NPS-AM-11-186

ACQUISITION RESEARCH SPONSORED REPORT SERIES

A Web Service Implementation for Large-Scale Automation, Visualization, and Real-Time Program-Awareness Via Lexical Link Analysis

27 September 2011

by

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

Prepared for: Naval Postgraduate School, Monterey, California 93943

The research presented in this report was supported by the Acquisition Chair of the Graduate School of Business & Public Policy at the Naval Postgraduate School.

To request Defense Acquisition Research or to become a research sponsor, please contact:

NPS Acquisition Research Program
Attn: James B. Greene, RADM, USN, (Ret.)
Acquisition Chair
Graduate School of Business and Public Policy
Naval Postgraduate School
555 Dyer Road, Room 332
Monterey, CA 93943-5103
Tel: (831) 656-2092
Fax: (831) 656-2253
E-mail: jbgreene@nps.edu

Copies of the Acquisition Sponsored Research Reports may be printed from our website www.acquisitionresearch.net

Abstract

DoD acquisition is an extremely complex system, comprised of myriad stakeholders, processes, people, activities, and organizational structures. Processes within this complex system are encumbered by the continuous creation of large amounts of unstructured and unformatted acquisition program data, which is narrowly useful, yet difficult to aggregate across the "enterprise." Acquisition analysts and decision-makers must analyze this available data to obtain a complete and understandable picture. This is a kind of systems *non-congruence* which has been difficult to overcome. For those embedded within the complexities of the acquisition community, this effort represents a daunting, if not impossible, task. We will apply a data-driven automation system, namely, Lexical Link Analysis (LLA), to facilitate acquisition researchers and decision-makers to recognize important connections (concepts) that form patterns derived from dynamic, ongoing data collection. The LLA technology and methodology is used to uncover and display relationships among competing programs and Navy-driven requirements. In the past year, we tested our method using samples of acquisition data for validity. LLA was demonstrated to discover statistically significant correlations, and automatically extract the links that might require expensive manpower to perform otherwise. This year, we started to develop LLA from a demonstration to an operational capability and facilitate a wider range of acquisition research applications. The resulting methodology can facilitate real-time awareness, reduce the workload of decision-makers, and make a profound impact on the long term success of acquisition strategies by revealing the current status of acquisition programs, and connections within and external to contributing or competing interests, as well as inform potential strategic choices available to decision-makers.

Keywords: Lexical Link Analysis, text mining, data mining, Program Elements, Major DoD Acquisition Programs, Universal Joint Task Lists, resource

allocation, warfighters’ requirement, Urgent Need Statements, unstructured data, data-driven automation

About the Authors

Dr. Ying Zhao is a research associate professor at the Naval Postgraduate School (NPS). Dr. Zhao joined NPS in May 2009. Her research is focused on knowledge management approaches such as data/text mining, Lexical Link Analysis, search and visualization for system self-awareness, decision-making, and collaboration. She received her PhD in mathematics from MIT and co-founded Quantum Intelligence, Inc. She was principal investigator (PI) for six contracts awarded by the DoD Small Business Innovation Research (SBIR) Program. She was the co-author of two U.S. patents in knowledge pattern search from networked agents and in fusion and visualization for multiple anomaly detection systems.

Dr. Ying Zhao
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-3789
Fax: (831) 656-3679
E-mail: ying.zhao@nps.edu

Dr. Shelley Gallup is a research associate professor at the Naval Postgraduate School's Department of Information Sciences, and the director of Distributed Information and Systems Experimentation (DISE). Dr. Gallup has a multidisciplinary science, engineering, and analysis background, including microbiology, biochemistry, space systems, international relations, strategy and policy, and systems analysis. He returned to academia after retiring from naval service in 1994 and received his PhD in engineering management from Old Dominion University in 1998. Dr. Gallup joined NPS in 1999, bringing his background in systems analysis, naval operations, military systems, and experimental methods first to the Fleet Battle Experiment series (1999–2002) and then to the FORCEnet experimentation in the Trident Warrior series of experiments (2003–present).

Dr. Shelley P. Gallup
Information Sciences Department
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-1040
Fax: (831) 656-3679
E-mail: spgallup@nps.edu

Dr. Doug MacKinnon is a research associate professor at the Naval Postgraduate School (NPS). Dr. MacKinnon is the Deputy Director of the Distributed Information and Systems Experimentation (DISE) research group where he leads multi-disciplinary studies ranging from Maritime Domain Awareness (MDA) to Knowledge Management (KM) and Lexical Link Analysis (LLA). He also led the assessment of Tasking, Planning, Exploitation, and Dissemination (TPED) process during field experiments Empire Challenge 2008 and 2009 (EC08/09). He holds a PhD from Stanford University, conducting successful theoretic and field research in Knowledge Management (KM). He has served as the Program Manager for two major government projects of over $50 million each, implementing new technologies while reducing manpower requirements. He has served over 20 years as a naval surface warfare officer, amassing over eight years at sea and serving in four U.S. Navy warships with five major, underway deployments.

Dr. Douglas J. MacKinnon
Information Sciences Department and Graduate School of Operational and Information Sciences
Naval Postgraduate School
Monterey, CA 93943-5000
Tel: 831-656-1005
Fax: (831) 656-3679
E-mail: djmackin@nps.edu

NPS-AM-11-186

ACQUISITION RESEARCH
SPONSORED REPORT SERIES

A Web Service Implementation for Large-Scale Automation, Visualization and Real-Time Program-Awareness Via Lexical Link Analysis

27 September 2011

by

Dr. Ying Zhao, Research Associate Professor,

Dr. Shelley P. Gallup, Research Associate Professor, and

Dr. Douglas J. MacKinnon, Research Associate Professor

Graduate School of Operational & Information Sciences

Naval Postgraduate School

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

DoD acquisition is an extremely complex system comprised of myriad stakeholders, processes, people, activities, and organizational structures. Processes within this complex system are encumbered by the development of large amounts of unstructured and unformatted acquisition program data, which, due to its enormity and complexity, is narrowly useful and difficult to aggregate across the enterprise. Acquisition analysts and decision-makers must, however, analyze all types and spectrums of the available data in order to obtain a complete and understandable picture. Considering the work that acquisitions systems must accomplish, there is a lack of internal congruence between multiple points at which the system should have knowledge of itself and of decision-makers who depend on aggregate information. Current information and decision support systems may not readily help overcome this difficulty, and they present users within the acquisition community with information overload and limited situational awareness. We believe that the application of a data-driven automation system—namely, Lexical Link Analysis (LLA)—can facilitate acquisition researchers' data sense-making dilemma and help reveal important connections (concepts) and patterns derived from dynamic, voluminous, and on-going data collection.

In the past two years, we have utilized the LLA method to discover valid associations among disparate, unstructured data sets that would have otherwise required lengthy and expensive man-hours to achieve. The LLA technology and methodology were used to uncover and graphically display relationships among competing programs and to compare their features with Navy-driven requirements. In the past year, we tested our method using samples of acquisition data for visualization and validity.

During the Phase II research period (begun in 2011), we proposed follow-on research to the NPS Acquisition Research Program using Lexical Link Analysis (LLA). The focus was to develop LLA from a demonstration to an operational capability, that is, a web service to facilitate a wider range of acquisition research applications. In Phase II of our research, we achieved the following:

- We developed a web service that integrated the capabilities we explored in Phase I of the research into an operational capability, which links the budgeting process through Program Elements (PEs) to the acquisition process via acquisition programs such as Major DoD Acquisition Programs (MDAPs) and Programs in Acquisition Category II (ACAT IIs), and to the warfighters' requirements such as Urgent Needs Statements (UNSs) and Universal Joint Task Lists (UJTLs). The web service is a real-time operational capability of program awareness, the results of which could be periodically updated and presented in dynamic, 3-D visualizations.
- We applied the LLA web service to authoritative and accurate data sources such as the Defense Technical Information Center (DTIC; http://www.dtic.mil/), Defense Acquisition Management Information Retrieval (DAMIR; http://www.acq.osd.mil/damir/), Acquisition Resources and Analysis (ARA), and Selected Acquisition Report (SAR; http://www.acq.osd.mil/ara/am/sar/).
- We communicated with a community of acquisition professionals at the annual symposium and researched wider applications of our system.

We summarized the LLA methodology into a journal paper (Zhao, Gallup, & MacKinnon, 2011c) in five dimensions: (1) System Self-awareness, (2) Lexical Link Analysis, (3) Visualization, (4) Agent Learning, and (5) Network Analysis. The first represents a global view of the issue, and the other four refer to a set of specific methods and intelligent agent tools we use to resolve analytic needs within very large data sets.

Significance of the Research

Acquisition research has increased in component, organizational, technical, and management complexity. It is difficult for acquisition professionals to remain continuously aware of their decision-making domains because information is overwhelming and dynamic. According to the *Chairman of the Joint Chiefs of Staff Instruction for Joint Capabilities Integration and Development System (JCIDS*; CJCS, 2009), there are three key processes in the DoD that must work in concert to deliver the capabilities required by the warfighters: the requirements process; the acquisition process; and the Planning, Programming, Budget, and Execution (PPBE) process.

