SYM-AM-18-080

PROCEEDINGS
OF THE
FIFTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

THURSDAY SESSIONS
VOLUME II

Acquisition Research:
Creating Synergy for Informed Change

May 9–10, 2018

March 30, 2018

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL

Utilizing Public Data for Data Enhancement and Analysis of Federal Acquisition Data

Ningning Wu—is Professor of Information Science at the University of Arkansas at Little Rock. She received a BS and an MS degree in Electrical Engineering from the University of Science and Technology of China and a PhD in Information Technology from George Mason University. Dr. Wu's research interests are data mining, network and information security, and information quality. She holds certificates of the IAIDQ Information Quality Certified Professional (IQCP) and the SANS GIAC Security Essentials Certified Professional. [nxwu@ualr.edu]

Richard Wang—is Director of the MIT Chief Data Officer and Information Quality Program. He is also the Executive Director of the Institute for Chief Data Officers (iCDO) and Professor at the University of Arkansas at Little Rock. From 2009 to 2011, Wang served as the Deputy Chief Data Officer and Chief Data Quality Officer of the U.S. Army. He received his PhD in information Technology from the MIT Sloan School of Management in 1985. [rwang@mit.edu]

M. Eduard Tudoreanu—is Professor of Information Science at University of Arkansas Little Rock. Professor Tudoreanu has expertise in human-computer interaction, information quality, advanced visualization of complex data, and virtual reality. He worked on visual data analysis, and has extensive experience in software development and user interface design. Professor Tudoreanu was the founding Technical Director of the Emerging Analytics Center. He has been the keynote speaker at ABSEL 2010, and served as a panelist for the National Science Foundation and Missouri EPSCoR. He earned his Doctor of Science degree in Computer Science in 2002 from the Washington University in St. Louis. [metudoreanu@ualr.edu]

Abstract

It is challenging to standardize data; yet, the capabilities to draw upon data across information systems hold huge potential for improving defense acquisition and procurement. Acquisition planning and management involves many decision-making and action-taking processes that cover a complex environment including actual acquisition, contracting, fiscal, legal, personnel, and regulatory requirements. A sound decision-making process has to rely on data—high quality data. Often the available data is dirty, outdated, incomplete, or insufficient for the expert to make a decision. On the other hand, there are enormous amounts of data on the web that can be utilized to crystalize the needed information. These data repositories are often publicly accessible and from a variety of sources including websites, government reports, news, wikis, blogs, online forums, and social media. This paper investigates how to leverage the information in public data sources to complement the internal data in order to support effective acquisition planning and management. This research is based on publicly available government acquisition databases at usaspending.gov and fpds.gov. It takes a data science approach for analyzing acquisition databases and focuses on two major tasks: (1) research on leveraging the web data for quality assessment and improvement of federal acquisition data and (2) research on appropriate data analytic techniques to discover useful information that can potentially help federal acquisition management and planning process.

Introduction

Military agencies collect, store, and integrate data from various sources in their acquisition and procurement decisions and management processes. However, data complexity is profound. Often, data are publicly available, but can be dirty, and become even dirtier due to biases during collection. Furthermore, acquisition, procurement, and contract data have varying data quality problems and can thus be difficult or even impossible to integrate.

Across the Department of Defense (DoD), there are hundreds of information systems that are drawn upon for defense acquisition and procurement tasks. It is challenging to standardize data across all of these information systems; yet, the capabilities to draw upon data from these systems not only are essential, but also hold huge potential for improving acquisition and procurement and reducing substantial costs across various acquisition and procurement programs. A critical challenge facing the DoD, and federal agencies in general, is how to develop data visibility capabilities to support various acquisition and procurement tasks without enforcing a single data standard across these hundreds of systems.

On the other hand, the vast quantity of online information provides great opportunities for us to harvest and enrich our data and knowledge. There are a variety of sources for the data including company and government websites and reports, news outlets, wikis, blogs, online forums, and social media. These sources contain rich information about almost everything and any subject we can think of. Indeed, searching for the needed information on the web has become a common practice for Internet users nowadays, thanks to the advancement of search engines and web technologies. If properly utilized, online information may help us assess and even improve the quality of the data we have. For instance, if a record contains a contractor's name but the address information is missing, we can fill the missing address by googling the contractor's address on the Internet. Similarly, if a contractor's DUNS number is found incorrect, then we may be able to find the right DUNS number by querying the websites that host DUNS number database.

A recent study by the Rand Corporation titled *Issues With Access Acquisition Data and Information in the Department of Defense* recommends several options for improving the DoD's acquisition data (McKernan et al., 2016). One option is to improve the quality and analytic value of acquisition data. It stated that according to information managers, **data verification and validation are top priorities** and the practice of building both manual and automated checks should be continued and expanded to other systems. Another option is to improve data analytic capabilities by continuing to collect both structured and unstructured data. It recommends that the DoD should try to come up with better ways of utilizing the unstructured data it collects.

This research aims to investigate appropriate data science approaches to improving the quality of federal acquisition data as well as discovering useful patterns that can further acquisition research. It will examine the feasibility of leveraging the information on the Internet for verification and validation of acquisition data. Utilizing online information faces several challenges. One of the key challenges is how to find the information that is credible and accurate from often an enormous amount of unstructured documents returned by a search. For instance, the information of an entity may spread out on various websites that have collected data from different sources and at different times. When searching the entity on the web, we may end up with thousands of if not millions of hits. Some of them may be incorrect and some out-of-date. Thus identifying the hits that contain both accurate and current information becomes a challenge. To make the problem even worse, the majority of online information is non-structured and textual. Thus, the question of how to extract the needed information from non-structured text becomes another challenge.

Research Issues

The web has greatly changed our ways of sharing and seeking information. At the same time, it has also altered traditional notions of trust due to the fact that the information can be published anywhere by anyone for any purpose, and there is no authority to certify the correctness of the information. It is up to the information consumers to make their own judgement about the credibility and accuracy of information they encountered online. To utilize online information effectively, this research needs to investigate appropriate methods to acquire valuable and reliable information online. Reliability of information can be measured from different aspects such as accuracy, timeliness, authority of information, trustworthiness of websites, and so forth. Consistency is another common issue with web data because the data on same subject might be different, or represented in different formats, scales, or metrics. Thus, resolving inconsistency of data from different sources and identifying the most accurate information become other key topics of this research.

As the majority of data on the web are unstructured text documents, it is challenging to identify, retrieve, and integrate the needed information from the web documents. Retrieval of desired information is not a trivial task and involves natural language processing, computational linguistics, text analysis, and entity identification and resolution. Other challenges of text analysis include complex and subtle relationships between concepts in text as well as ambiguity and context sensitivity of terms in text. The research will examine ways to identify and collectively integrate the needed information from both public and internal sources, and to leverage them for further acquisition research.

Research Methodology

This study is based on Federal Acquisition databases at USAspending.gov, which contains spending information of all U.S. departments between the years 2000 and 2018 for a selected state or all states. The data can be downloaded in different formats, such as CSV, TSV, and XML. Spending data are further categorized under prime award and sub-award. The types of spending include contracts, grants, loans, and other financial assistance. Our downloaded data contains 47GB data in total, covers the DoD budget between 2000 and 2017 including each type of spending data for both prime award and sub-award. We set up a database system to host the data.

