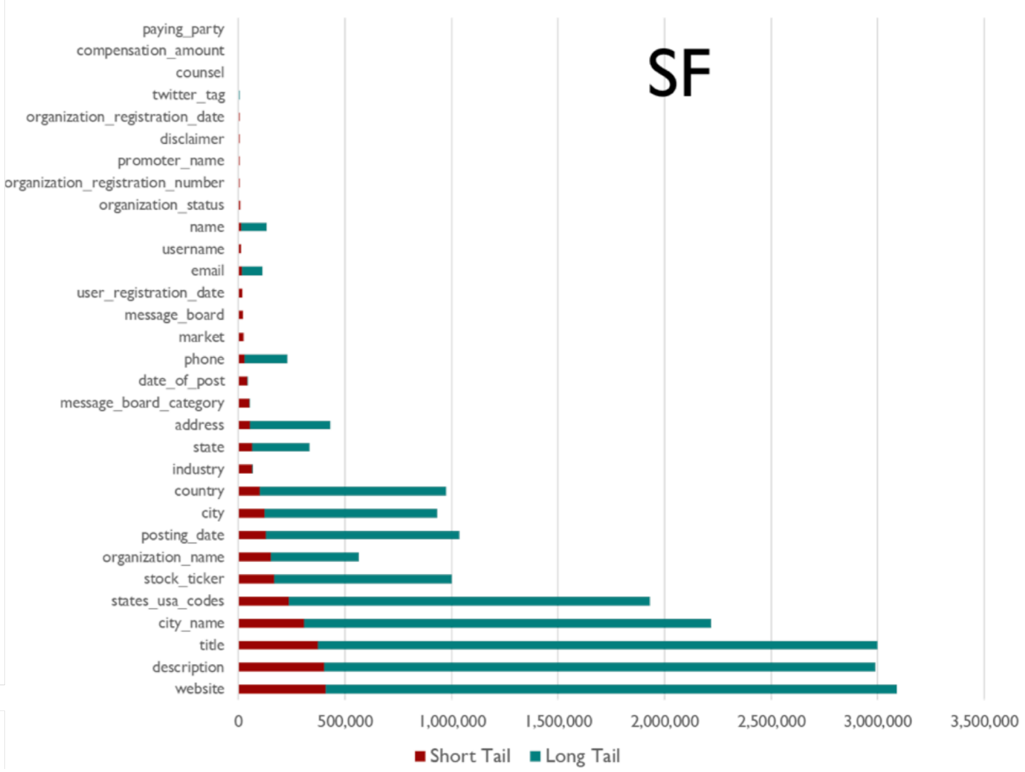


Figure 10. For the SF domain, a comparison of web page count (x-axis) such that at least one value per field (y-axis) was extracted from that website (green bar/long tail), against the short tail/red bar wherein only Inferlink extractions are considered.