Each process produces a large amount of data in an unstructured manner; for example, the warfighters' requirements are documented in UJTLs, Joint Capability Areas (JCAs), and UNSs. These requirements are processed in the JCIDS to become projects and programs, which should result in products such as weapon systems that meet the warfighters' needs. Program data are stored in the Defense Acquisition System (DAS). Programs are divided into MDAPs, ACATIIs, and so forth. PEs are the documents used to fund programs yearly through the congressional budget justification process. All the data is too voluminous, too unformatted, and too unstructured to be easily digested and understood—even by a team of acquisition professionals. There is a critical need for automation to help reveal to decision-makers and researchers the interrelationships within these processes (see Figure 1).

We have attempted to develop and frame our research efforts around research questions in the following categories: conceptual, focused, theory development, and methodology.

Conceptual

- How can the information that emerges from the acquisition process be used to produce overall awareness of the fit between programs, projects, and systems and of the needs for which they were intended?
- If a higher level of awareness is possible, how will that enable system-level regulation of programs, projects, and systems for improvement of the acquisition systems?

Focused

- Based on the normal evolution of documentation and on the current data-based program information, how can requirements (needs) be connected to system capabilities via automation of analysis?
- Can requirements gaps be revealed?

Theory Development

- How can a correlation between system interdependency (links/relationships) and development costs be shown, if present?

Methodology

- How can we use natural language and other documentation (roughly, unformatted data) to produce visualization of the internal constructs useful for management through Lexical Link Analysis (LLA)?

Lexical analysis ("Lexical Analysis," 2010) is a form of text mining in which word meanings are developed from the context from which they are derived. Link analysis, a subset of network analysis that explores associations among objects, reveals the crucial relationships between objects when collected data may not be complete. LLA is an extended lexical analysis and link analysis. LLA can also be used in a learning mode in which such features and contextual associations are initially unknown and are constantly being learned, discovered, updated, and improved as more data become available.

We consider that the cognitive interface between decision-makers and a complex system may be expressed in a range of terms or features (i.e., a specific vocabulary or lexicon) to describe attributes and the surrounding environment of a system. Here, system self-awareness, or program-awareness (Gallup, MacKinnon, Zhao, Robey, & Odell, 2009) allows decision-makers to be aware of what systems, programs, and products are available for acquisition; to understand how the systems match warfighters' needs and requirements; to recognize relationships among them; to improve efficiency of available collaboration; to reduce duplication of effort; and to reuse components to support cost-effective management with greater immediacy, possibly in real-time.

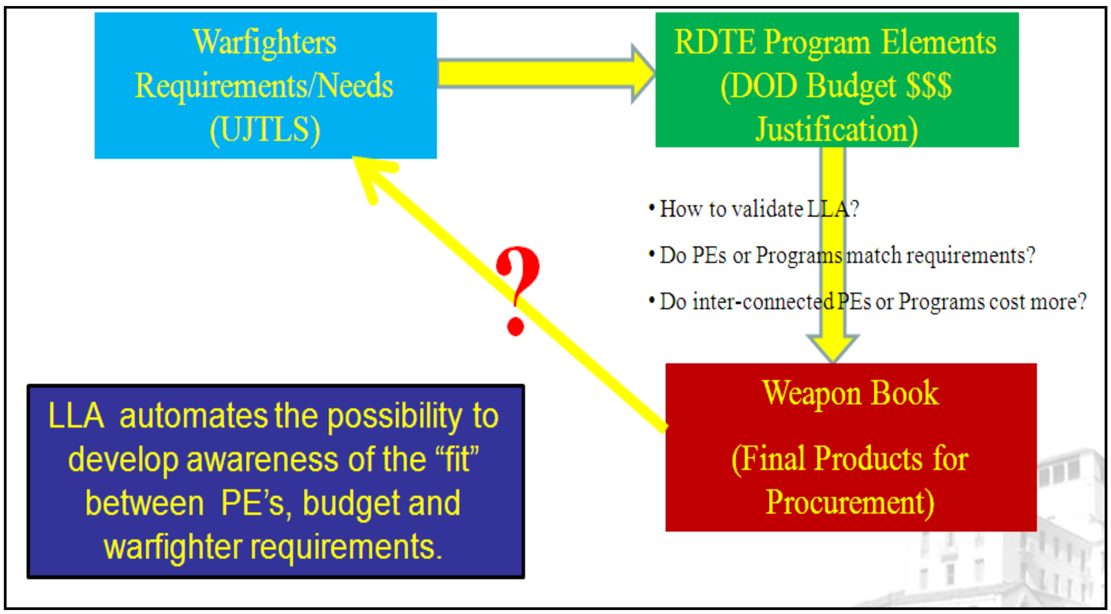


Figure 1. LLA Seeks to Inform the Business Processes Links (e.g., From Requirements to DoD Budget Justification to Final Products) That Are Critical for DoD Acquisition Research

In precise terms, we observed that there were three important processes that seem fundamentally disconnected. They were the congressional budgeting justification process (such as information contained within the PEs), the acquisition process (such as information in the MDAP and ACATII), and the warfighters' requirements (such as information in UNSs and in UJTLs). They were not analyzed and compared to each other in a dynamic, holistic methodology that could keep up with changes and reflect patterns of relationships.

There had been little previous effort to integrate the data in these three components. In Phase I of the project (2009 to 2010), we analyzed in detail samples in the three components, validated the LLA method using the large-scale data sets, and also successfully applied the method to discover the patterns in the data that were interesting and previously unknown to many acquisition professionals (Zhao, Gallup, & MacKinnon, 2010, 2011a, 2011b).

Results for Phase II

During the Phase II research period, begun in 2011, we proposed follow-on research. Our goals for Phase II were as follows:

- Apply LLA to larger-scale data and wider applications and employ parallel computing and dynamic, 3-dimensional (3-D) visualizations.
- Apply LLA to become a real-time operational capability of program awareness, the results of which could be periodically updated and presented in a web service.

We started developing a web service that was designed to integrate the capability we explored in Phase I of the research into an operational capability, which links the budgeting process through PEs, to the acquisition process via acquisition programs (MDAPs, ACATIIs), and to the warfighters' requirements (UNS, UJTL, etc.). We implemented an LLA platform from which to periodically present all the information in a single view so that users can view the trends based on the data in each of the three areas. We gathered the most recent documents in three areas from the following sources:

1. PEs:

   http://www.dtic.mil/descriptivesum/

2. MDAPs & ACATIIs:

   http://comptroller.defense.gov/defbudget/fy2008/fy2008_weabook.pdf

   http://www.fas.org/man/dod-101/sys/land/wsh2007/index.html

   http://www.acq.osd.mil/ara/am/sar/

3. UJTLs:

   http://www.dtic.mil/doctrine/jel/cjcsd/cjcsm/m350004d.pdf

The web service described in Figure 2 dramatically speeded up efforts to collect the data. For example, each of the 24 sets of PE documents contained about 200 PDF PEs from http://www.dtic.mil/descriptivesum/, totaling about 5,000 documents. Manually downloading and extracting desired links would be very time intensive. By submitting several parallel jobs to the Naval Postgraduate School (NPS) High Performance Computing (HPC) Center, the download took approximately six hours.

Web Service Design

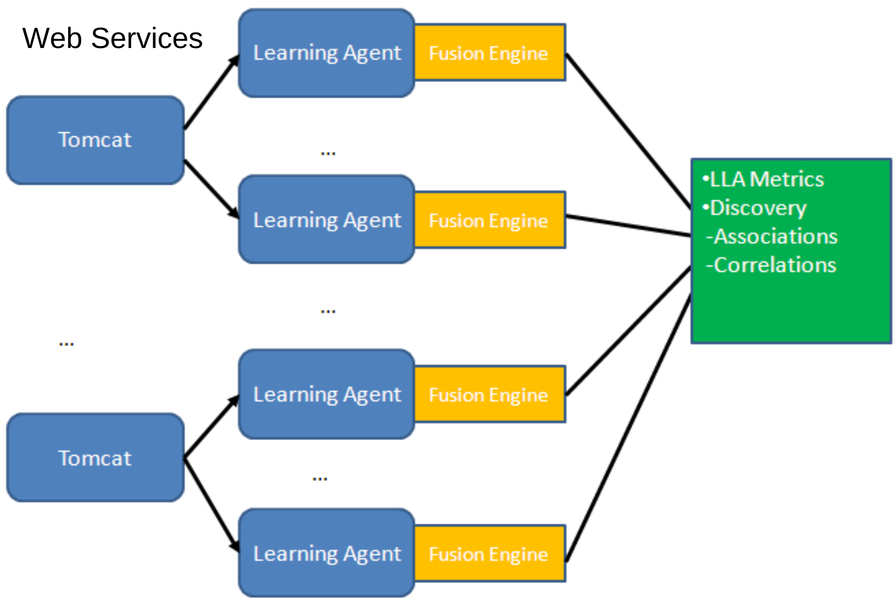


Figure 2. Initial Web Service Design

Figure 2 shows the initial web service design, detailed as follows:

- Tomcat (http://tomcat.apache.org/index.html) was used as the infrastructure to host multiple learning agents for the web service. A Collaborative Learning Agent system (CLA; Quantum Intelligence [QI], 2009) of multiple agents was installed in a single or multiple Tomcat(s). In Figure 3, the ARP web service is shown, hosted via http://disedev4.ern.nps.edu:8080/ARP, which is a dedicated server for this project at the NPS DISE lab. Eventually, we will move the service to the NPS HPC Center, where hundreds of learning agents will be hosted in the cloud computing environment to gather, analyze, and disseminate information in a massive, parallel fashion. The web service administration function includes the following capabilities:
    - Peer List: allows the current agent to list the peers with which it shares index and learning models