Figure 1 shows the framework of the proposed Data Enhancement and Analytics System. The system has four major components, namely Quality Assessment engine (QA), Data Cleaning engine (DC), Data Enhancement and Analytics engine (DAE), and Text Retrieval and Analysis engine (TRA). The key component is Text Retrieval and Analysis engine as it supports the functionalities of the rest three components. TRA is responsible for four tasks: (1) performing searches on the Internet, (2) identifying the websites that contain the most reliable data, (3) extracting the desired information from the text, and (4) information fusion by collectively integrating information from multiple sources. When information needed for quality assessment and data cleaning is not available, TRA will search and extract the needed information online. Data Analytics and Enhancement engine aims to enhance our knowledge about data by discovering hidden and interesting patterns in the data as well as complementing the internal data with the information that is not found in the database but is potentially useful for advanced data analytics.

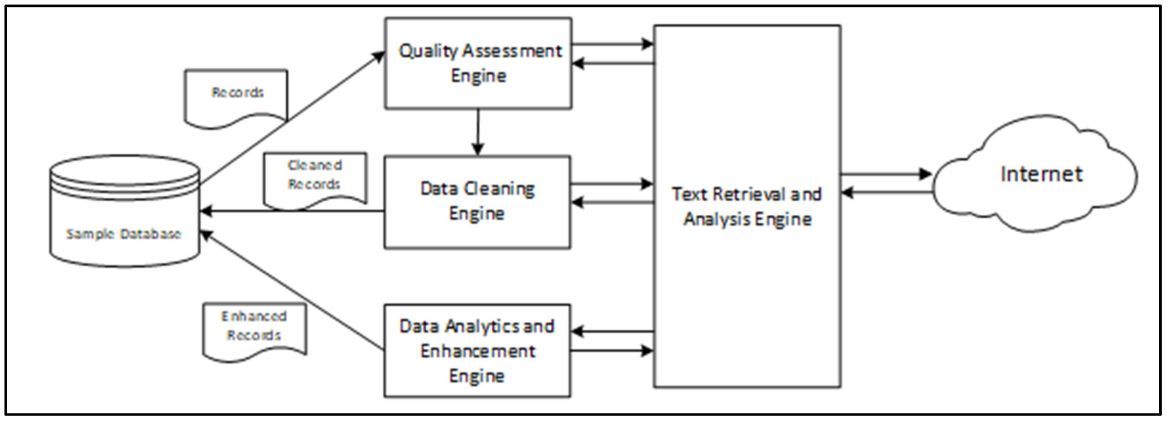

Figure 1. Framework of Data Enhancement and Analytics System

Our research methodology contains following steps:

- Assess the quality of the sample acquisition database in terms of accuracy, consistency, and completeness. Assessment on completeness is rather straightforward; however, assessment on accuracy and consistency is not, as it requires the extra knowledge about data and their semantics. For instance, to decide whether a value is accurate or not, we need to know what the expected correct value is. To evaluate whether the two values are consistent, we need to know their semantics and relationship. If they are not consistent, then we need to know which value is wrong and causing the inconsistency issue. Unfortunately, we do not always have the information we need for the quality assessment.
- Based on the quality assessment findings obtained in step 1, identify the fields for quality improvement. The key task of this step is to investigate the feasibility of leveraging the information online for both quality assessment and improvement. It will research on effective ways to evaluate the credibility of websites and to extract reliable information from a large amount of web pages.
- Apply appropriate data analytics methods to discover useful patterns from the data.
- Utilize online data to complement the information of the sample database. The primary objective of this step is to research appropriate text mining methods to retrieve the information for the purpose of advanced data analytics. Examples of information may include a business's product/service information, location, business type, business size, business relationship networks. This information can help us estimate the uniqueness of a business as well as the level of risk it might potentially pose to a project if it fails. The findings of this step can further acquisition research by identifying the room for improvement of a project.

Preliminary Research Findings

Quality Assessment

The data in sample database are organized into four tables: primeContracts, subContracts, PrimeGrants, and SubGrants. Table 1 shows general information about the tables, where *RecCnt* and *ColCnt* represent the number of records and number of columns in a table respectively; *CompleteCols* and *SingleValCols* represent the number of columns with no missing values and number of columns with only a single value across all records; and *EmptyCols* and *IncompleteCols* represent the number of empty columns and the number of columns with missing values respectively.

Table 1. Table Information of the Sample Database

Table Name	RecCnt	ColCnt	CompleteCols/ SingleValCols	EmtpyCols	IncompleteCols
PrimeContracts	23,677,787	212	50/1	0	162
SubContracts	395,569	101	41/0	3	57
PrimGrants	202,166	67	32/5	2	33
SubAGrants	11,115	101	29/4	25	47

For the quality assessment purpose, attributes are classified into two categories: identity attributes and non-identity attributes. Identity attributes provide identifier information for a contract or a contractor including project identifiers, contractor identifiers, address, telephone, and so forth. The rest attributes are non-identity attributes that do not provide identifier information. This study focuses on the quality assessment of only identity attributes on three dimensions: column completeness, accuracy, and field length consistency. Only PrimeContracts and SubContracts tables are used in this study as they have relatively more quality issues.

Column Completeness

Completeness can be measured in different aspects including column completeness, schema completeness, and population completeness. Column completeness measures the degree to which there exist missing values in a column of a table. Schema completeness measures the degree to which entities and attributes are missing from the schema. Population completeness measures the degree to which members of the population that should be present but are not present. Since there is not enough information for assessing schema and population completeness, the study will focus only on column completeness, which is measured by the percentage of non-missing values in the column.

Figures 2 and 3 show the completeness measures for identity attributes of PrimeContracts and SubContracts tables respectively. PrimeContracts table has perfect or near perfect completeness on three attributes. SubContracts table has 100% completeness on prime_award_piid and subawardee_dunsnumber, but it has missing values on both prime awardee's and subawardee's parent dunsnumbers. A possible reason might be some contractors may not have a parent company.