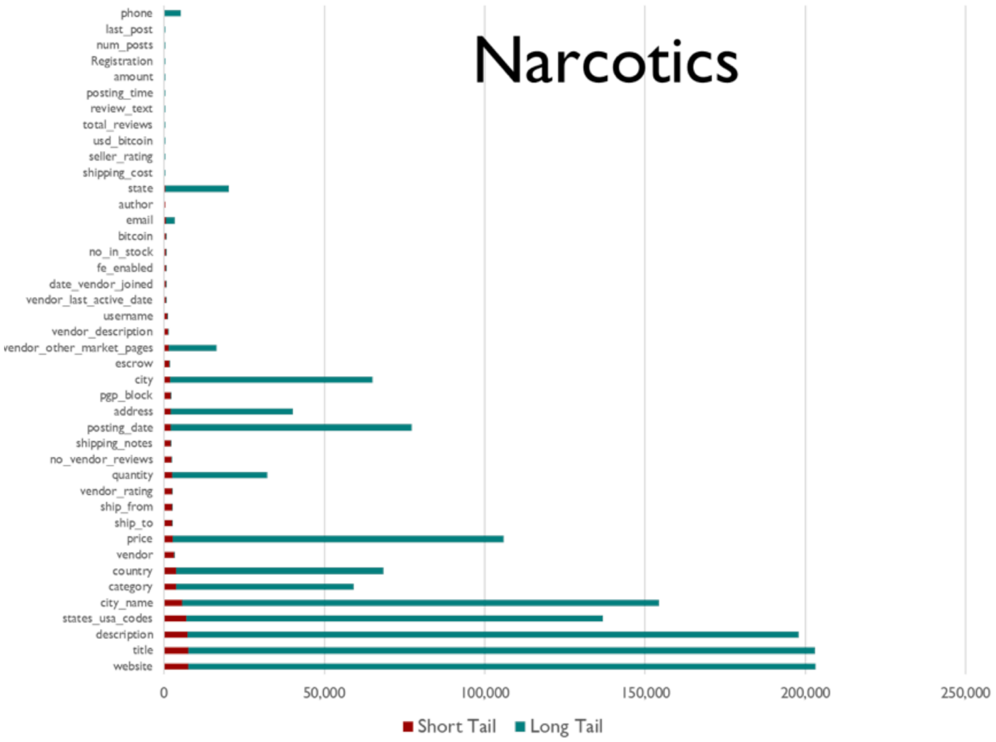


Figure 11. For the Narcotics domain, a comparison of web page count (x-axis) such that at least one value per field (y-axis) was extracted from that website (green bar/long tail), against the short tail/red bar wherein only Inferlink extractions are considered.

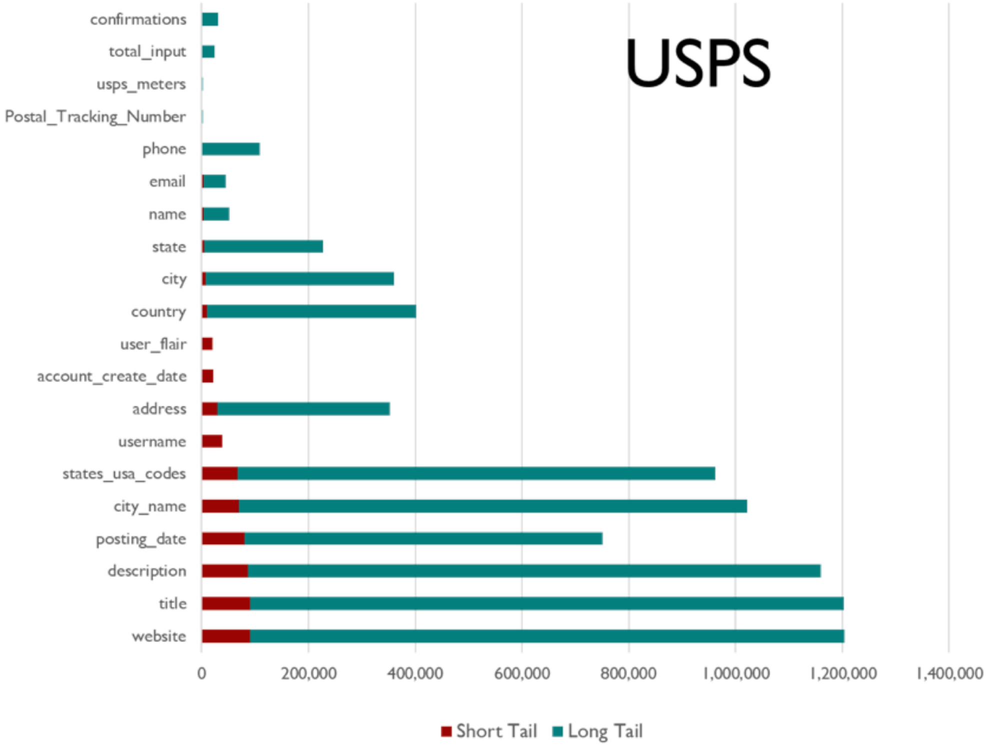


Figure 12. For the USPS domain, a comparison of web page count (x-axis) such that at least one value per field (y-axis) was extracted from that website (green bar/long tail), against the short tail/red bar wherein only Inferlink extractions are considered.