- o One-click mining: uses only one click to index and mine the data stored locally
- o Properties: specifies parameters used in the one-click mining
- o Dashboard monitor: displays *lexical links* discovered from the mining process continuously
- o Back to search: provides the capability to allow a basic search

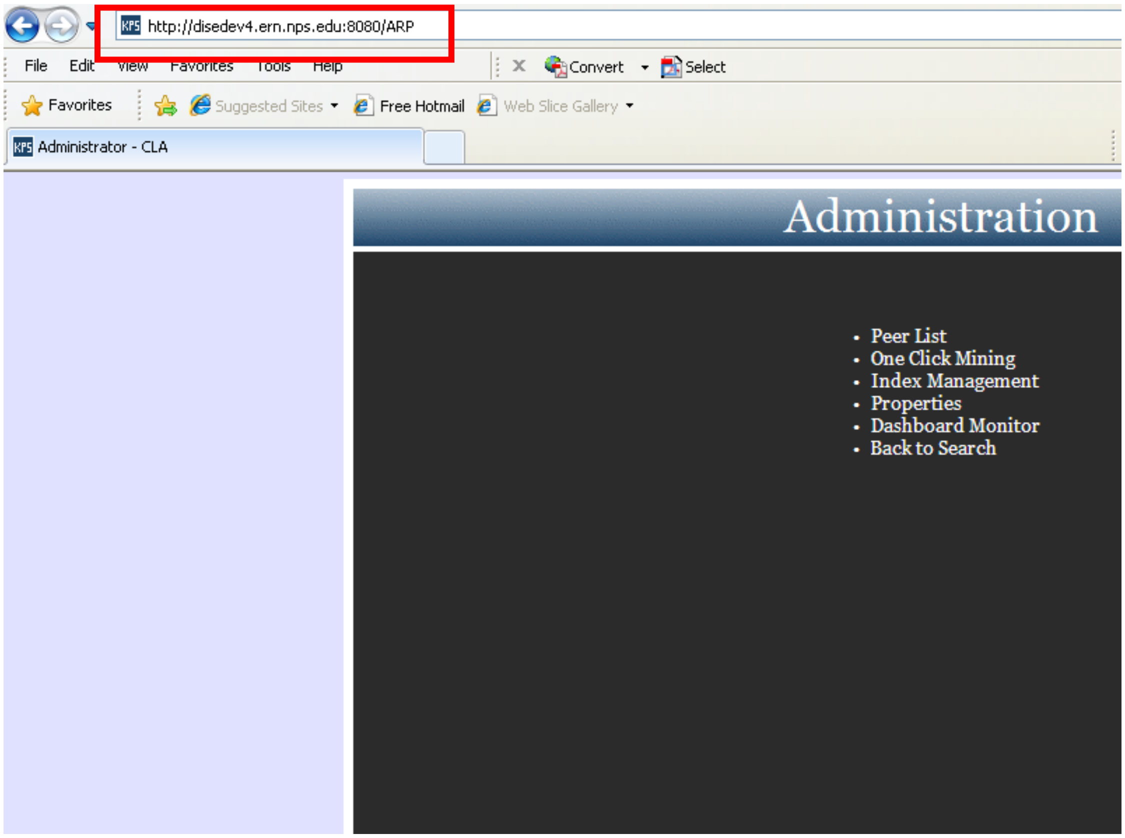


Figure 3. Web Service Hosted Using Tomcat

- A single learning agent was implemented to mine the data that were gathered in each of the categories, for example, PEs of the Air Force in 2011, as shown in the one-click mining capability in Figure 4. “Path to Data” was used to point to the data stored locally. “Index Name” was used to store the search index and learning model generated from the data.

Figure 4. One Click Mining

- The indexes or learning models generated from Figure 4 are stored locally in each learning agent, as shown in the "Index Management" in Figure 5. A fusion engine is attached to a learning agent. The function of the fusion engine is to combine lexical links discovered from the local index/learning model with the lexical links discovered from its peers in a recursive manner, thus forming a combined view of all the data from the total learning agent network. As shown in Figure 5, when "Fuse" is clicked, the indexes/learning models selected (e.g., navy_2009, navy_2010, and navy_2011) were combined into one model.

Figure 5. Fusion Engine

An index or learning model contains the following functions:

- Lexical links are highlighted in the search results, as shown in the dashboard display in Figure 6. When a lexical link is clicked via "Investigate," a search is invoked and the source documents containing the link are listed and highlighted.

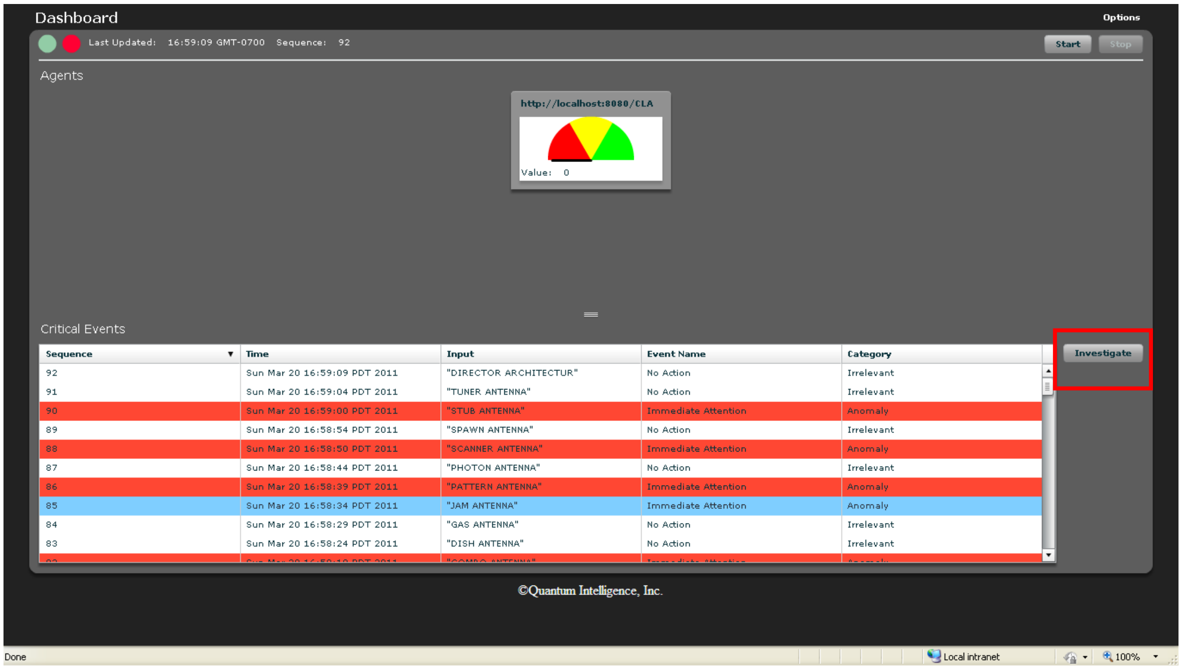


Figure 6. Dashboard to Display Lexical Links Discovered

- The key metrics of *lexical link* counts are used to measure overlaps and gaps between PEs, PEs and other categories of information such as MDAPs, UNS/UJTLs, and changes over time.

The fusion engine described in Figure 5 fuses the learning models and then groups the lexical links into categories to look at the links and overlaps among different services and over years in detail. As shown in Figure 7, a single category (theme), using a triple of word hubs of *Tactic, Combat*, and *Effort* as the category title, contains lexical links related to the category. These lexical links are generated from different data sources of PEs from 2009 to 2011: red—links only in 2011; green—links only in 2010; and blue—links only in 2009. The purple links are the ones that are in more than two sources.