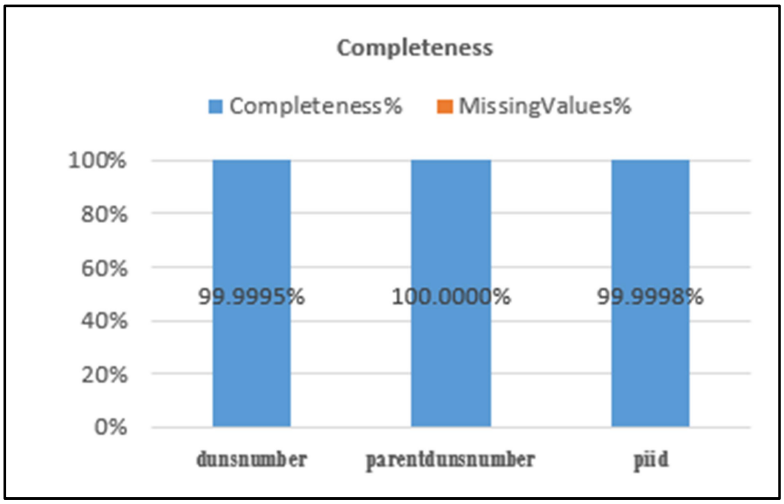

Figure 2. Completeness Measure of Identity Attributes for PrimeContracts Table

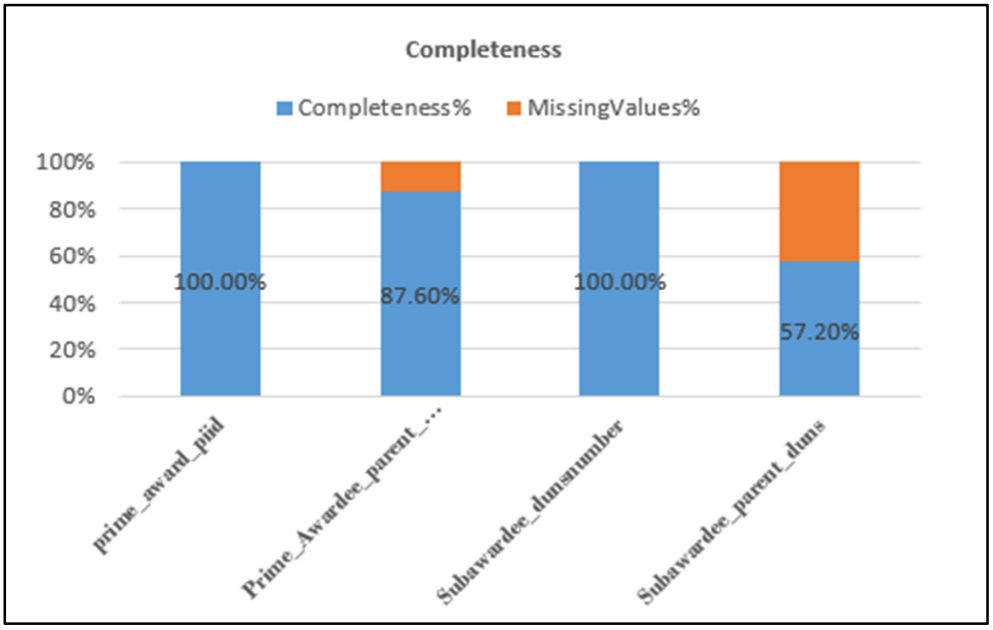

Figure 3. Completeness Measures of Identity Attributes for SubContracts Table

Attribute Length Consistency

Attribute length consistency measures how consistent lengths of an attribute's values are. Each identity attribute of the PrimeContrats table is supposed to have fixed-length values, as are the identity attributes of the SubContracts table. For example, a DUNS number, provided by Dun & Bradstreet (D&B), is a unique nine-digit identification number for each physical location of a business. Thus a DUNs number of other than nine digits is problematic. Figures 4 and 5 show the assessment of attribute length consistency of both tables. DUNS numbers in PrimeContracts table have a variety lengths; while DUNS numbers in SubContracts are consistently of nine digits.

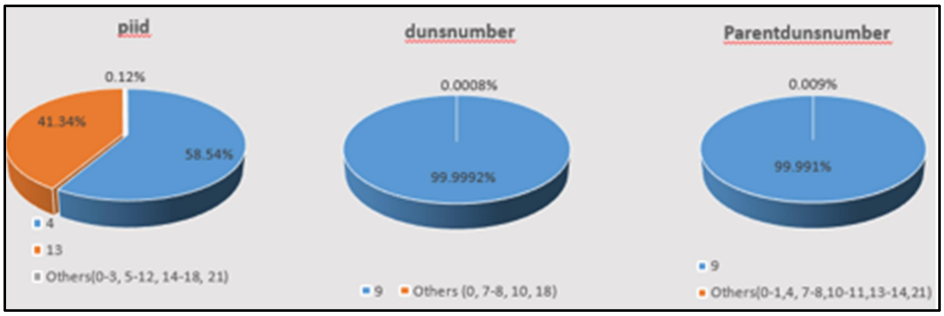

Figure 4. Field Length Consistency Assessment of PrimeContracts Table

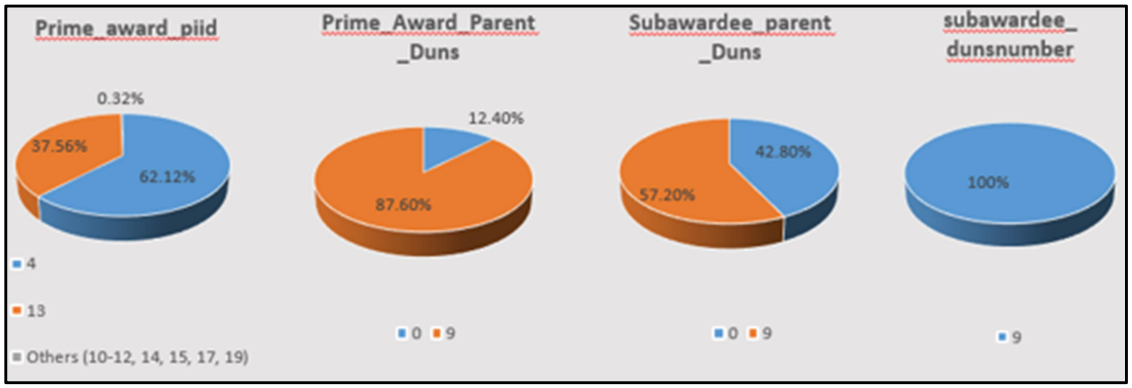

Figure 5. Field Length Consistency Assessment of SubContracts Table

Accuracy

Data is accurate if it is free of error and conforms to gold standards of data. Accessing accuracy is not an easy task as it requires the knowledge of correct data and needs to compare each data item with the known correct value, which is often not available. The accuracy assessment of this study will focus on the DUNS numbers first, because we can use the free service provided by Duns & Bradstreet database to search a business's DUNS number based on its name.

Quality Improvement

For proof of concept, the first phase of this research will focus on fixing incorrect DUNS numbers, and the online Duns & Bradstreet database will be used for this purpose. Duns & Bradstreet database contains DUNS numbers of 285 million commercial entities and 100 million associated contacts. It provides the services that allow users to search a company's information by name, telephone number, or DUNS number. When searching for a DUNS number, the database doesn't return the query results immediately, instead the results are sent through email. This somehow discourages the use of a script program to automatically submit a query and retrieve the results.

Given a business's DUNS number, the Duns & Bradstreet database can be queried for the corresponding business name, address, and telephone number. It is a bit tricky when querying the DUNS number based on the company name, as a company may have more than one DUNs number with one for each branch. To identify the right DUNS number for a branch, extra information such as the street address, zip code, and telephone number is needed.

The research started with the *dunsnumber* attribute of the PrimeContracts table. There are a total of 26 records in the table, all with a 10-digit *dusnumber*. A closer study on those DUNS numbers revealed that most of them are actually the phone numbers. Among those records, only three have information on *vendorname* attribute. The rest of the records misplace the vendor names into other fields. Since a company may have multiple branches with each located at a different address, searching the DUNs database by a business name may result in multiple matches. Thus address information is critical for finding the best match. Unfortunately, address information is misplaced in all 26 records. Figure 6 shows a few columns of the 26 records with misplaced values for vendor name and address.