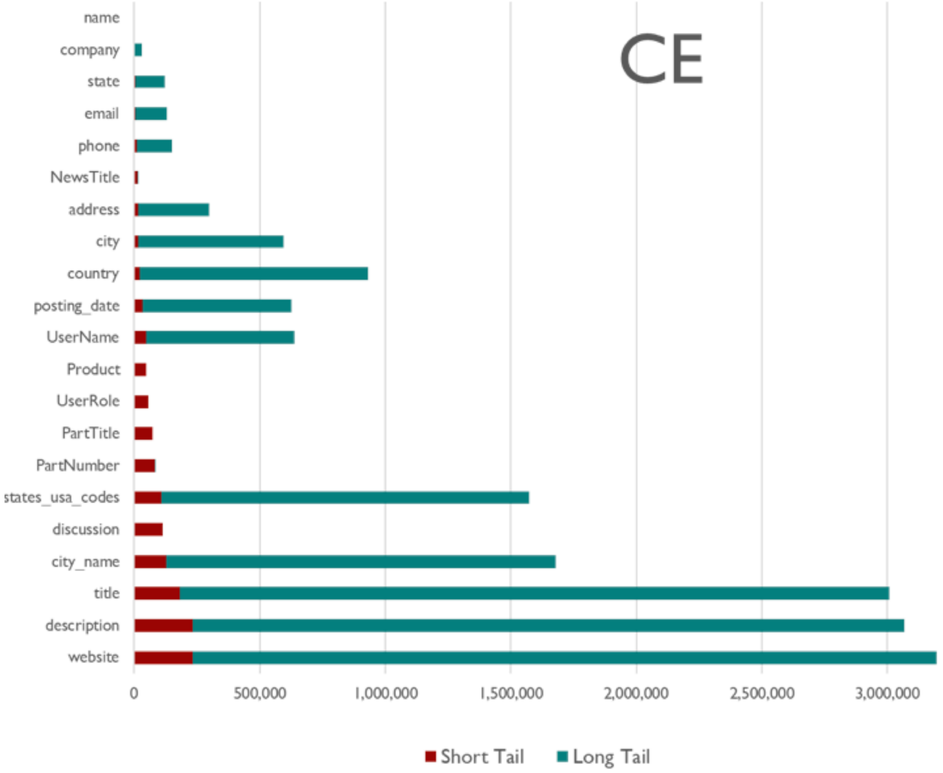


Figure 13. For the CE domain, a comparison of web page count (x-axis) such that at least one value per field (y-axis) was extracted from that website (green bar/long tail), against the short tail/red bar wherein only Inferlink extractions are considered.

5.4. Evaluation of Knowledge Graph Quality

5.4.1. Tasks and Materials

The aim of this evaluation was to assess the quality of the KG; specifically, the precision of the extractions for fields in the KG. We conducted this task by sampling 25 random web pages per domain and assigning each domain to an annotator from our research group. Each extracted element (for each field) reported by myDIG was given a 0/1 score for correctness. Each annotator made this assessment by referring back to the original cached web page to verify whether the answer was correct. We report these precision metrics, along with the total numbers of extractions evaluated for correctness, and comment on the results.

5.4.2. Results

Results in Table 2 show that myDIG was able to achieve reasonable levels of precision for the collective sets of extractions. For some domains, like CE and USPS, precision was close to (or above) 90%, while for others, it tended to be between 70 and 80%. While there is room for improvement, these precision numbers show that, on average, there was considerably more signal than noise in facets and extractions.

Table 2. Manually-judged precision of field extractions stored in the knowledge graph for each domain.

Domain	Number of Extractions	Correct	Precision
IFS	168	125	74%
CE	85	75	89%
USPS	116	109	94%
N	211	156	74%
SF	67	48	72%

6. Qualitative Feedback and Analysis

Much feedback was elicited during the testing of myDIG from users of all five domains. In the interest of space and the contributions that we claimed in the introductory section of this paper, we focus on how myDIG helped users obtain insights that were otherwise not achievable using generic search technology like Google.

6.1. Insights beyond Google

In response to the question, *did the tool ever help you achieve an insight that would have been otherwise difficult to identify without the tool? If so how?* posed by NIST evaluators, users had the following comments (comments are not syntactically modified in any way):

[IFS] Yes, it allowed us to see possible connections to phone numbers, email addresses, locations, etc., that could not necessarily be verified/discovered using Google or site-led searches within Armslist, backpage, etc.

[CE] SAVES SO MUCH TIME. Aggregates data that would be searched using Google.