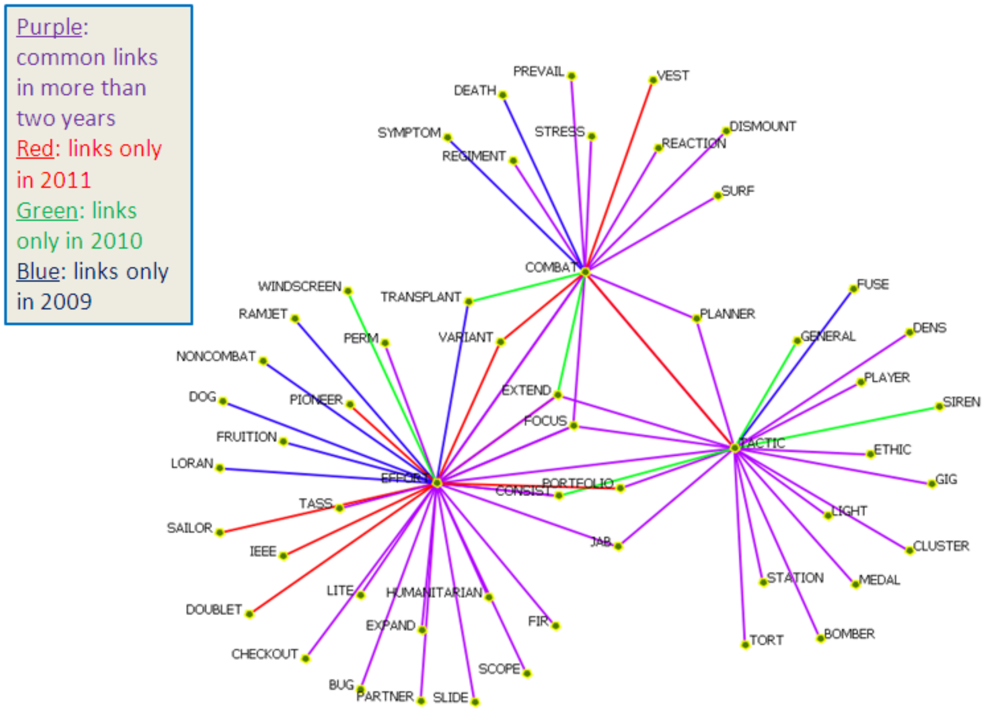


Figure 7. Lexical Links Grouped Into Categories

Figure 8 shows all the groups in one view. Each of the connected links represents a set of features that belong to a group, such as "Tactic–Combat–Effort," shown in Figure 7. As shown in Figure 8, the total number of features, features deleted, and features added (2009 to 2011) were computed, respectively, from the lexical links.

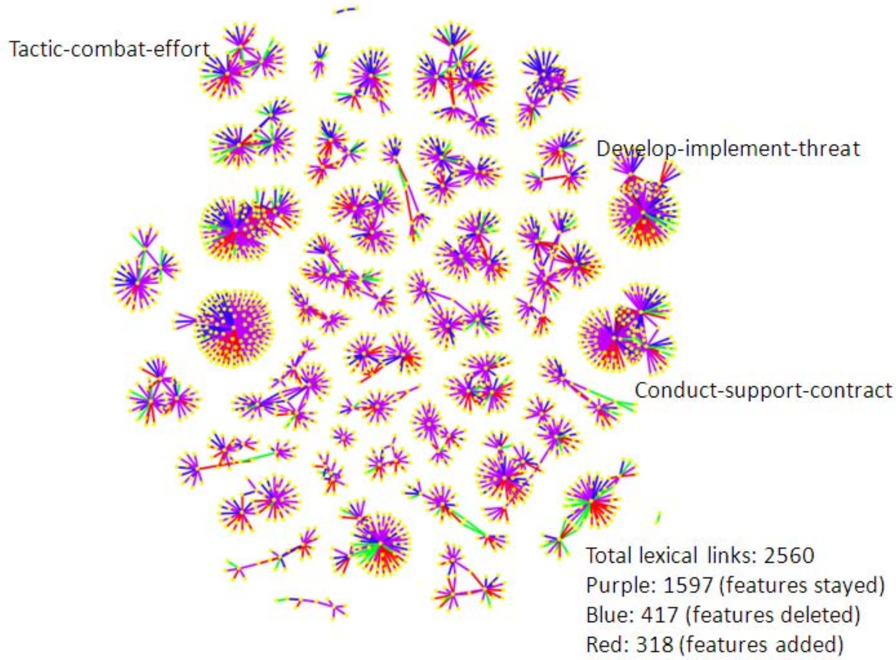


Figure 8. Overall View of Three Years of PEs

LLA networks are visualized using a set of commonly known social network tools such as Organizational Risk Assessment (ORA), shown in Figure 8. Another tool we explored is Pajek (Networks/Pajek, 20081), which is able to export a network in an X3D format and then display it in 3-D. X3D is a product from the Modeling, Virtual Environments, and Simulation (MOVES) Institute at NPS for 3-D visualization and navigation.

Social Network of PEs

We have been using the initial implementation of the LLA web service in the workflow that benefits acquisition professionals. As an example, the fusion engine was used to construct a social network view of PEs. Figures 9 and 10 illustrate the differences between LLA discovered linkages and those found by human analysts. In Figure 9, PE 0603721N is linked to PEs 0602435N, 0602782N, 0601153N, and 0603235N. Figure 10 indicates Program Elements (PEs) identified by human analysts. Titles for the PEs are as follows:

- 0602435N: Ocean Warfighting Environment Applied Research;
- 0602782N: Mine and Expeditionary Warfare Applied Research;
- 0601153N: Defense Research Sciences; and
- 0603235N: Common Picture Advanced Technology.

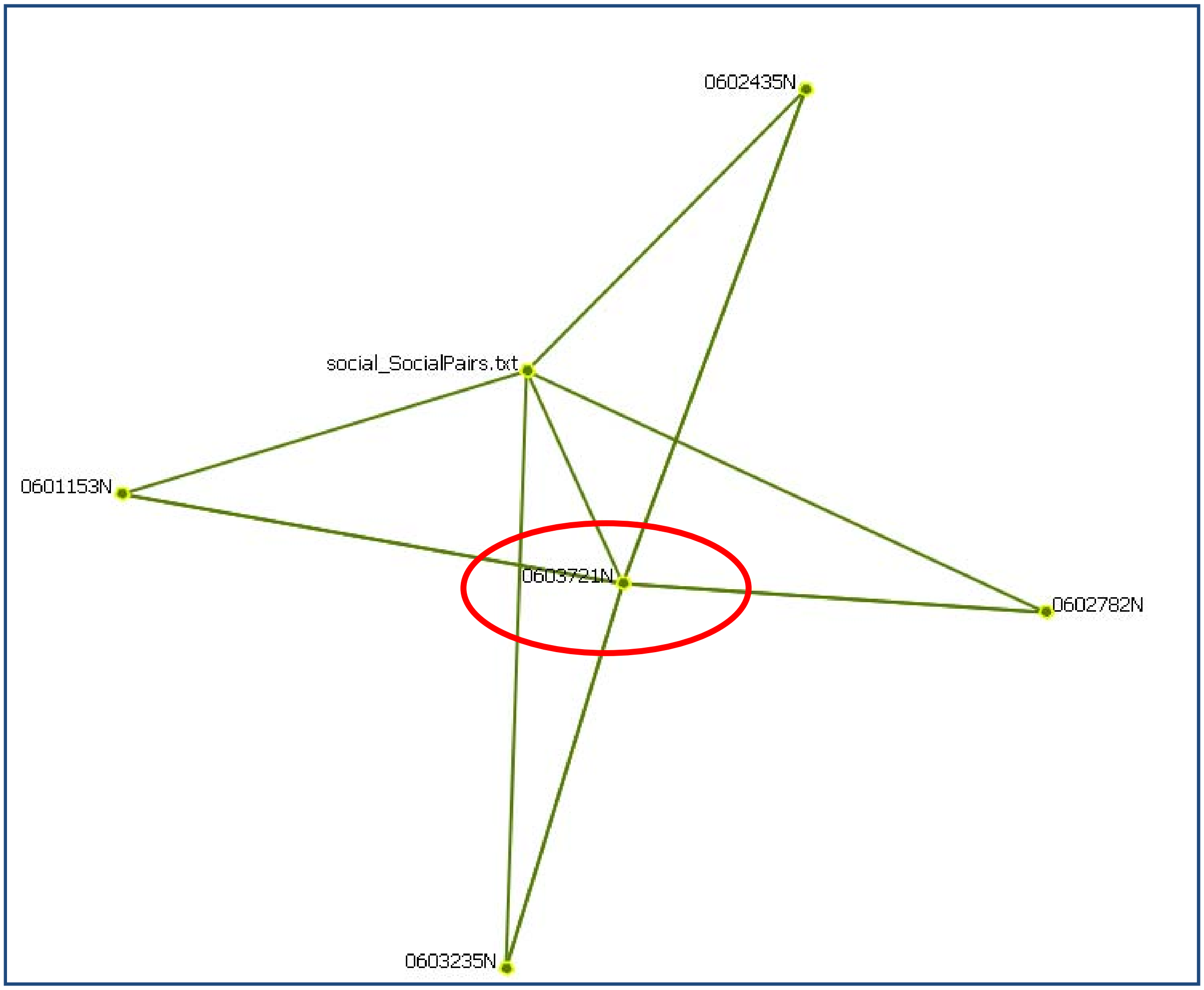
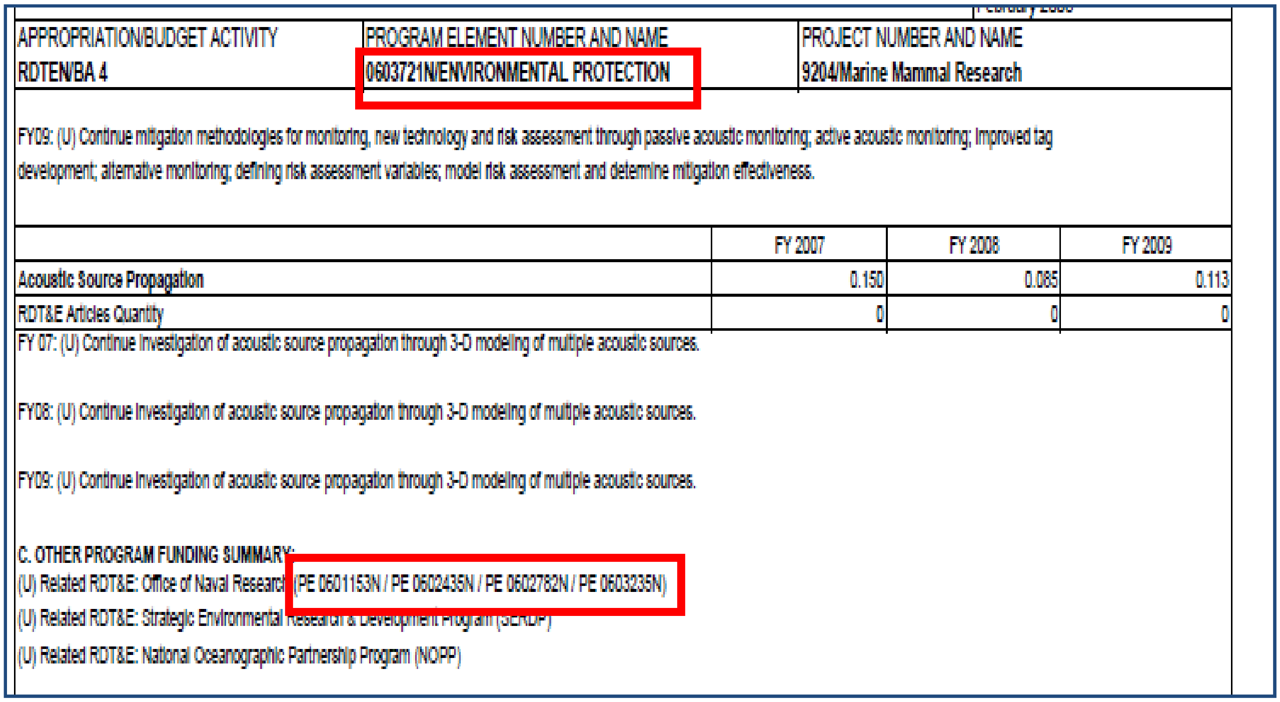


