IP DATA REPOSITORY STUDY

Laura Freeman
VIRGINIA TECH HUME ISL
Brian Mayer
VIRGINIA TECH SANGHANI CENTER

SPONSORS:
OFFICE OF THE UNDER SECRETARY OF DEFENSE
FOR ACQUISITION AND SUSTAINMENT
THE OFFICE OF THE UNDER SECRETARY OF DEFENSE
FOR RESEARCH AND ENGINEERING

EXECUTIVE SUMMARY AND BRIEF | OCTOBER 2021

DISCLAIMER

The Acquisition Innovation Research Center is a multi-university partnership led by the Stevens Institute of Technology and sponsored by the U.S. Department of Defense.

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Under Secretary of Defense for Acquisition and Sustainment (OUSD(A&S)) and the Office of the Under Secretary of Defense for Research and Engineering (OUSD(R&E)) under Contract HQ0034-19-D-0003, TO#0653.

Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

No Warranty.

This Material is furnished on an "as-is" basis. Virginia Tech and the Stevens Institute of Technology make no warranties of any kind, either expressed or implied, as to any matter including, but not limited to, warranty of fitness for purpose or merchantability, exclusivity, or results obtained from use of the material. Virginia Tech and the Stevens Institute of Technology do not make any warranty of any kind with respect to freedom from patent, trademark, or copyright infringement.

EXECUTIVE SUMMARY

In support of the Acquisition Innovation Research Center (AIRC), Virginia Tech was tasked to develop a data and electronic information repository for models, samples, templates, and exemplars of Intellectual Property (IP). Through the use of the "FAIR" principles (Findability, Accessibility, Interoperability, and Reusability), the team researched government data using a two-tiered approach that explored the requested data fields in the AIRC Performance Work Statement (PWS) and publicly available metadata to create a longitudinal topology of DoD contracting activity.

From the analysis of the available data on SAM.gov, the team constructed queries to build and update the tasked repository. The team created a semi-automated search tool for the SAM.gov database to structure acquisition data and enable further exploratory analysis of data on the DoD's Major Defense Acquisition Programs (MDAPs), Other Transaction Authorities (OTAs), Vendor Contracted Organizations, Funding Organizations, and Patents. This semi-automated search tool enabled the team to explore more than 15M unique contracting activities across 12M unique PIIDs containing obligations of $1,797B in funding (2018-2021). This semi-automated tool may be applicable to SAM.gov as well as other download-limited APIs for future AIRC projects.

Although the team accomplished analysis on open-source SAM.gov data, full access to DoD contracting data is necessary for future AIRC Digital Data Strategy development. The team found access to DoD contracting data to be restricted and "view only", which limited the application of analytics for improving decision making, curating data and developing an IP repository.

REPORT FINDINGS

Virginia Tech researched, assessed and analyzed historic data sets for the development of a data repository that, along with the team's recommendations, support research and new policy and practice development. The team developed the following major findings:

- This project created a semi-automated search and extraction tool of the SAM.gov database, demonstrating the potential of analytical tools that consolidate, curate, and make the vast amounts of decentralized DoD acquisition data ready for analysis. A new database created by this tool structures acquisition data in a way that enables analysis and academic engagement that lead to acquisition innovation.
- Given availability of further acquisition data, there is potential for adding to the repository and furthering analysis and insights. This additional data will require support for ingest, processing, quality control, and potential automation. Further research is required to identify the potential impact or meaning of analyses that leverage this data.

Solutions to two major findings are imperative to accomplish the long-term goal of AIRC and the construction of a Digital Data Strategy:

1. *The Common Access Card (CAC) and overall access issues obstruct the analysis needed for an IP repository and future DoD AIRC Digital Data Strategy efforts.*
2. *Data access will enable the development and application of analytics to the contracting subset and provide the DoD an operational use case for future AIRC Digital Data Strategy efforts.*

DATA SOURCE REVIEW

As experienced during the course of this task, the team was not able to use or assess the data referenced, limiting its ability to provide insight into IP contracting and decision making.