Comments from SF users revealed, however, that the actual corpus crawled can sometimes be an important bottleneck:

[SF] Conceptually, maybe it would, but for the use cases proposed by SEC, there were no insights found; data extraction/crawling issues.

6.2. Productivity

Users also felt that myDIG improved their productivity and helped them to focus on relevant information. In response to the question, *did the tool identify any new and relevant sources of information that you had not found through other methods previously? If so, what were they?* users had the following comments:

[IFS] It was able to categorize postings by location, which allowed us to focus our investigation.

[CE User 1] -Spartan6 xlinx (email addresses); -domain search and aggregation on email addresses; -filling in relevant domain endings to match (while hard to interpret post-hoc, we believe these indicate relevant information sets discovered by User 1).

However, once again, the crawled data were noted to be an issue by both SF and CE User 2.

[SF] No; index properly; Investorshub was not indexed correctly: some posts are crawled and scraped correctly, most are not.

[CE User 2] Was really unhappy with data that was pulled in by the tool. A better (more efficient) domain discovery process...

Despite the comments by SF users, we note that they explicitly indicated to us that they would like the system to be transitioned to their office and that they would like to keep using myDIG to curate their domain and applying it in actual on-the-ground investigations.

6.3. Useful Features

Users were also asked the following: *were there any features within the tool that stood out as particularly useful?* Most users expressed that the glossary feature was very useful, and ECS was also singled out by some; specific responses are noted below:

[CE] -Really liked the glossary feature; -Can have multiple domains; -Inferlink integration is Really straightforward; -Status bar was helpful to know when a process is running vs. when it is not; -subset of TLDs helpful.

[SF] -Being able to 'click' on the entity; -documents that relate to an address; find relevant information associated with a particular entity.

7. Discussion: Implications for Future Design

Based on the results and analyses in the previous sections, we summarize some important implications uncovered over the course of testing myDIG in five different domains.

Automatic data acquisition is the new bottleneck. Many, if not most, of the users expressed satisfaction with the facilities offered by myDIG and felt that the potential of myDIG would have been more fully realized if the quality of the data ingested had been better. Unfortunately, we had no control over this issue, although the DARPA MEMEX program used state-of-the-art data acquisition, relevance modeling, and crawling systems. One area of future work currently being considered is the integration of domain discovery and knowledge discovery. While such an integration would introduce new complexities and dependencies into the system, it could potentially improve data acquisition quality by incorporating richer user feedback into the domain discovery process.

Users should be empowered to specify their own vocabularies. In the AI literature, there are powerful data models and representations for specifying fine-grained ontologies [11,49,50]. These have merits in certain fields, such as biomedicine, but are not suitable for users who do not explicitly deal with vocabularies, ontologies, or data models on a regular basis [51]. Instead, systems like myDIG allow users to crystallize their internal vocabularies by exploring the data and specifying fields as they go along; for example, many of our users interleaved Inferlink annotations while defining new fields. This methodology is in sharp contrast to the more traditional view of specifying a domain model upfront before data acquisition, processing, or modeling begins.

Provenance, going all the way to the original source, is key to fostering trust. We are, by no means, the first to make this observation, but believe that it is important to (re-)state owing to the rapid progress in deep learning technologies for information extraction, without the same progress in interpreting or explaining the results. Users expressed trust in our system because of the provenance offered both at the web page and extraction levels. We note that, recently, the AI community has recognized this issue and is tackling the explainability problem with vigor.

8. Conclusions and Future Work

There is a need to democratize and personalize machine learning and search technologies as they become ever more complicated, such that users and analysts in specific domains can interact with these tools without sacrificing the benefits of years of domain expertise. The myDIG knowledge discovery system is an effort funded under the DARPA MEMEX program to facilitate this high-level goal. User experience while using myDIG can be roughly separated into two phases: setting up the domain and conducting search. The myDIG system does not require specifying vocabularies or making ontological commitments in advance and is not specialized for a single domain. More specifically, in all five investigative domains where myDIG was evaluated, users were able to build a personalized, domain-specific search engine over corpora that contained hundred of thousands, and in some cases, millions, of raw HTML web pages.

Feedback from investigative users shows that the system has potential for tackling some difficult questions, with the quality of domain discovery and data acquisition proving to be an important bottleneck that we are looking to address in future work. Most encouragingly, users (particularly in the securities fraud domain) have explicitly indicated to us that they would like to keep using the system and would like the system to be transitioned to them.