ID	dunsnumber	multipleorsingleawardidc	vendorname	vendorenabled	vendorlocationdisableflag	streetaddress	streetaddress2	streetaddress3	city	state	zipcode	vendoralternatesitecode
1	2022611902	CB&I FEDERAL SERVICES LLC			2370 TOWNE CENTER BLVD		BATON ROUGE	LA	708068172	UNITED ST	LA	386491765
2	2082332929	SUNDANCE-TLI			275 S 5TH AVE_ STE 215		POCATELLO	ID	832016429	UNITED ST	ID	967391967
3	2164515588	LUCAS PRECISION_ LIMITED	LUCAS PRECISION_ LIMITED		13020 SAINT CLAIR AVE		CLEVELAND	OH	441082033	UNITED ST	OH	622006591
4	2536803243	U.S. OIL TRADING LLC			3001 MARSHALL AVE		TACOMA	WA	984213116	UNITED ST	WA	784187226
5	3053925669	WORLD FUEL SERVICES FL		9800 NW 41ST ST STE 400		MIAMI	FL	331782980	USA: UNITEI	FL	25	3054288000
6	3256738838	AVFUEL CORPORATION			47 W ELLSWORTH RD		ANN ARBOR	MI	481082206	UNITED ST	MI	131423808
7	4103792800	MABEY BRIDGE & SHORE_ INC			6770 DORSEY ROAD		ELKRIDGE	MD	210756205	UNITED ST	MD	296358146
8	4106942749			1580A W NURSERY RD		LINTHICUM	MD	210902202	USA: UNITEI	MD	2	8004439219
9	4155675899	UNITED BLOOD SERVICES		6210 E OAK ST		SCOTTSDALI	AZ	852571101	USA: UNITEI	AZ	9	4159010740
10	4794524727	GREENSCAPES TOTAL LAWNCAR			3118 S 64TH CIR		FORT SMITH	AR	729034971	UNITED ST	AR	799080705
11	4806000407			2423 W MINERAL RD		PHOENIX	AZ	850419559	USA: UNITEI	AZ	7	4806000407
12	6034311331	HECKLER & KOCH DEFENSE IN			19980 HIGHLAND VISTA DR S		ASHBURN	VA	201474189	UNITED ST	VA	166031588
13	6103858200	STV INCORPORATED			205 W WELSH DR		DOUGLASSVILLE	PA	195188713	UNITED ST	PA	106768252
14	6192388341			3589 DALBERGIA ST		SAN DIEGO	CA	921132123	USA: UNITEI	CA	51	6192388338
15	7038491000			6200 GUARDIAN GATEWAY DR		ABERDEEN F	MD	210051327	USA: UNITEI	MD	2	7036413735
16	7578734959	AH-BC NAVY JV_ A JOINT VE			804 OMNI BLVD STE 201		NEWPORT NEWS	VA	236064422	UNITED ST	VA	830647272
17	7736371666	DEHLER MANUFACTURING CO._			5801 W DICKENS AVE		CHICAGO	IL	606394030	UNITED ST	IL	5069661
18	7818717449	NOBLE SUPPLY AND LOGISTIC		302 WEYMOUTH ST		ROCKLAND	MA	23701171	USA: UNITEI	MA	9	7818711911
19	8044847840			12650 E ARAPAHOE RD BLDG		ENGLEWOO	CO	801123901	USA: UNITEI	CO	6	8044847839
20	8082451911	SENTER PETROLEUM INC			3011 AUKELE ST STE C		LIHUE	HI	967661465	UNITED ST	HI	606679520
21	8085229712	ALOHA PETROLEUM LTD	ALOHA PETROLEUM LTD		1132 BISHOP STREET STE 17		HONOLULU	HI	968132807	UNITED ST	HI	81909046
22	8088335825			707 KAKOI ST		HONOLULU	HI	968192017	USA: UNITEI	HI	1	8088361957
23	8177636775	LOCKHEED MARTIN CORPORA	LOCKHEED MARTIN CORPORATI		1 LOCKHEED BLVD		FORT WORTH	TX	761083619	UNITED ST	TX	834951691
24	9036650030	TIGER LAWN AND LANDSCAPE			1389 MCR 3222		JEFFERSON	TX	756575878	UNITED ST	TX	969073886
25	9184266191			209 SW 7TH ST		ANTLERS	OK	745233834	USA: UNITEI	OK	2	9184262871
26	9519409686	PERRIS WIND TUNNEL		2093 GOETZ RD		PERRIS	CA	925709315	USA: UNITEI	CA	41	9518057728

Figure 6. Partial View of Records With Misplaced Values

Figure 7 shows an example of using online database to retrieve the DUNS number for the business with a single DUNS number. Figure 8 shows an example of extracting the DUNS numbers of companies that have multiple DUNS numbers. The vendor name, address, zip code, and telephone numbers are used to identify the highlighted best match, then the corresponding DUNS number is retrieved.