Figure 9. Social Network of PE 0603721N



APPROPRIATION/BUDGET ACTIVITY	PROGRAM ELEMENT NUMBER AND NAME	PROJECT NUMBER AND NAME
RDTEN/BA 4	0603721N/ENVIRONMENTAL PROTECTION	9204/Marine Mammal Research

FY09: (U) Continue mitigation methodologies for monitoring, new technology and risk assessment through passive acoustic monitoring; active acoustic monitoring; Improved tag development; alternative monitoring; defining risk assessment variables; model risk assessment and determine mitigation effectiveness.

	FY 2007	FY 2008	FY 2009
Acoustic Source Propagation	0.150	0.085	0.113
RDT&E Articles Quantity	0	0	0

FY 07: (U) Continue investigation of acoustic source propagation through 3-D modeling of multiple acoustic sources.

FY08: (U) Continue investigation of acoustic source propagation through 3-D modeling of multiple acoustic sources.

FY09: (U) Continue investigation of acoustic source propagation through 3-D modeling of multiple acoustic sources.

C. OTHER PROGRAM FUNDING SUMMARY:
(U) Related RDT&E: Office of Naval Research (PE 0601153N / PE 0602435N / PE 0602782N / PE 0603235N)
(U) Related RDT&E: Strategic Environmental Research & Development Program (SERDP)
(U) Related RDT&E: National Oceanographic Partnership Program (NOPP)

Figure 10. PE 0603721N Linked to PEs Identified by Human Analysts

Semantic Network of PEs

Compared to the links identified by human analysts, LLA was used to look into the links among PEs from all of the Services as a whole system; therefore, the links discovered were cross-Service and potentially overcame the cognitive *blind spots* of human analysts. For example, Table 1 lists the semantic network for PE 0603721N discovered by LLA. Three of four human identified links showed up in the top 100 of the LLA links, with 0601153N, 0602435N, and 0603235N ranked 33, 35, and 58, respectively.

Figure 11 shows a total social network view of the PEs using the links identified by human analysts for all the PEs for the 2008 data and a 3-D view from Pajek. PEs ending with an *A* were Army PEs, PEs ending with an *F* were Air Force PEs, and PEs ending with an *N* were Navy PEs. As one can observe, the links in Figure 11 tended to be within the Services; for example, analysts tended to identify Army PEs linked to Army PEs, Air Force to Air Force, and Navy to Navy. The cost of each PE in 2008 is illustrated as the bubble size. As seen in Figure 11, PEs within the Services were more cross-referenced, and the cost seemed inversely correlated to the links.

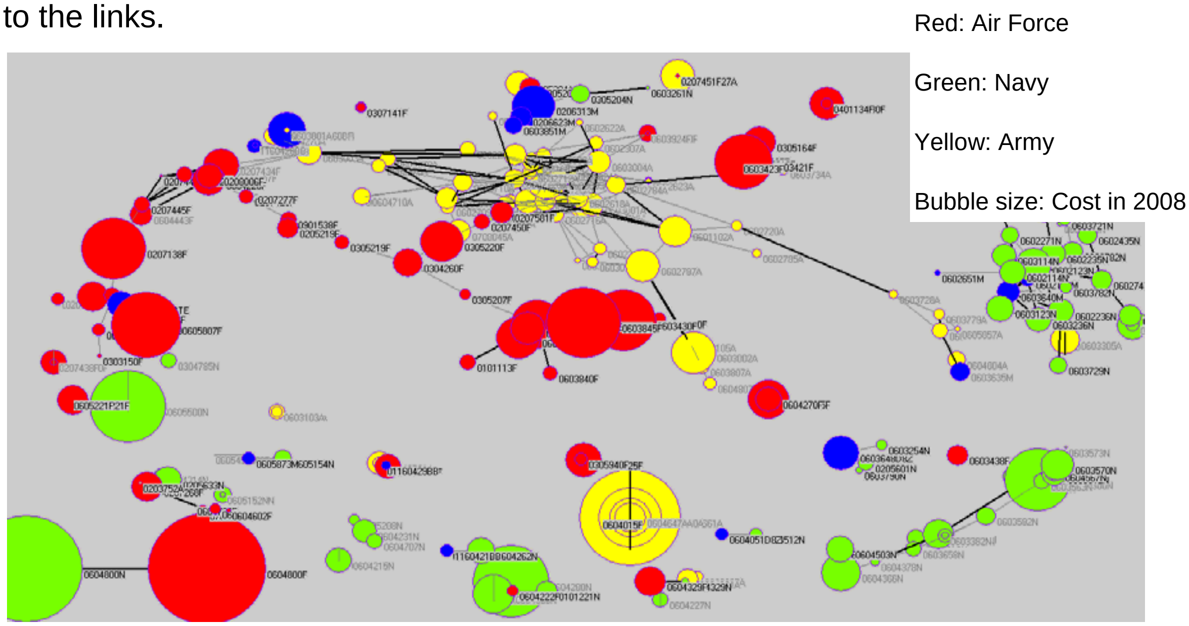


Figure 11. A Social Network View of PEs With the Links Identified by Human Analysts—A 3-D View From Pajek

Figure 12a and Figure 12b show the social network and semantic network 3-D views of all the PEs for the 2008 and 2009 data using Pajek. The cost ratio of each PE in 2009 and 2008 is illustrated as the bubble size. The purple box shows a program that has a ratio of 1, indicating no changes of cost from 2008 to 2009. As shown in Figure 12b, which is laid out by the free energy of the network connections, with the more connected programs in the middle, larger sizes of nodes tend to be on the outside, indicating the correlation between independencies of programs and cost increases. The social network links marked by human analysts in Figure 12a do not reveal this pattern.