In keeping with the democratic principles espoused in this work, and by the DARPA MEMEX program, myDIG was packaged, documented, and released to the public in November, 2017 (with regular updates), shortly after the MEMEX program held its final Quarterly Period Review (QPR). We are already experimenting with myDIG in the classroom and laboratory settings and using it to explore domains other than the investigative domains described in this paper. A specialized instance of myDIG has been transitioned to a state-level law enforcement agency for investigating crimes such as human trafficking.

Author Contributions: Conceptualization, M.K. and P.S.; methodology, P.S.; software, M.K.; validation, M.K. and P.S.; formal analysis, M.K.; investigation, M.K.; resources, P.S.; data curation, M.K.; writing, original draft preparation, M.K.; writing, review and editing, P.S. and M.K.; visualization, P.S. and M.K.; supervision, P.S. and M.K.; project administration, P.S.; funding acquisition, P.S.

Funding: This research is supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under Contract Number FA8750- 14-C-0240. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

Acknowledgments: We gratefully acknowledge our collaborators and all (former and current) members of our team who contributed their efforts and expertise to DIG, particularly during the dry and final evaluation runs: Amandeep Singh, Linhong Zhu, Lingzhe Teng, Nimesh Jain, Rahul Kapoor, Muthu Rajendran R. Gurumoorthy, Sanjay Singh, Majid Ghasemi Gol, Brian Amanatullah, Craig Knoblock, and Steve Minton.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DIG	Domain-specific Insight Graphs
myDIG	my Domain-specific Insight Graphs
HTML	HyperText Markup Language
BI	Business Intelligence
KG	Knowledge Graph
SEC	Securities and Exchange Commission
IR	Information Retrieval
OTC	Over-The-Counter
KGC	Knowledge Graph Construction
AI	Artificial Intelligence
IE	Information Extraction
NER	Named Entity Recognition
TLD	Top-Level Domain
ECS	Entity-Centric Search
IFS	Illegal Firearms Sales
AWS	Amazon Web Services
DARPA	Defense Advanced Research Projects Agency
NIST	National Institute of Standards and Technology
CE	Counterfeit Electronics
FPGA	Field-Programmable Gate Array
SF	Securities Fraud

References

1. Elbashir, M.Z.; Collier, P.A.; Davern, M.J. Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *Int. J. Account. Inf. Syst.* **2008**, *9*, 135–153. [CrossRef]
2. Olszak, C.M.; Ziemba, E. Approach to building and implementing business intelligence systems. *Interdiscip. J. Inf. Knowl. Manag.* **2007**, *2*, 135–148.
3. Veak, T.J. *Democratizing Technology: Andrew Feenberg's Critical Theory Of Technology*; Suny Press: Albany, NY, USA, 2012.
4. Tanenbaum, J.G.; Williams, A.M.; Desjardins, A.; Tanenbaum, K. Democratizing technology: Pleasure, utility and expressiveness in DIY and maker practice. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; ACM: New York, NY, USA, 2013; pp. 2603–2612.
5. Lakkaraju, K.; Yurcik, W.; Lee, A.J. NVisionIP: Netflow visualizations of system state for security situational awareness. In Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security, Washington, DC, USA, 29 October 2004; ACM: New York, NY, USA, 2004; pp. 65–72.