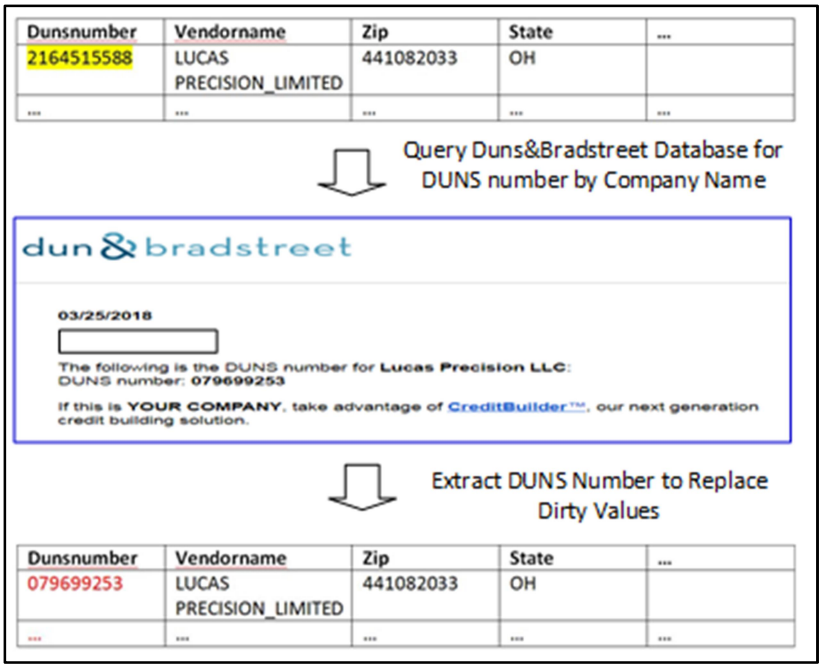

Figure 7. Examples of DUNS Number Extraction Using Online DUNS Database

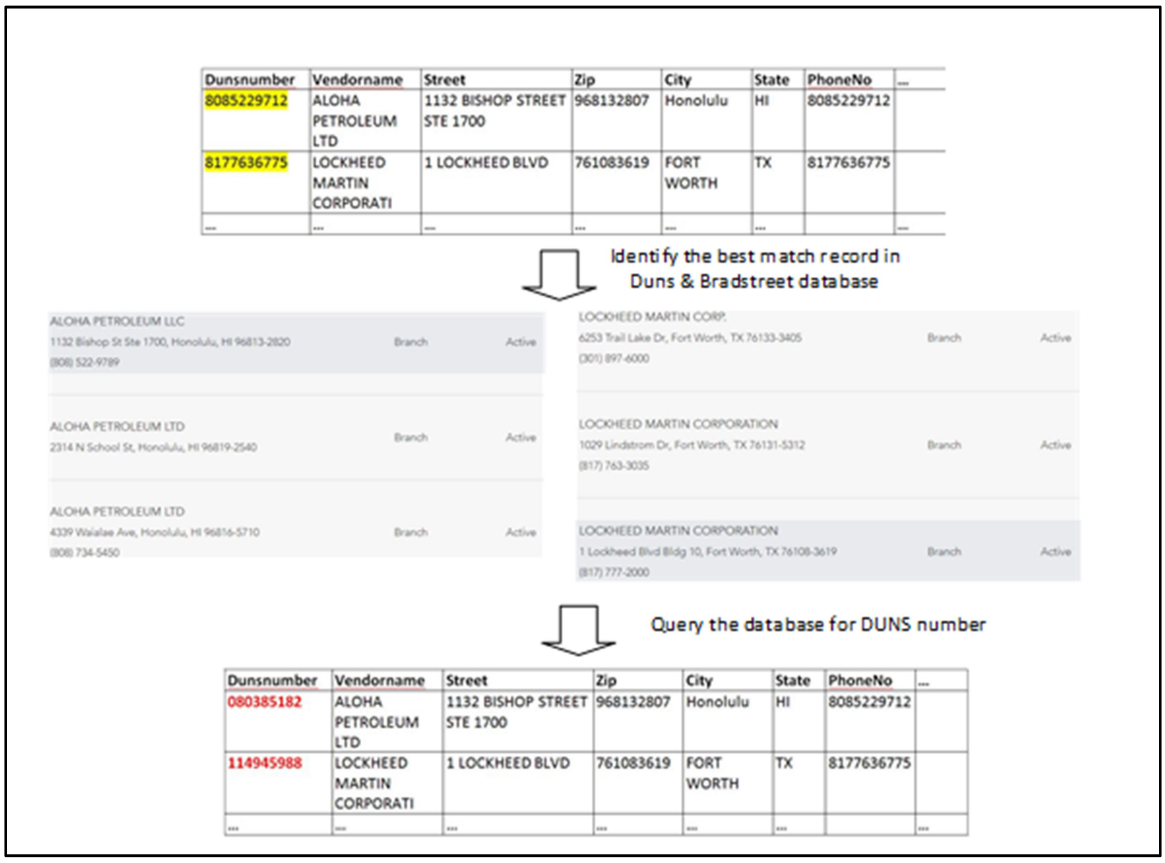

Figure 8. Examples of Identifying the Best Match in Online DUNS Database

For records with both incorrect DUNS number and missing vendor name, it is still possible to find the correct DUNS information if those records have correct and current phone numbers and address information.

Figure 9 shows the DUNS numbers, addresses, and telephone numbers that are retrieved from Duns & Bradstreet database for all 26 records. Both old and new DUNS numbers are displayed. For records with IDs 2 and 16, Duns & Bradstreet returns only a single match based on the business name; however, the returned street address is different from the one in the acquisition database. For records with IDs 8 and 11, Duns & Bradstreet returns multiple matches, all with the identical address and business name. So the DUNs number of the headquarters branch is chosen. Overall, 22 out of 26 records have *dunsnumber* filled by using the online DUNS database. The next phase of this study will be the validation and verification of the DUNS numbers and address information.

We are in the process of automating the DUNS number retrieval process, which is also a component of the Text Retrieval and Analysis engine. The Duns & Bradstreet database will be used due to its authority and reputation. The main challenge of DUNS number retrieval is to identify the best match for a company in the online database. This is indeed an entity resolution problem. Due to possible data quality problems in the sample database such as misplaced fields, typos, and missing and dirty values, identification of an exact match in DUNS database might not always be possible. The probabilistic matching methods appears promising as they can handle fuzzy matches more effectively. Another challenge is to identify the misplaced attribute values. The research will explore the methods for automatically sensing the semantics of a field based on its syntactic features and relationship with other fields and records.

ID	dunsnumber(old)	dunsnumber(new)	vendorname	Street	city	state	zip	tel
1	2022611902	079811713	CB*I Mahan LLC	2370 Towne Center Blvd	Baton Rouge	LA	70806-8172	(225)932-6500
2	2082332929	967391967*	sundance-tli	305 N 3rd Ave Ste B	pocatello	ID	83201-6306	(208) 233-2929
3	2164515588	079699253	LUCAS PRECISION_ LIMITED	13020 Saint Clair Ave	Cleveland	OH	44108-2033	(216) 451-5588
4	2536803243	784187226	U.S. OIL TRADING LLC	3001 Marshall Ave	Tacoma	WA	98421-3116	(253) 383-1651
5	3053925669	131504342	WORLD FUEL SERVICES FL	9800 Nw 41st St Ste 400	Miami	FL	33178-2980	(305) 428-8000
6	3256738838	020829396	AVFUEL CORPORATION	47 W Ellsworth Rd	Ann Arbor	MI	48108-2206	(734) 663-6466
7	4103792800	171902265	MABEY INC	6770 Dorsey Rd	Elkridge	MD	MD	(410) 379-2800
8	4106942749	005128988*	Northrop Grumman Systems Corporation	1580a W Nursery Rd	Linthicum Heights	MD	21090-2202	(410) 765-1000
9	4155675899	006902498	BLOOD SYSTEMS INC	6210 E Oak St	Scottsdale	AZ	85257-1101	(480) 946-4201
10	4794524727	799080705	GREENSCAPES TOTAL LAWNCARE& LANDSCAPING SI	3118 S 64th Cir	Fort Smith	AR	72903-4971	(479) 452-4727
11	4806000407	363821294*	ORBIT INDUSTRIAL SERVICE & MAINTENANCE	5316 W Missouri Ave	Glendale	AZ	85301-6006	(480) 704-4849
12	6034311331	134466999	HECKLER & KOCH DEFENSE INC.	19980 Highland Vista Dr Ste 190	Ashburn	VA	20147-4189	(703) 450-1900
13	6103858200	059946819	STV INCORPORATED	205 W Welsh Dr	Douglassville	PA	19518-8713	(610) 385-8200
14	6192388341	080090672	SOUTHBAY SANDBLASTING & TANK CLEANING INC	3589 Dalbergia St	San Diego	CA	92113-3810	(619) 238-8338
15	7038491000	827488979	ABERDEEN PROVING GROUND US ARM	6200 Guardian Gtwy	Aberdeen Proving Ground	MD	21005-1327	(410) 273-2640
16	7578734959	830647272*	AH/BC NAVY JV LLC	11837 Rock Landing Dr Ste 300	Newport News	VA	23606-4493	(757) 873-4959
17	7736371666	059530805	DEHLER MANUFACTURING COM	5801 W Dickens Ave	Chicago	IL	60639-4030	(773) 637-0615
18	7818717449	107910259	NOBLE SALES CO. INC	302 Weymouth St	Rockland	MA	02370-1171	(781) 871-1911
19	8044847840	079427300	PERFORMANCE FOOD GROUP	12650 E Arapahoe Rd	Englewood	CO	80112-3901	(303) 662-7100
20	8082451911	069891588	SENTER PETROLEUM INC	3011 Aukele St Ste C	Lihue	HI	96766-1430	(808) 245-1911
21	8085229712	080385182	ALOHA PETROLEUM LTD	1132 Bishop St Ste 1700	Honolulu	HI	96813-2820	(808) 522-9789
22	8088335825	007050586	GARLOW PETROLEUM	707 Kakoi St	Honolulu	HI	96819-2017	(808) 836-1957
23	8177636775	008016958	Lockheed Martin Corporation	1 Lockheed Blvd Bldg 10	Fort Worth	TX	76108-3619	(817) 777-2000
24	9036650030	969073886	CHRIS GIBBONS	1389 Mcr 3222	Jefferson	TX	75657	(903) 665-8190
25	9184266191	619338010	CHOCTAW MANUFACTURING DEFENSE CONTRACTO	209 Sw 7th St	Antlers	OK	74523-3834	(580) 298-2203
26	9519409686	141883517	PERRIS SKYVENTURE	2093 Goetz Rd	Perris	CA	92570-9315	(951) 940-4290

Figure 9. DUNs Numbers and Addresses Retrieved From Duns & Bradstreet

Data Mining

Data mining is the process of examining large data sets to uncover hidden but interesting patterns such as unknown correlations, market trends, customer preferences, and other useful business information. The analytical findings can shed significant insights to help add perspective to use the data and to lead to more effective decision makings. Some major data mining techniques include association discovery, classification, clustering, regression, sequence or path analysis, and structure and network analysis.