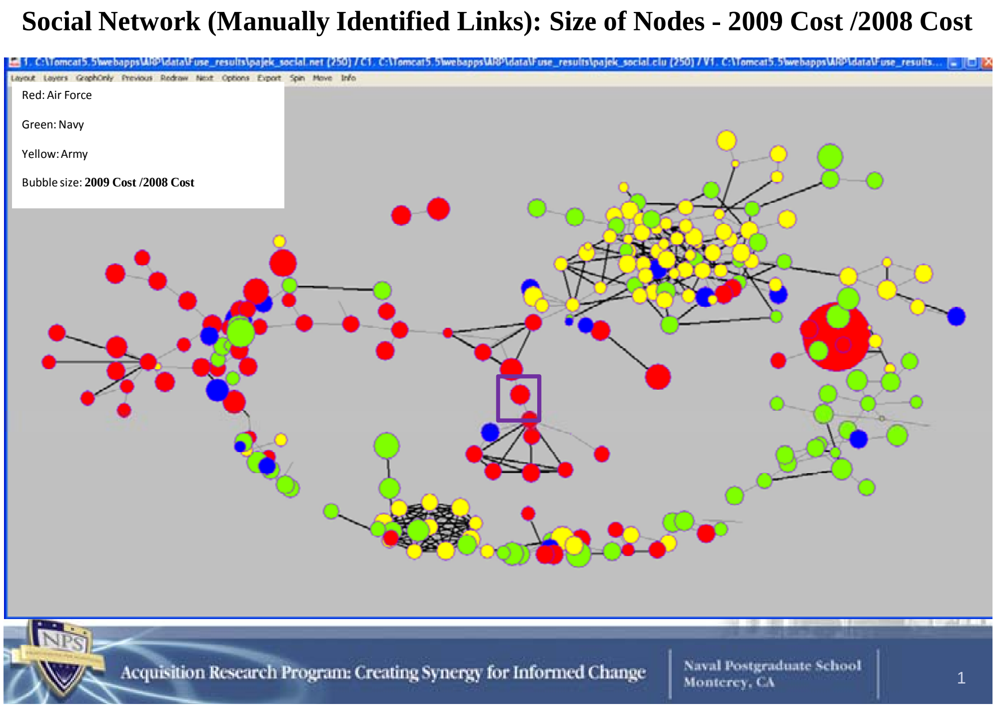


Figure 12a. A 3-D View of PEs Identified by the Human Social Network

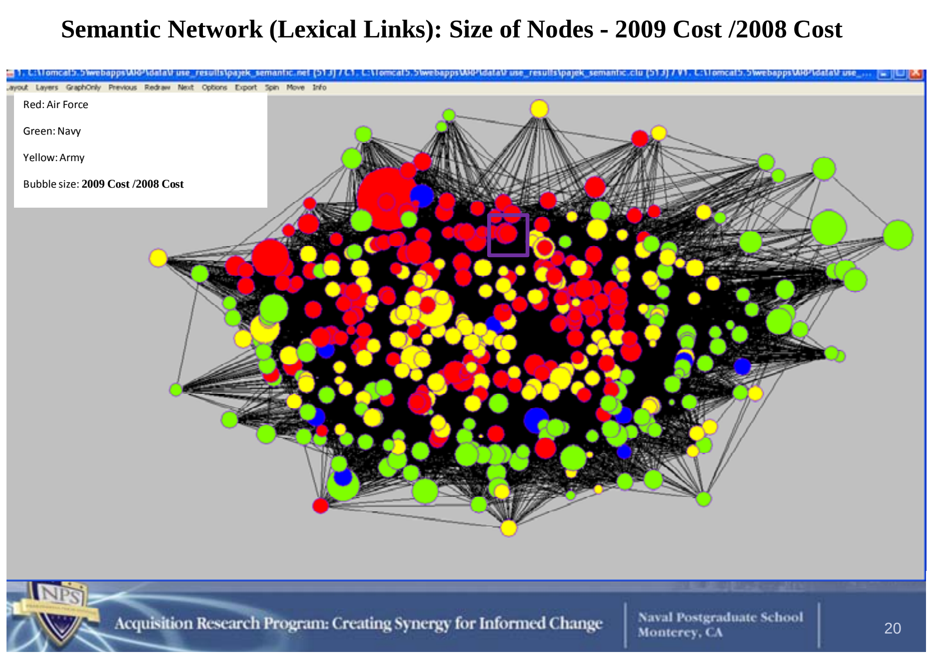


Figure 12b. A 3-D View of PEs Identified by the LLA Semantic Network

Table 1. Semantic Network for PE 06043721N

1	0603721N	
2	0000012.74;JET--	0602787A
3	0000012.52;SEDIMENT--	0601102A
4	0000012.43;CLEAN--	0603804A
5	0000011.78;DESTRUCT--	0602203F
6	0000010.77;SHIELD--	0602102F
7	0000010.48;JET--	0601102F
8	0000010.33;CLEAN--	0604804A
9	0000009.67;DESTRUCT--	0604645A
10	0000008.90;SHIELD--	0604231N
11	0000008.49;UNIFORM--	0206313M
12	0000008.15;GAIN--	0603789F
13	0000008.06;ARLINGTON--VA,NJ;RIGOR--	0605013A
14	0000007.94;ARLINGTON--VA,NJ;LABOR--	0603747A
15	0000007.93;CONTENT--	0602202F
16	0000007.72;GAIN--	0603001A
17	0000007.69;JET--	0603790F
18	0000007.68;SHIELD--	0603640M
19	0000007.36;JET--	0603216F
20	0000007.11;SHIELD--	0602601F
21	0000007.11;CLEANUP--	0603728A
22	0000007.05;FINISH--	0605857A
23	0000006.98;UNIFORM--	0601104A
24	0000006.89;SHIELD--	0603005A
25	0000006.73;SEDIMENT--	0602236N
26	0000006.63;ARLINGTON--	0305204A
27	0000006.62;UNIFORM--	0604601A
28	0000006.62;CONTENT--	0602702F
29	0000006.60;DESTRUCT--	0604759F
30	0000006.59;FINISH--	0604661A
31	0000006.55;GOVT--	0604240F
32	0000006.43;GAIN--	0602120A
33	0000006.34;DESTRUCT--	0601153N
34	0000006.22;GAIN--	0604321A
35	0000006.13;SEDIMENT--	0602435N
36	0000005.98;EXPECT--	0602204F
37	0000005.88;STORM--	0207601F
38	0000005.85;GAIN--	0603231F
39	0000005.85;GAIN--	0303140F
52	0000005.01;NORMAL--	0207410F
53	0000004.98;DESTRUCT--	0603004A
54	0000004.93;LABOR--	0605801A
55	0000004.87;CONCERN--	0602747N
56	0000004.86;SHIELD--	0603561N
57	0000004.86;JET--	0603236N
58	0000004.80;SHIELD--	0603235N
59	0000004.76;JET--	0602618A
60	0000004.76;DESTRUCT--	0604660A
61	0000004.75;NORMAL--	0305206F
62	0000004.65;AGREEMENT--	0207418F
63	0000004.61;GAIN--	0604805A
64	0000004.56;JET--	0605805A
65	0000004.54;ARLINGTON--	0203758A
66	0000004.45;BREED--	0207451F
67	0000004.40;GOVT--	0604215N

In addition to the potential to discover human analysts' blind spots in connecting PEs across the Services, we also observed that LLA might discover rare features that two PEs might share. Table 2 shows examples of these links using the highlighted word hubs in Table 1 for the top four PEs linked to PE 06043721N.

Table 2. Unique and Rare Semantic Links

Top 4 PEs linked to PE 06043721N	Titles	Semantic Links
0602787A	Medical Technology	Jet lag, jet fuel exposure
0601102A	Defense Research Sciences	Destruction, containment in water, soil, and sediments resulting from military activities
0603804A	Logistics and Engineer Equipment	The Army fights with clean fuel and drinking water
06032203F	Aerospace Propulsion	Non-destructive test, fuels and lubrication

Observations for the RDT&E Budget Justification Process

We took a detailed look at the Research, Development, Test, and Evaluation (RDT&E) budget modification practice from 2008 to 2009, in an effort to see if LLA links identified among PEs and JTLs are correlated with the changes in the budget allocation from 2008 to 2009. Our observations are summarized in Table 3.

We observed that from 2008 to 2009, as shown in Table 3, the average 2009 budget change, in terms of percentage change for each PE whose number of LLA links to other PEs was larger than 10, was 14%, compared to 40% whose number of LLA links to other PEs was fewer than 10. The total 2009 cost change was $558 million for the former, and $434 million for the latter. This indicates the practice tended to reduce the budget for PEs with more links to other PEs, and to increase the budget for the ones with less links, allocating resources to avoid overlapping efforts and to fund new and unique projects.

Table 3. Budget Change Sorted Using LLA Links From PEs to PEs

LLA links from PE to PE	Average Budget Change from 2008 to 2009 (in terms of percentage change for each PE)	Total Budget Change in Millions
>10	14%	($558)
<=10	40%	$434

In contrast, the same 450 PEs sorted according to the numbers of LLA links with respect to UJTLs are shown in Table 4. Overall, there were fewer numbers of LLA links observed, meaning that there were gaps between the RDT&E resource allocation and the warfighters' requirements. For PEs which had at least one LLA match to UJTLs, the average percentage cost change was 10%, compared to 29% for PEs which had no matches. This indicated a need to consider gaps and warfighters' requirements as priorities in the RDT&E investment.

We found that the total cost change for PEs with at least one match to the UJTLs was $735 million, compared to $859 million for PEs with no matches. We found this was due to the current practice which tended to cut the budget of the more expensive programs, such as MDAPs, rather than the less expensive ones.

Table 4. Budget Change Sorted Using LLA Links From PEs to UJTLs

LLA links of PE to UJTL	Average Budget Change from 2008 to 2009 (in terms of percentage change for each PE)	Total budget change in millions
>1	10%	$735
<=1	29%	($859)

These findings can be useful as validation and guidance for implementing Secretary of Defense Gates' defense cutting plan. For example, Secretary Gates said the Pentagon must get "more bang for its buck and shift its focus to the military's needs for the future" (Hedgpeth, 2010, p. 1). Top acquisition officials in the nation have been looking for ways to limit spending, identify efficiencies, and eliminate unnecessary cost. Secretary Gates also planned to add 20,000 acquisition workers to implement the cost reduction. The program awareness implemented via the LLA method can link warfighters' requirements to the budget and to final weapon products, and can help all the acquisition workers in their decision-making. The use of the LLA method creates an opportunity for new acquisition workers to reduce the overall inefficiency of the 10% cost change, as opposed to the 29% cost change, as illustrated in Table 4, which focused mainly on the big ticket items such as MDAPs.