The goal of the research proposed under this task is to identify data that can be used to evaluate IP contracting, which requires identification and access to the appropriate data, followed by confirmation of this data as an indicator of IP contracting effectiveness. The government requested identification of data to evaluate IP contracting but did not provide specific data sources.

In spite of these limitations, a major contribution of this work is the acquisition data topology review, which enabled identification of sources that contained structured or unstructured information in some way associated to IP and contract effectiveness.

TWO-TIERED APPROACH

To support the development of a DoD-wide repository regarding IP considerations, data identification efforts focused on two areas: 1. exploration of data fields referenced by the government within the Performance Work Statement (PWS); and 2. consideration of publicly available metadata to create a longitudinal topology of DoD contracting activity.

With millions of contracting activities occurring annually involving hundreds of billions of dollars, establishing a topology facilitates the effective targeting of data of interest across multiple attributes, over time, and hierarchically.

The term "model" was mentioned on multiple occasions within the PWS and referenced data fields. This term has different definitions related to developing a data repository.

- **PWS Reference**. Paired with "model contracting language" or "model solicitation", model describes an exemplar of one or multiple attributes. Given the volume and diversity of DoD contracting activities, there would likely be multiple exemplars conditioned on type of work, organization, etc.
- **Analytic Approaches for Data Curation**. Given that some of the desired information comes from unstructured text, specialized data processing models are needed to make this data analysis-ready.

IP-RELATED GOVERNMENT REFERENCED INFORMATION

Research efforts focused on government-identified data elements of interest with potential IP considerations and binned these data elements into four categories based on source location.

- **Category #1** are items within an executed contract document, referenced by a Procurement Instrument Identifier (PIID) with a date signed. Items contained within these documents are not indexed or referenceable in any existing database requiring a per contract inspection, defined by the PIID and date signed. These documents are contained in the DoD Electronic Data Access (EDA) document repository and locally in a contracting organization document storage system.
- **Category #2** are items contained within a pre-solicitation announcement such as a Request for Proposal (RFP) or Request for Quote (RFQ) that, similar to the contract document items, are not referenceable in a DoD database. Complicating the process of attribution, some solicitations do not result in a contract award. Elements within the announcements and documents are written in paragraph form, or unstructured text, which requires NLP and a defined ontology regarding attributes of interest to be extracted.
- **Category #3** is Service Contract Reporting (SCR), which applies only to a subset of awarded contracts and is entered by a prime vendor through the SAM.gov portal.
- **Category #4** includes data for which no authoritative source or documents were provided by the government and that, as a result, could not be considered.

DoD CONTRACT ACTIVITY PROCESS

Figure 1 illustrates the three main entities involved in a contracting activity as well as categorized attributes regarding that activity.