Association discovery aims to find frequent patterns that represent the inherent regularities in the datasets. Applications of association discovery include association, correlation, and causality analysis; basket data analysis; and cross-marketing, and so forth. Classification, also called supervised learning, is the task of inferring a function from labeled training dataset. The function can then be used to classify new data instances. Decision tree, Bayesian networks, support vector machine, and neural networks are some of the commonly used models for classification. Clustering, also called non-supervised learning, group a collection of data objects into groups according a predefined distance function. Clustering can be employed as a stand-alone tool to get insights about data or as a preprocessing tool for other algorithms. Sequence analysis discovers patterns among sequences of ordered events or elements. Application of sequence patterns include customer shopping sequence, DNA sequences and gene structures, sequences of stock market changes, and so forth. Graph and network analysis aims to discover frequent subgraphs, trees, or substructures. It has been used for social networks analysis and web mining.

Cluster and Network Analysis

As the first phase of the research, network analysis is performed on prime contractors and their subcontractors of the sample database in a hope to discover the business networks among contractors. Figure 10 visualizes some findings from the network analysis. It shows the top three big contractors that have the largest number of subcontractors, and top three highly demanded subcontractors who are working for the largest number of different contractors.

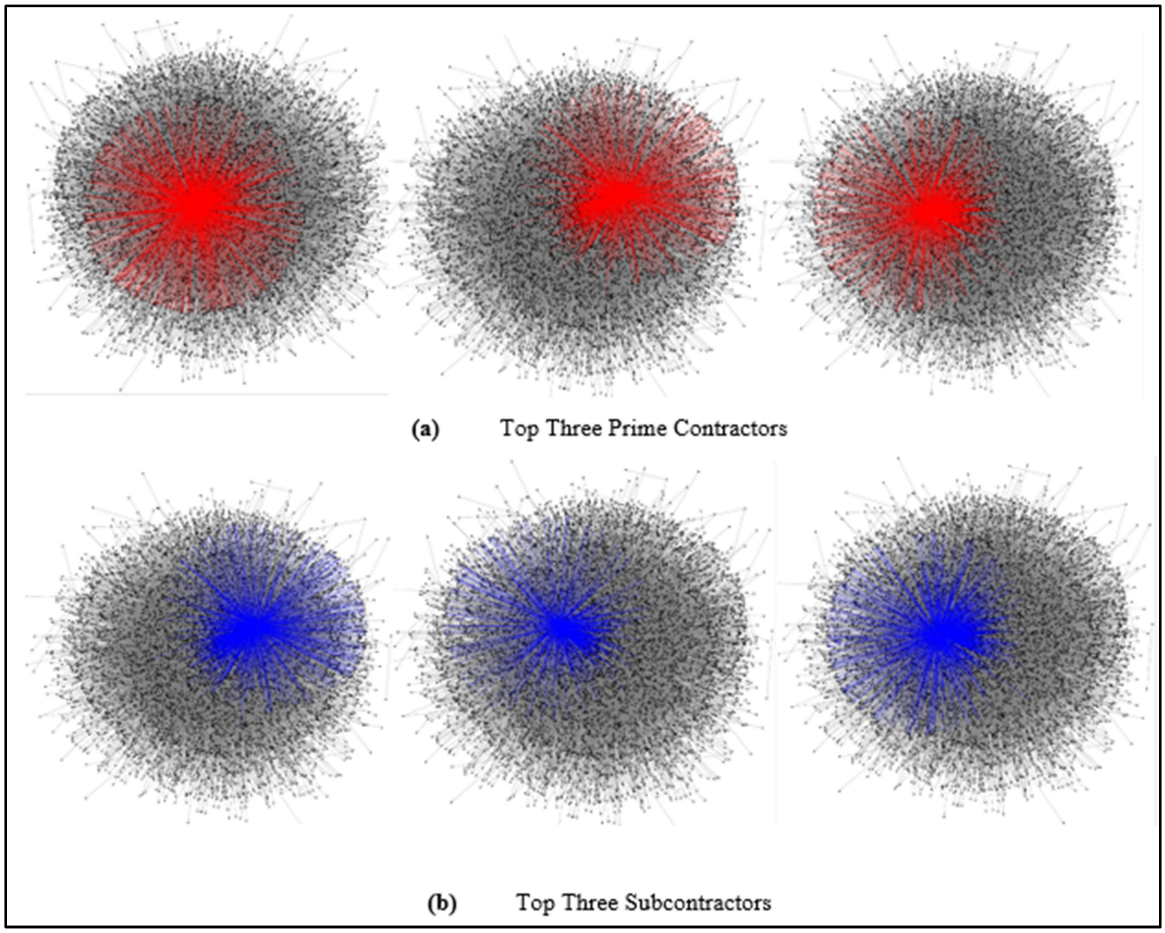

Figure 10. Cluster Analysis of Contractors

Figure 11 shows the clustering results of only contractors that worked with at least five subcontractors. Figure 11(a) shows overall clustering result, where each dot represents a primary contractor. The dots in orange are “big” primary contractors with many subcontractors. The dots in purple are relatively “small” primary contractors. Figure 11(b) shows zoomed-in clusters for two big prime contractors.

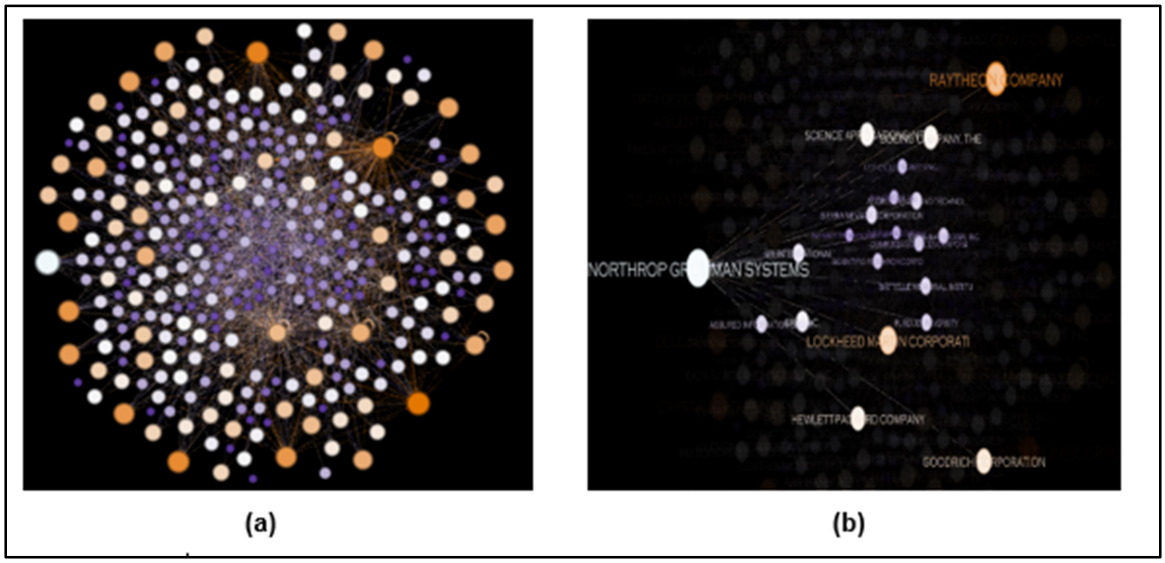

Figure 11. Clustering Results for Contractors Involved in More Than Five Projects

Figure 12 shows the analysis of the relationship between contractors and subcontractors by state. Each dot represents a state. The size of a dot is determined by the number of contracts awarded to a state. A directed edge represents the relationship between primary contractors and their subcontractors (pointed by an arrow). The thicker an edge is, the more contracts there are between the primary contractors and their subcontractors. The figure shows that some states, like California, get more contracts than others, and some states, like Illinois, tend to subcontract their projects to the other states.