THIS PAGE INTENTIONALLY LEFT BLANK

Status After the Symposium

Since the annual acquisition research symposium in May 2011, we have accomplished the following items:

- We summarized the LLA methodology in a journal paper (Zhao et al., 2011c) in five dimensions that are briefed in the following sections: 1) System Self-awareness, 2) Lexical Link Analysis, 3) Visualization, 4) Agent Learning, and 5) Network Analysis. The first represents a global view of an issue, and the other four refer to a set of specific methods and intelligent agent tools we use to resolve analytic needs within very large data sets.
- We prepared a Phase III proposal and tasks for FY2012.
- We started to work with potential case studies contacts to gather the data and prepare the analysis.

Summary of the Methodology

System Self-Awareness

We borrow from notions of *awareness* and advance the term *self-awareness* of a complex system as the collective and integrated understanding of system capabilities, or *features*. A related term, *situational awareness*, is used in military operations and carries with it a sense of immediacy and cognitive understanding of the warfighting situation. Here, system self-awareness, in the acquisition context, is a *program-awareness* (Gallup et al., 2009; Zhao et al., 2010, 2011a, 2011b), which allows decision-makers to be aware of the systems, programs, and products that are available for acquisition; to recognize relationships among them; to improve the efficiency of available collaboration; to reduce the duplication of effort; and to re-use components to support cost effective management—with greater immediacy, possibly in real-time.

Lexical Link Analysis

Lexical Link Analysis (LLA) is an innovative extension of lexical analysis combined with link analysis, and employs enabled agent learning technology. The following are the steps for performing an LLA:

1. Read each set of documents.
2. Select feature-like word pairs.
3. Apply a social network algorithm to group the word pairs into clusters or themes. A theme includes a collection of lexical word pairs connected to each other.
4. Compute a "weight" for a theme for the information of a time period, that is, the number of word pairs that belong to a theme for that time period.
5. Sort theme weights by time, and study the distributions of the themes by time.

Visualization

We have been generating visualizations, including a lexical network visualization using various open source tools. We began by using the Organizational Risk Assessment (ORA; Center for Computational Analysis of Social and Organizational Systems [CASOS], 2009) tool and expanded to other tools. For example, in the past year, we developed 3-D network views using Pajek (Networks/Pajek, 2008) and X3D (X3D, 2011). We also developed our visualizations Radar view and Match view (Zhao et al., 2010).

Unsupervised Agent Learning

LLA uses a computer-based learning agent called Collaborative Learning Agents (CLA; QI, 2009) to employ an unsupervised learning process that separates patterns and anomalies. CLA is a computer-based learning agent, or agent collaboration, capable of ingesting and processing data sources, leveraged via an educational license with Quantum Intelligence, Inc. The unsupervised agent learning is outlined in the following steps:

1. Index each set of documents separately and in parallel using multiple learning agents. Multiple agents can work collaboratively and in parallel. We set up a cluster utilizing Linux servers in the NPS High Performance Computing (HPC) Center to handle the large-scale data and secure environment in the NPS Secure Technology Battle Laboratory (STBL).

2. Apply context lists for entity extraction: using word juxtaposition, context lists are provided initially to specify the contexts for who (people), where (location), and what (action).

3. Generate social networks based on entities extracted. The relation types are people-to-people, location-to-location, action-to-action, people-to-location, people-to-action, and location-to-location. Each relationship is linked with a set of lexical terms that are discovered automatically from the data.

4. Generate semantic networks based on lexical links from the text documents that do not contain the entities extracted from the previous steps.

5. Apply visualization and network analysis highlighted to analyze the extracted networks from Steps 1 to 4. Semantic networks, combined with the people social networks, will characterize the behavior, such as actions and events, of potential high-value targets.

Social and Semantic Network Analysis

Current research of social network analysis mostly focuses on people or organizations of direct associations, regardless of the contents linked. The so-called study of centrality (Feldman, 2007; Girvan, 2002) has been a focal point for the social network structure study. Finding the *centrality* of a network lends insight into the various roles and groupings, such as the connectors, the clusters, the network core, and its periphery. We have been working towards the following three areas of innovations in the network analysis:

- Extract social networks based on the entity extraction.
- Extract semantic networks based on the contents and word pairs using LLA.
- Apply characteristics and centrality measures from the semantic networks and social networks to predict latent properties such as

emerging techniques that might dominate in the future. The characteristics are further categorized into themes and time-lined trends for the prediction of future events.

Anticipated Benefits of Our Approach

The LLA method provides the solutions to meet the critical needs of acquisition research. The key advantage is to provide an innovative, near real-time self-awareness system to transfer diversified data services into strategic decision-making knowledge, detailed as follows:

- Automation: High correlation of LLA results with the link analysis done by human analysts makes it possible for automation, saving human power, and improving responsiveness. Automation is achieved via computer program or software *agent(s)* to perform LLA frequently—and in near real-time. Agent learning makes it possible to reach real-time; visualization correlates lexical links to core measures; features and patterns are discovered over time for the system as a whole. We can take advantage of the data in motion (Twitter and social media sites) and RSS feed data to build a better picture of real-time program awareness.
- Discovery: It "discovers" and displays a network of word pairs. These word pair networks are characterized by one-, two-, or three-word themes. The weight of each theme is determined based on its frequency of occurrence. It may also discover blind spots of human analysis that are caused by the overwhelming data for human analysts to go through.
- Validation: As we continue validating LLA by direct correlation with human analysts' results, we recognize that using LLA to validate human analysis is yet another advantage of our methodology. For instance, LLA may provide different perspectives of links. In the acquisition context, links discovered by human analysts may emphasize component/part connections. They do not necessarily reflect the content overlaps; therefore, interdependencies of the programs identified by human analysts, for example, program managers, might help the programs to stay funded from year to year for the benefit of continuing the program itself, not cost reduction for the government. LLA looks for overlapping of the contents in order to improve affordability and meet the requirements of warfighters. Consequently, it provides better results in terms of trust, quality of association discovery, breakthrough in the taxonomy of ignorance,

organizational boundaries, and organizational reach (Denby & Gammack, 1999).

LLA is related to a number of extant tools for text mining, including keyword analysis and tagging technology (Foltz, 2002), and intelligence analysis ontology for cognitive assistants (Tecuci et al., 2007). What results from this process is a learning model—like an ethnographic *code book* (Schensul, Schensul, & LeCompte, 1999). LLA, conducted over time, is related to the discourse space using quadratic assignment procedures (QAP; Hubert & Schultz, 1976).

A similar approach, such as the AutoMap (Carley, 2007), uses dynamic network analysis tools to process unstructured data. Although it provides a user friendly interface to visualize social networks and compute various methods related to the dynamic network analysis, speed and scalability is the problem of AutoMap, which was tested on small data sets.

LLA is unique in the ability to construct these linkages discovered via intelligent agents using social network grouping methods, thus revealing underlying themes found within structured and unstructured data. When compared with static word ontology for matching meaning, such as WordNet (2011), developed at Princeton University, a lexical dictionary of English terms and their relationships derived manually as a static database over a period of time, our approach is dynamic, data-driven, and domain-specific. Our methods, if conducted frequently and automatically, can reveal trends of the central themes over time, thus providing much needed situational awareness.

Another common approach in text analysis is Latent Semantic Analysis (LSA; (Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988; Gorman, Foltz, Kiekel, Martin, & Cooke, 2003; Letsche & Berry, 1997) and Probabilistic Latent Semantic Analysis (PLSA). A document is considered to be composed of a collection of words—a "bag of words," where word order and grammar are not considered important. A recent development related to this method is called Latent *Dirichlet* allocation (LDA; Blei & Lafferty, 2007; Blei, Ng, & Jordan, 2003;), which is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a statistical distribution (Dirichlet distribution) over the corpus. Our theme generation from LLA is different than LDA, in which a collection of lexical terms are connected to each other semantically, as if they are in a social community, and social network grouping methods are used to group the words.

Plan for FY2012

The research we have proposed for FY2012 will extend our previous work in the following ways:

1. Build at least two use cases of applications of Lexical Link Analysis Web Service for large-scale automation, validation, discovery, visualization, and real-time program awareness.
2. Demonstrate the methodology for assisting the DoD-wide effort of integrating and maintaining authoritative and accurate acquisition data services in both legacy and new platforms.

The following are potential use cases for FY12:

1. Integrate with authoritative and accurate data. We plan to work with Mr. Mark Krzysko, who is the Deputy Director from the Enterprise Information & OSD Studies, Office of the Under Secretary of Defense for Acquisition, Technology & Logistics (OUSD[AT&L]). The OUSD[AT&L] provides the DoD-wide acquisition community with authoritative and accurate data services. Mr. Krzysko mentioned that currently, the DTIC, DAMIR (http://www.acq.osd.mil/damir/), ARA (http://www.acq.osd.mil/ara), and SAR (http://www.acq.osd.mil/ara/am/sar/) are good sources. *Requirements* data are not included yet. Krzysko stated that applying analytic tools such as LLA to data services will dramatically improve the quality of data, because the automatic analytic methods will not only discover new patterns that are previously unknown, but will also be able to examine the quality of existing data services systematically. It helps identify *bad* data and data independencies that could result from poorly collected field data and integration processes. The OUSD[AT&L] is also interested in semantic links discovered and correlated to numerical measures. We will work with the organization to improve web services, including the capabilities, as follows:

- Ingest authoritative, accurate data sources from legacy and new platforms.