6. Goldstein, J.I.; Ramshaw, P.D.; Ackerson, S.B. An Investment Masquerade: A Descriptive Overview of penny stock fraud and the federal securities laws. *Bus. Lawyer* **1992**, *47*, 773–835.
7. Lavrenko, V.; Allan, J.; DeGuzman, E.; LaFlamme, D.; Pollard, V.; Thomas, S. Relevance models for topic detection and tracking. In Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, USA, 24–27 March 2002; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2002; pp. 115–121.
8. Niu, F.; Zhang, C.; Ré, C.; Shavlik, J. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *Int. J. Semant. Web Inf. Syst. (IJSWIS)* **2012**, *8*, 42–73. [CrossRef]
9. Craven, M.; DiPasquo, D.; Freitag, D.; McCallum, A.; Mitchell, T.; Nigam, K.; Slattery, S. Learning to construct knowledge bases from the World Wide Web. *Artif. Intell.* **2000**, *118*, 69–113. [CrossRef]
10. Singhal, A. Introducing the Knowledge Graph: Things, Not Strings. Available online: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/ (accessed on 27 February 2019).
11. World Wide Web Consortium. RDF 1.1 Concepts and Abstract Syntax 2014. Available online: https://www.w3.org/TR/rdf11-concepts/ (accessed on 27 February 2019).
12. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 28–37. [CrossRef]
13. Sarawagi, S. Information extraction. *Found. Trends® Databases* **2008**, *1*, 261–377. [CrossRef]
14. Aggarwal, C.C.; Zhai, C. *Mining Text Data*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
15. Chang, C.H.; Kayed, M.; Girgis, M.R.; Shaalan, K.F. A survey of web information extraction systems. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 1411–1428. [CrossRef]
16. Kushmerick, N.; Weld, D.S.; Doorenbos, R. *Wrapper Induction for Information Extraction*; University of Washington: Seattle, WA, USA, 1997.
17. Flesca, S.; Manco, G.; Masciari, E.; Rende, E.; Tagarelli, A. Web wrapper induction: A brief survey. *AI Commun.* **2004**, *17*, 57–61.
18. Riloff, E.; Jones, R. *Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping*; AAAI/IAAI: Menlo Park, CA, USA, 1999; pp. 474–479.
19. Rebele, T.; Suchanek, F.; Hoffart, J.; Biega, J.; Kuzey, E.; Weikum, G. YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In Proceedings of the International Semantic Web Conference, Kobe, Japan, 17–21 October 2016; pp. 177–185.
20. Fagin, R.; Kimelfeld, B.; Reiss, F.; Vansummeren, S. A Relational Framework for Information Extraction. *ACM SIGMOD Rec.* **2016**, *44*, 5–16. [CrossRef]
21. Niu, F.; Zhang, C.; Ré, C.; Shavlik, J.W. DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS* **2012**, *12*, 25–28.
22. SpaCy Natural Language Package. 2017. Available online: https://SpaCy.io/ (accessed on 19 September 2017).
23. Finkel, J.R.; Grenager, T.; Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd Annual Meeting on Association For Computational Linguistics, Ann Arbor, MI, USA, 25–30 June 2005; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 363–370.
24. Ratner, A.J.; De Sa, C.M.; Wu, S.; Selsam, D.; Ré, C. Data programming: Creating large training sets, quickly. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3567–3575. [PubMed]
25. Rieh, S.Y.; Xie, H. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Inf. Process. Manag.* **2006**, *42*, 751–768. [CrossRef]
26. Viswanathan, A.; Michaelis, J.R.; Cassidy, T.; de Mel, G.; Hendler, J. In context query reformulation for failing SPARQL queries. *Proc. SPIE* **2017**, *10190*, 101900M.
27. Muslea, I. Machine learning for online query relaxation. In Proceedings of the Tenth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, Seattle, WA, USA, 22–25 August 2004; ACM: New York, NY, USA, 2004; pp. 246–255.
28. Mirzadeh, N.; Ricci, F.; Bansal, M. Supporting user query relaxation in a recommender system. In Proceedings of the Ec-Web, Zaragoza, Spain, 31 August–3 September 2004; Springer: Berlin, Germany, 2004; pp. 31–40.
29. Gormley, C.; Tong, Z. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*; O'Reilly Media, Inc.: Newton, MA, USA, 2015.