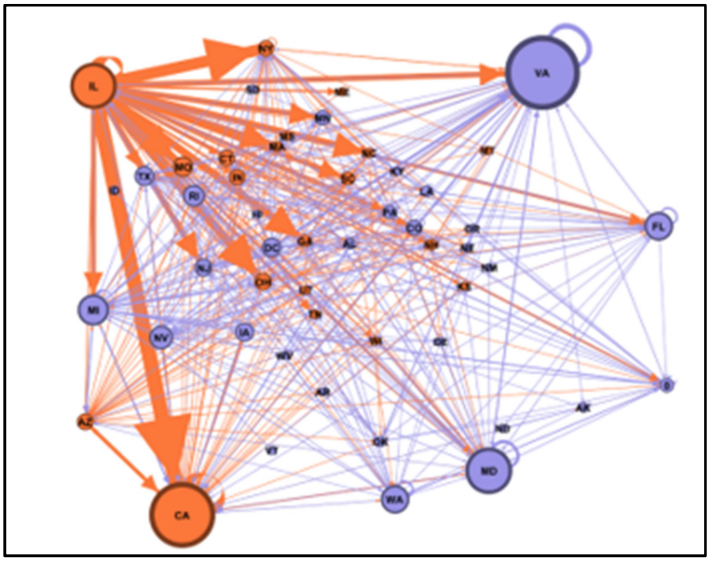

Figure 12. Clustering Results by State

Figure 13 shows the relationship of contractors of different business types. Each dot represents contractors of the same business type. A line between two dots indicates companies of two different business types are related by a contract. The figure reveals that companies tend to give subcontracts to the companies of the same business type. There are only two outliers that relate companies of different types.

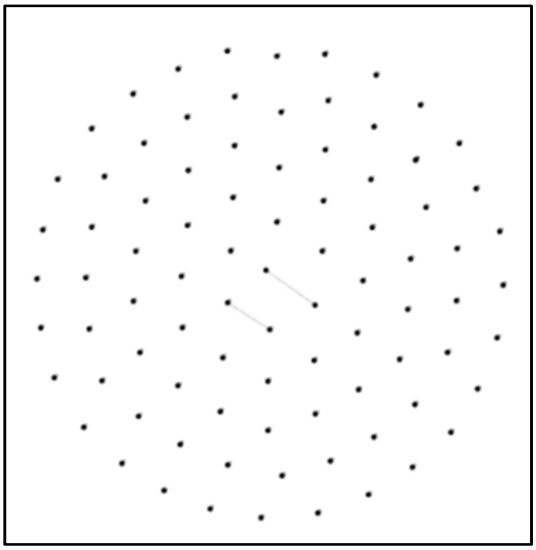

Figure 13. Relationship of Companies of Different Business Types

Pattern Discovery

A preliminary data analysis was performed and aimed to discovery patterns that may shed insights into possible areas for improvement in acquisition projects. For instance, small contractors are usually less robust and easier to fail compared to large contractors when facing natural or man-made disasters. Projects with subcontractors located in places that have a high risk of natural disasters such as earthquakes may have risks of a potential delay in delivery time. By taking into account of the risk factors in planning a project can help identify the room for improvement to ensure the successful and prompt delivery of the project.

The first round of analysis focused on finding the following patterns in the existing projects: (1) small-business subcontractors that are involved in different projects led by some key primary contractors and (2) projects that have multiple subcontractors located in a place that has a high risk of natural disasters such as earthquakes, hurricanes, flooding, or wild fires. The following section discusses two examples of our findings.

Finding 1

PTB is a small, single-location company with less than 200 employees. It was involved in six different projects led by some key primary contractors including Boeing, Lockheed Martin, and L-3 Communications. The average award amount is about $5,400. A close study of the company's website, shown as Figure 14, revealed it may provide some important services to its primary contractors. Since company websites and the acquisition database are all publicly accessible, they might be used by enemies for inferring sensitive information on a project or planning attacks to make the project fail.

Figure 14. Snapshot of PTB Webpages

Finding 2

We retrieved locations of 7.0-magnitude quake epicenters in United States from the U.S. Geological Survey website, www.usgs.gov, and identified 118 subcontractors located nearby an epicenter of 7.0-magnitude quake. In *SubContracts* table, 984 awards have at least one subcontractor located in the high-risk earthquake areas; 41 of them have at least two subcontractors located in the high-risk areas. Table 3 shows the top five contracts with the most number of subcontractors in high-risk earthquake areas.

Table 2. Top Five Contracts With the Most Number of Subcontractors in High Risk Earthquake Areas

Project ID	#Subcontractors
1	15
2	15
3	11
4	10
5	8

Finding 3

We did a preliminary exploration of whether a correlation exists between procurement data and employment information. The result shows that large reductions in federal contracts are correlated in a majority of cases (66% or 75%, depending on the metric used) to drops in employment in a given region and industry. This finding shows that it is possible to determine the location of an undisclosed contractor by examining public employment data at the times when large contracts are reduced or simply reach the end of their period. Such undisclosed contractors are typically employed by larger government contractors to achieve confidentiality, security, or a competitive advantage. Depending on the situation, acquisition experts may need additional planning to protect such hidden

contractors if security is desired, or may rely on data science to identify these contractors and avoid them becoming a weak link in the acquisition process.

Conclusion and Future Work

The paper proposed a Data Enhancement and Analytics framework that is designed to use public data for improving quality and data analytic capabilities of the acquisition data. A proof of concept analysis was conducted to show the feasibility of using web information for quality assessment and improvement of the sample database. Still more needs to be done to implement the framework.

The future work will focus on the following two directions. First, we will research effective text analysis and trust evaluation techniques to identify credible and valuable information from the web. Second, we will research appropriate big data analytic techniques that can help enhance decision-making capabilities for acquisition management and planning.

Literature Review

This section summarizes some related work in the fields of federal acquisition data analysis, trust in web information, and Defense Acquisition Visibility Environment (DAVE).

Federal Acquisition Data Analysis

Apte, Rendon, and Dixon (2015) explored the use of big data analytic techniques to explore and analyze large dataset that are used to capture information about DoD services acquisitions. The paper described how big data analytics could potentially be used in acquisition research. As the proof of concept, the paper tested the application of big data analytic techniques by applying them to a dataset of Contractor Performance Assessment Report System (CPARS) ratings of 715 acquired services. It also created predictive models to explore the causes of failed services contracts. Since the dataset used in the research was rather small and far from the scope of big data, the techniques explored by the paper mainly focus on traditional data mining techniques without taking into account of big data properties.

Black, Henley, and Clute (2014) studied the quality of narratives in Contract Performance Assessment Reporting System (CPARS) and their value to the acquisition process. The research used statistical analysis to examine 715 Army service contractor performance reports in CPARS in order to understand three major questions: (1) To what degree are government contracting professionals submitting to CPARS contractor performance narratives in accordance with the guidelines provided in the CPARS user's manual? (2) What is the added value of the contractor performance narratives beyond the value of the objective scores for performance? (3) What is the statistical relationship between the sentiment contained in the narratives and the objective scores for contractor evaluations?

Our proposed research focuses on a much broader scope of acquisition projects. The research starts with cleaning and enhancing the acquisition data first. Once data is clean enough and has sufficient information, then advanced data analytic techniques will be applied in hopes of discovering interesting patterns that can be used to further acquisition management and planning research.