- Visualize and report analytics including lexical, semantic, and social links for the data. Correlate with core numerical metrics (costs, schedules) periodically and in real-time.

- Influence how data are gathered and collected in the future, identify core metrics, and identify bad data links and program interdependencies.

2. Analysis of the Acquisition Research Program data: We will work with the NPS Acquisition Research Program. We will build a use case of Lexical Link Analysis using all the acquisition research publications; for example, we will build acquisition lexicons, links, and themes over time (i.e., from 2003 to now). We have downloaded about 740 publications from the website http://www.acquisitionresearch.net and prepared for the analysis.

3. Acquisition risk analysis: We will work with the MITRE Corporation for acquisition. We will work with the organization using LLA for the MITRE's projects, for example, Experimenting with Acquisition Strategies Using Gaming, and Composable Capability on Demand (CCOD) applications. MITRE has a list of keywords and requirements that they believe could form the basis of match matrix summaries derived from large collections of program documents. These documents will be categorized into risk areas that might contribute to the ultimate success of acquisition, which can be detected earlier.

THIS PAGE INTENTIONALLY LEFT BLANK

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent *Dirichlet* allocation. *Journal of Machine Learning Research, 3*, 993–1022. Retrieved from http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

Carley, K. M. (2007). Improved data extraction and assessment for dynamic network analysis. In *U.S. Army Research Institute Program in Basic Research—FY 2007* (Special Report 66). Retrieved from http://www.hqda.army.mil/ari/pdf/S_66.pdf

Center for Computational Analysis of Social and Organizational Systems (CASOS). (2009). AutoMap: Extract, analyze and represent relational data from texts. Retrieved from http://www.casos.cs.cmu.edu

Chairman of the Joint Chiefs of Staff (CJCS). (2009). *Chairman of the Joint Chiefs of Staff instruction for joint capabilities integration and development system (JCIDS*; J-8 CJCSI 3170.01G). Retrieved from http://www.intelink.sgov.gov/wiki/JCIDS

Dahmann, J., Baldwin, K., Bergin, D., Choudhary, A., Dubon, A., & Eiserman, G. (2005). Matrix mapping tool (MMT; White paper). Washington, DC: OUSD AT&L/Defense Systems.

Denby, E., & Gammack, J. (1999). Modelling ignorance levels in knowledge-based decision support. Retrieved from http://wawisr01.uwa.edu.au/1999/DenbyGammack.pdf

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing* (pp. 281–285). New York, NY: Association for Computing Machinery.

Feldman, R. (2007). Link analysis and text mining: Current state of the art and applications for counter terrorism. In *NATO Advanced Study Institute on Mining Massive Data Sets for Security*. Video retrieved from http://videolectures.net/mmdss07_feldman_latm/

Foltz, P. W. (2002). Quantitative cognitive models of text and discourse processing. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *The handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum.

Gallup, S. P., MacKinnon, D. J., Zhao, Y., Robey, J., & Odell, C. (2009, October 6–8). Facilitating decision making, re-use and collaboration: A knowledge management approach for system self-awareness. In *Proceedings of the* International Joint Conference on Knowledge Discovery, Knowledge

Engineering, and Knowledge Management (IC3K). Madeira, Portugal: INSTICC Press.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. A., & Cooke, N. J. (2003). Evaluation of latent semantic analysis-based measures of communications content. In *Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Hedgpeth, D. (2010, September 14). Gates starts outlining cuts to save $100 billion for defense. *Washington Post*. Retrieved from http://www.washingtonpost.com

Hubert, L. & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychology*, *29*, 190–241.

Letsche, T .A., & Berry, M. W. (1997). Large-scale information retrieval with latent semantic indexing. *Information Sciences, 100*, 105–137.

Lexical analysis. (2010). In *Wikipedia*. Retrieved September 2010 from http://en.wikipedia.org/wiki/Lexical_analysis

Networks/Pajek: Program for large network analysis. (2008). Retrieved from http://vlado.fmf.uni-lj.si/pub/networks/pajek/

Quantum Intelligence (QI). (2009). Collaborative learning agents (CLA). Retrieved from http://www.quantumii.com/qi/cla.html

Schensul, J. J., Schensul, S. L., & LeCompte, M. D. (1999). *Essential ethnographic methods: Observations, interviews and questionnaires*. Lanham, MD: Rowman Altamira.

Tecuci, G., Boicu, M., Marcu, D., Boicu, C., Barbulescu, M., Ayers, C., & Cammons, D. (2007). Cognitive assistants for analysts. In J. Auger & W. Wimbish (Eds.), *Proteus futures digest: A compilation of selected works derived from the 2006 Proteus workshop*. Carlisle Barracks, PA: National Intelligence University, Office of the Director of National Intelligence, and U.S. Army War College Center for Strategic Leadership.

WordNet: A lexical database for English. (2011). Retrieved from http://wordnet.princeton.edu/

X3D. (2011). Open standards for real-time 3D communication. Retrieved from http://www.web3d.org

Zhao, Y., Gallup, S., & MacKinnon, D. (2010). Towards real-time program awareness via lexical link analysis. In *Proceedings of the Seventh Annual Acquisition Research Program*. Monterey, CA: Naval Postgraduate School.

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011a, May). A web service implementation for large-scale automation, visualization and real-time program-awareness via lexical link analysis. In *Proceedings of the Eighth Annual Acquisition Research Program*. Monterey, CA: Naval Postgraduate School.

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011b). *Towards real-time program awareness via lexical link analysis* (NPS-AM-10-174). Monterey, CA: Naval Postgraduate School.

Zhao, Y., Gallup, S. P., & MacKinnon, D. J. (2011c, September). System self-awareness and related methods for improving the use and understanding of data within DoD. *Software Quality Professional, 13*(4), 19–31. Retrieved from http://asq.org/pub/sqp/

THIS PAGE INTENTIONALLY LEFT BLANK

2003 – 2011 Sponsored Research Topics

Acquisition Management

- Acquiring Combat Capability via Public-Private Partnerships (PPPs)
- BCA: Contractor vs. Organic Growth
- Defense Industry Consolidation
- EU-US Defense Industrial Relationships
- Knowledge Value Added (KVA) + Real Options (RO) Applied to Shipyard Planning Processes
- Managing the Services Supply Chain
- MOSA Contracting Implications
- Portfolio Optimization via KVA + RO
- Private Military Sector
- Software Requirements for OA
- Spiral Development
- Strategy for Defense Acquisition Research
- The Software, Hardware Asset Reuse Enterprise (SHARE) repository

Contract Management

- Commodity Sourcing Strategies
- Contracting Government Procurement Functions
- Contractors in 21st-century Combat Zone
- Joint Contingency Contracting
- Model for Optimizing Contingency Contracting, Planning and Execution
- Navy Contract Writing Guide
- Past Performance in Source Selection
- Strategic Contingency Contracting
- Transforming DoD Contract Closeout
- USAF Energy Savings Performance Contracts
- USAF IT Commodity Council
- USMC Contingency Contracting

Financial Management

- Acquisitions via Leasing: MPS case
- Budget Scoring
- Budgeting for Capabilities-based Planning
- Capital Budgeting for the DoD
- Energy Saving Contracts/DoD Mobile Assets
- Financing DoD Budget via PPPs
- Lessons from Private Sector Capital Budgeting for DoD Acquisition Budgeting Reform
- PPPs and Government Financing
- ROI of Information Warfare Systems
- Special Termination Liability in MDAPs
- Strategic Sourcing
- Transaction Cost Economics (TCE) to Improve Cost Estimates

Human Resources

- Indefinite Reenlistment
- Individual Augmentation
- Learning Management Systems
- Moral Conduct Waivers and First-term Attrition
- Retention
- The Navy’s Selective Reenlistment Bonus (SRB) Management System
- Tuition Assistance

Logistics Management

- Analysis of LAV Depot Maintenance
- Army LOG MOD
- ASDS Product Support Analysis
- Cold-chain Logistics
- Contractors Supporting Military Operations
- Diffusion/Variability on Vendor Performance Evaluation
- Evolutionary Acquisition
- Lean Six Sigma to Reduce Costs and Improve Readiness
- Naval Aviation Maintenance and Process Improvement (2)

- Optimizing CIWS Lifecycle Support (LCS)
- Outsourcing the Pearl Harbor MK-48 Intermediate Maintenance Activity
- Pallet Management System
- PBL (4)
- Privatization-NOSL/NAWCI
- RFID (6)
- Risk Analysis for Performance-based Logistics
- R-TOC AEGIS Microwave Power Tubes
- Sense-and-Respond Logistics Network
- Strategic Sourcing

Program Management

- Building Collaborative Capacity
- Business Process Reengineering (BPR) for LCS Mission Module Acquisition
- Collaborative IT Tools Leveraging Competence
- Contractor vs. Organic Support
- Knowledge, Responsibilities and Decision Rights in MDAPs
- KVA Applied to AEGIS and SSDS
- Managing the Service Supply Chain
- Measuring Uncertainty in Earned Value
- Organizational Modeling and Simulation
- Public-Private Partnership
- Terminating Your Own Program
- Utilizing Collaborative and Three-dimensional Imaging Technology

A complete listing and electronic copies of published research are available on our website: www.acquisitionresearch.org

THIS PAGE INTENTIONALLY LEFT BLANK

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

www.acquisitionresearch.net