30. Han, J.; Haihong, E.; Le, G.; Du, J. Survey on NoSQL database. In Proceedings of the 2011 6th International Conference on Pervasive Computing and Applications (ICPCA), Port Elizabeth, South Africa, 26–28 October 2011; pp. 363–366.
31. Amitay, E.; Carmel, D.; Golbandi, N.; Har'el, N.Y.; Ofek-Koifman, S.; Yogev, S. Information Retrieval with Unified Search Using Multiple Facets. U.S. Patent 8,024,324, 20 September 2011.
32. Banks, K.; Hersman, E. FrontlineSMS and Ushahidi-a demo. In Proceedings of the 2009 International Conference on Information and Communication Technologies and Development (ICTD), Doha, Qatar, 17–19 April 2009; p. 484.
33. Jadhav, A.S.; Purohit, H.; Kapanipathi, P.; Anantharam, P.; Ranabahu, A.H.; Nguyen, V.; Mendes, P.N.; Smith, A.G.; Cooney, M.; Sheth, A.P. Twitris 2.0: Semantically Empowered System for Understanding Perceptions From Social Data. Available online: https://works.bepress.com/amit_sheth/284/ (accessed on 27 February 2019).
34. Abel, F.; Hauff, C.; Houben, G.J.; Stronkman, R.; Tao, K. Twitcident: Fighting fire with information from social web streams. In Proceedings of the 21st International Conference on World Wide Web, Lyon, France, 16–20 April 2012; ACM: New York, NY, USA, 2012; pp. 305–308.
35. Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; Vieweg, S. AIDR: Artificial intelligence for disaster response. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; ACM: New York, NY, USA, 2014; pp. 159–162.
36. Rogstadius, J.; Vukovic, M.; Teixeira, C.; Kostakos, V.; Karapanos, E.; Laredo, J.A. CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM J. Res. Dev.* **2013**, *57*, 4:1–4:13. [CrossRef]
37. Kumar, S.; Barbier, G.; Abbasi, M.A.; Liu, H. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM), Barcelona, Spain, 17–21 July 2011.
38. Choi, S.; Bae, B. The real-time monitoring system of social big data for disaster management. In *Computer Science and Its Applications*; Springer: Berlin, Germany, 2015; pp. 809–815.
39. Thom, D.; Krüger, R.; Ertl, T.; Bechstedt, U.; Platz, A.; Zisgen, J.; Volland, B. Can twitter really save your life? A case study of visual social media analytics for situation awareness. In Proceedings of the 2015 IEEE Pacific Visualization Symposium (PacificVis), Hangzhou, China, 14–17 April 2015; pp. 183–190.
40. Gao, T.; Hullman, J.R.; Adar, E.; Hecht, B.; Diakopoulos, N. NewsViews: An Automated Pipeline for Creating Custom Geovisualizations for News. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14), Toronto, ON, Canada, 26 April–1 May 2014; ACM: New York, NY, USA, 2014; pp. 3005–3014. [CrossRef]
41. Krishnamurthy, Y.; Pham, K.; Aécio, S.; Freire, J. Interactive Exploration for Domain Discovery on the Web. In Proceedings of the KDD IDEA, San Francisco, CA, USA, 14 August 2016.
42. Liu, L.; Peng, T.; Zuo, W. Topical Web Crawling for Domain-Specific Resource Discovery Enhanced by Selectively using Link-Context. *Int. Arab J. Inf. Technol. (IAJIT)* **2015**, *12*, 196–204.
43. Lopez, L.A.; Duerr, R.; Khalsa, S.J.S. Optimizing apache nutch for domain specific crawling at large scale. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 1967–1971.
44. Ramnandan, S.K.; Mittal, A.; Knoblock, C.A.; Szekely, P. Assigning semantic labels to data sources. In Proceedings of the European Semantic Web Conference, Portoroz, Slovenia, 31 May–4 June 2015; Springer: Berlin, Germany, 2015; pp. 403–417.
45. Inferlink. 2017. Available online: http://www.inferlink.com/ (accessed on 19 September 2017).
46. Zaharia, M.; Xin, R.S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; Franklin, M.J.; et al. Apache Spark: A unified engine for big data processing. *Commun. ACM* **2016**, *59*, 56–65. [CrossRef]
47. Crockford, D. *The Application/json Media Type for Javascript Object Notation (Json)*; The Internet Society: Reston, VA, USA, 2006.
48. McCandless, M.; Hatcher, E.; Gospodnetic, O. *Lucene in Action: Covers Apache Lucene 3.0*; Manning Publications Co.: Shelter Island, NY, USA, 2010.
49. Knublauch, H.; Fergerson, R.W.; Noy, N.F.; Musen, M.A. The Protégé OWL plugin: An open development environment for semantic web applications. In Proceedings of the International Semantic Web Conference, Hiroshima, Japan, 7–11 November 2004; Springer: Berlin, Germany, 2004; Volume 3298, pp. 229–243.

50. McGuinness, D.L.; Van Harmelen, F. OWL web ontology language overview. *W3C Recomm.* **2004**, *10*, 2004.
51. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25. [CrossRef] [PubMed]