Trust in Web Information

Extensive research has been conducted on evaluating credibility and trust of online information, primarily textual information. The research by Corritore, Kracher, and Wiedenbeck (2003) identified three factors that impact trust in online environment: perception of credibility, ease of use, and risk. Cheskin (1999) identified six major features that encouraged trust in websites. The features are brand (the reputation of a company), seals of approval (icons from companies that certify a site as following security measures), ease of navigation, fulfillment (trust or distrust developed in using the web), presentation, and technology. In addition to the six features, the trust is expected to develop over time. The more interaction between the user and a website, the more information the user would gain to decide how much to trust it. Fogg et al. (2001) studied what makes a website credible. It defined credibility as believability and considered trustworthiness a major component of credibility. The paper also identified four factors as contributing to trustworthiness, namely linking (where the user was linked from and where the site links), policy statement, social recommendations, and business interest.

The commonly recommended approaches to online information evaluation include five criteria, including checking the accuracy, authority, objectivity, currency, and coverage or scope of the information and/or its source (Metzger, 2007). Accuracy refers to the degree to which a website is error free. The authority can be about a website or an author of information. The authority of a website is usually measured by its reputation and authority. The authority of an author is measured by the author's credentials and qualifications on the specific subject of information. Objectivity measures whether the information is fact or opinion. Currency refers to how up-to-date the information is, and coverage refers to the comprehensiveness or depth of the information provided.

Recent research shows that people tend to use verification strategies that require the least effort to perform. For instance, instead of using the recommended five criteria in evaluation, they opted to base decisions on factors like website design and navigability (Fogg et al., 2003). These findings are consistent with some recent credibility studies (Hilligoss & Rieh, 2008) and with theories from information processing and cognitive science (Sundar, 2008; Taraborelli, 2008). These theories stipulate that people have constraints on their ability to process information, and they tend to use cognitive resource that is just enough for a sufficiently optimal outcome for the evaluation task (Lang, 2000; Fogg et al., 2003).

Defense Acquisition Visibility Environment (DAVE)

Acquisition Visibility (AV) is the concept of providing the DoD with data and analysis support capabilities to inform the acquisition community. DAVE establishes a framework for improved and expanded support to the Under Secretary of Defense for Acquisition, Technology, and Logistics (USD[AT&L]). As shown in Figure 15, DAVE employs a three-tiered architecture that contains the *DAVE portal, DAVE Platform*, and AV Data Framework.

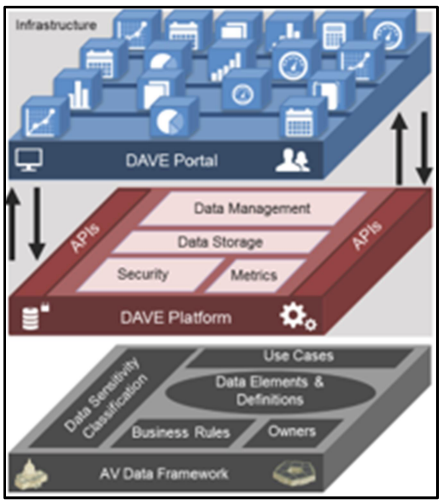

Figure 15. Defense Acquisition Visibility Environment (DAVE)

The *DAVE portal* is a synthesis of interactive infrastructure including data visualizations, calendars, and project management tools that are set to continue to grow in scope and capability as DAVE expands. These diverse tools with analysis capabilities will help users answer such questions as, “Are we solving a business problem by assessing the efficiency and effectiveness of the project?” and “What value does the project add to acquisitions in the Department of Defense?”

The DAVE platform includes the Application Programming Interfaces (APIs) for data management, data storage, metrics, and security. The DAVE platform determines the APIs for facilitating data access, and determines to which party the information can be shared. The APIs are the building blocks that allow for the integration of features or data, and the platform itself processes the data to get it to the state users require, as well as coordinating internal processes.

The AV Data Framework is the foundation on which the portal and platform are built and provides a number of essential elements including use cases, data elements and definitions, business rules, guidelines and markers regarding ownership of data, and data sensitivity classifications.

The proposed Data Cleansing and Enhancement System in this research can be used to support part of the functionalities of the DAVE platform as it prepares the data to the state that is suitable for user consumption or further analysis by the data leaning and enhance processes.

References

Apte, U., Rendon, R., & Dixon, M. (2016). Big data analysis of contractor performance information for service acquisition in DoD: A proof of concept. In *Proceedings of the 13th Annual Acquisition Research Symposium.* Monterey, CA: Naval Postgraduate School.

Augustine, N. R. (1997). *Augustine's laws*. AIAA.

Black, S., Henley, J., & Clute, M. (2014). *Determining the value of Contractor Performance Assessment Reporting System (CPARS) narratives for the acquisition process* (NPS-CM-14-022). Monterey, CA: Naval Postgraduate School.

Brown, B. (2010). *Introduction to defense acquisitions management* (10th ed.). Defense Acquisition University. Retrieved from http://www.dau.mil/publications/publicationsDocs/Intro%20to%20Def%20Acq%20Mgmt%2010%20ed.pdf

Cheskin, S. (1999). *Ecommerce trust: Building trust in digital environments*. Archetype/Sapient.

Cilli, M. Parnell, G. S., Cloutier, R., & Zigh, T. (2015). A systems engineering perspective on the revised defense acquisition system. *Systems Engineering, 18*(6), 584–603. doi:10.1002/sys.21329

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737–758.

Defense Acquisition University (DAU). (2016). *DAU Center for Defense Acquisition: Research agenda 2016–2017*. Retrieved from http://dau.dodlive.mil/files/2016/01/ARJ-76_ONLINE-FULL.pdf

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., … Treinen, M. (2001). What makes web sites credible?: A report on a large quantitative study. In CHI '01: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 61–68). New York, NY: ACM.

Gaither, C. C. (2014). Incorporating market based decision making processes in defense acquisitions. *International Journal of Defense Acquisition Management, 6*, 38–50.

Gallup et al. (2015, May). *Lexical link analysis (LLA) application: Improving web service to defense acquisition visibility environment*. Distributed Information Systems Experimentation.

Golbeck, J. (2008). Trust on the world wide web: A survey. *Foundations and Trends® in Web Science, 1*(2), 131–197.

Hagan, G. (1998). *Glossary: Defense acquisition acronyms and terms*. Fort Belvoir, VA: DoD, Defense Systems Management College, Acquisition Policy Department.

Krzysko, M. (2012, February). The need for acquisition visibility. *Journal of Software Technology*, 4–9.

Krzysko, M. (2016). *Acquisition decision making through information and data management.* Retrieved from http://www.digitalgovernment.com/media/Downloads/asset_upload_file917_5737.pdf

McKernan, M., Moore, N. Y., Connor, K., Chenoweth, M. E., Drezner, J. A., Dryden, J., … Szafran, A. (2016). *Issues with access to acquisition and information in the Department of Defense*. Santa Monica, CA: RAND.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*(2013), 210–220.

Miller, A., & Ray, J. (2015, January 1). Moving from standard practices to best practices in defense acquisition. *Defense ARJ, 22*(1), 64–83.

Pennock, M. J. (2008). *Defense acquisition: A tragedy of the commons*.

Under Secretary of Defense. (2007, November). *Operation of the defense acquisition system* (DoDI 5000.01). Washington, DC: Author.

Under Secretary of Defense. (2015). *Operation of the defense acquisition system* (DoDI 5000.02). Washington, DC: Author.

Acknowledgment

We would like to thank Jamoris Miller and Leonardo Vieira for their help in visualization of some results.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net