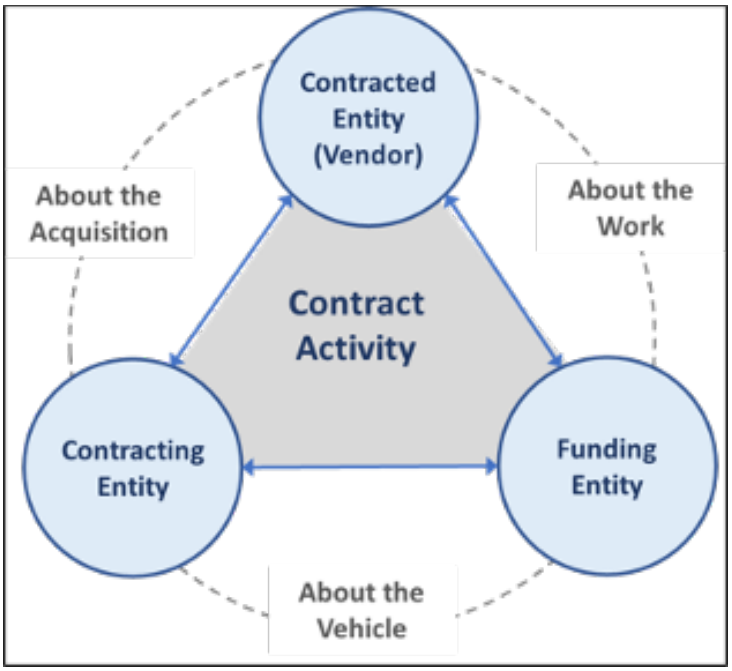

Figure 1: Contracting Activity Network Graph

GOVERNMENT REPORTING

When a contract activity is executed, two electronic reporting actions are initiated from the supporting contracting organization via their contracting system: 1. transmission of metadata is reported to the SAM Databank and stored locally. Research showed that there is up to a 90-day reporting delay from the DoD to SAM databank; and 2. a copy of the contract execution document(s), in PDF format, is transmitted to the EDA (Electronic Data Access) system. Research and lack of system access could not determine whether supporting contract execution documents are uploaded.

These reporting requirements are associated with a contract "activity" that also includes a broader area of contract reporting than contract award. Contract awards can be 1 year, 3 years, or up to 25 years and changes are common over the duration of the contract, all of which constitute activities reported into SAM.gov. Of the IP-related government referenced information, data elements in Category #1 are susceptible to change over the duration of the contract. Capturing these data elements only at the time of award is insufficient, and inspection of activities for each PIID is necessary to determine whether an activity caused a change in a CLIN, CDRL, clause, or other data element of interest. Each of these inspections requires some combination of NLP algorithms to evaluate the PDF documentation associated with that activity.

AFFECTED ENTITIES

Figure 1 illustrates the three entities required to execute a contracting activity. While the Contracting Entity is responsible for activity reporting to systems described, each entity will have copies of all relevant contract execution documents. For the Funding Entity, additional documents retained locally can include: industry responses to a Source Sought (SS) or Request For Information (RFI) that inform a subsequent solicitation; Q&A responses to a solicitation; Industry Day slides (if conducted); and Proposal reviewer comments and grading. Documents unique to and retained by the Vendor can include separate Annex 2 subcontracting agreements with their former and current teammates, as well as distribution of work overall or by CLIN. A small subset of contracting activities executing sensitive national security missions will be classified at the Secret or Top-Secret Level.

GENERALIZED DATA OWNERSHIP & ACCESS

Each entity involved in a contracting activity receives copies of all relevant documents. The Contracting Entity has specific reporting responsibilities with regards to activity metadata and document upload into EDA. The Contracted Entity has a reporting responsibility related to SCR for designated contracts. Ownership of contract documents is controlled by the three entities and each can choose to release none, some, or all of that information to designated parties with or without restrictions.

For the purpose of establishing an IP repository for reference and usage across the DoD, research focused on the authoritative systems used to store this data, whether in a database or a document repository. Access to both EDA and the SRC portion of SAM Databank require a CAC and/or approval permission from the respective data owner. Even with permission, there may be additional requirements and/or restrictions as to what can be viewed.

In the case of EDA, document access is assumed to be viewable only with a download capability from an individual and includes bulk files. A separate approval process is required to move documents to another environment. The research indicates the front-end of DoD systems do not directly support the use of specialized NLP and ML algorithms for data analysis. This particular contract, WRT-1037, did not authorize CACs for research personnel nor provide necessary access permission to perform analysis for repository development. *The team identified the CAC and access issues as obstructions to enable the necessary analysis for an IP repository and for future DOD Acquisition Innovation Research Council (AIRC) Digital Data Strategy efforts*

PUBLICLY AVAILABLE DoD CONTRACTING METADATA

The SAM Databank is a portal to multiple databases providing an array of information regarding contracts, contract activities, pricing, solicitations, and company registration to do business with the Federal Government. Some information is publicly available, which enabled the team to focus on contracting activities, the foundation from which any IP implications may arise. The critical field within this data is the PIID, a unique ID field similarly used to search with EDA and for SRC information. Interest centered on contracting activity specific to the DoD and resulted in three different scenarios included in the data collection process:

- Scenario #1: DoD funding using a DoD contracting organization
- Scenario #2: DoD funding using a non-DoD contracting organization
- Scenario #3: Non-DoD funding using a DoD contracting organization

SOURCE & DATA FIELDS

The data extracted from the SAM Databank covered different aspects of the contracting process, listed below:

1. The Contract Activity
2. The Contracted Entity, or Vendor
3. The Funding Entity
4. The Contracting Entity
5. The About the Work data
6. The About the Vehicle data
7. The About the Acquisition data

PUBLICLY AVAILABLE DATA COLLECTION PROCESS

The work to extract the whole set of DoD-related contract activities for a particular quarter is split into four types of queries:

1. Indefinite Delivery Vehicle (IDV) activities with a DoD contracting organization.
2. "Historical" IDV activities with a DoD contracting organization.
3. IDV activities with a non-DoD contracting organization and a DoD funding organization.
4. Other Transaction Authority (OTA) activities with a DoD contracting or funding organization.

Query Types

SAM.gov has separate query interfaces for IDV and OTA activities so those must be separate queries. The primary challenge for extracting the whole set of DoD-related contract activities from the SAM databank is that SAM limits results of a report to 150,000 records.

Building and Updating the Repository

The most efficient solution for SAM.gov's 150,000 record limit is to filter each query with a short date range. The date signed field, representing when a contract activity went into effect, is appropriate. Because running one query per week for the entire date range of this project (FY18-Q2FY21) would be time consuming and because repetitive tasks are error prone, the team automated this part of the SAM.gov data query using a package that automatically finds and clicks buttons in a browser. Queries were run at the end of a quarter for the preceding quarter to keep the database current.

Ingest and Quality Control

The data repository refactors SAM data into third normal form by splitting the details of organizations into their own table and referencing them by their identifier. The repository stores IDV and OTA activities in the same table. During ingest of the SAM data into the database, it is important to ensure each contract activity is represented once. The solution used relies on a unique identifier for each activity, a process that can detect duplicates and ignore them or update the database as appropriate.

DATA AVAILABILITY

Currently the database contains all activities for any contract that had activity any time starting in FY18 through the second quarter of 2021. There is a total of:

- 15,396,422 unique contract activities across 12,262,027 unique PIIDs obligating $1,797B in funding
- 28,543 funding offices
- 1,776 contracting offices
- 143,904 vendors
- 4.4 GB of data

ENABLED DATA ANALYSIS

(DESCRIPTIONS OF PROCESSES AND FINDINGS ARE FOUND WITHIN THE APPENDICES INCLUDED WITH THE FINAL REPORT)

SAM Databank is a portal to multiple databases and repositories. Once registered, users have access to 37 "standard" reports, the format of which does not allow for customization. From the portions a public user could access, the team found rudimentary search capabilities, no analytic capability supporting frequency counting, no histograms or normality tests, and no visualization capabilities, only basic tabular displays.

The semi-automated search tool to extract and re-ingest in an integrated schema has addressed data download challenges, with the immediate ability to do longitudinal and hierarchical analysis at a scale previously not supported. Perhaps the biggest benefit from data accessed in this manner is it enables exploratory research and analysis as a targeting mechanism to identify contracting activities of interest based on a single or combination of attributes. The constructed data set enabled exploratory analysis focused on the following:

- The DoD's Major Defense Acquisition Programs (MDAPs)
- Contacting activities using Other Transaction Authorities (OTAs)
- Top Contracted Organizations (Vendors)
- Funding Organization
- Patents

CONCLUSION

- Extramural research would be accelerated by the development of tools to consolidate and make analysis-ready the vast types and storage locations of acquisition data. This work is foundational to the analysis of acquisition strategies and techniques.
- Additional data and data fields will further the need for a scalable system to handle this data.
- More data will drive additional support requirements for data ingestion, processing, quality control, potential automation.
- CAC and overall access issues obstruct the necessary analysis for an IP repository and for future DoD AIRC Digital Data Strategy efforts.
- Access to data will allow the development and application of analytics to the contracting subset and provide the DoD an operational use case to build upon for future AIRC Digital Data Strategy efforts.