Generating a DoD Ontology and Dictionary from the Joint Doctrine documents

Chuck Rehberg & Dave Decker

Semantic Insights

Presented on at the Big Data Working Group

07-Dec-15

Structure of this Presentation

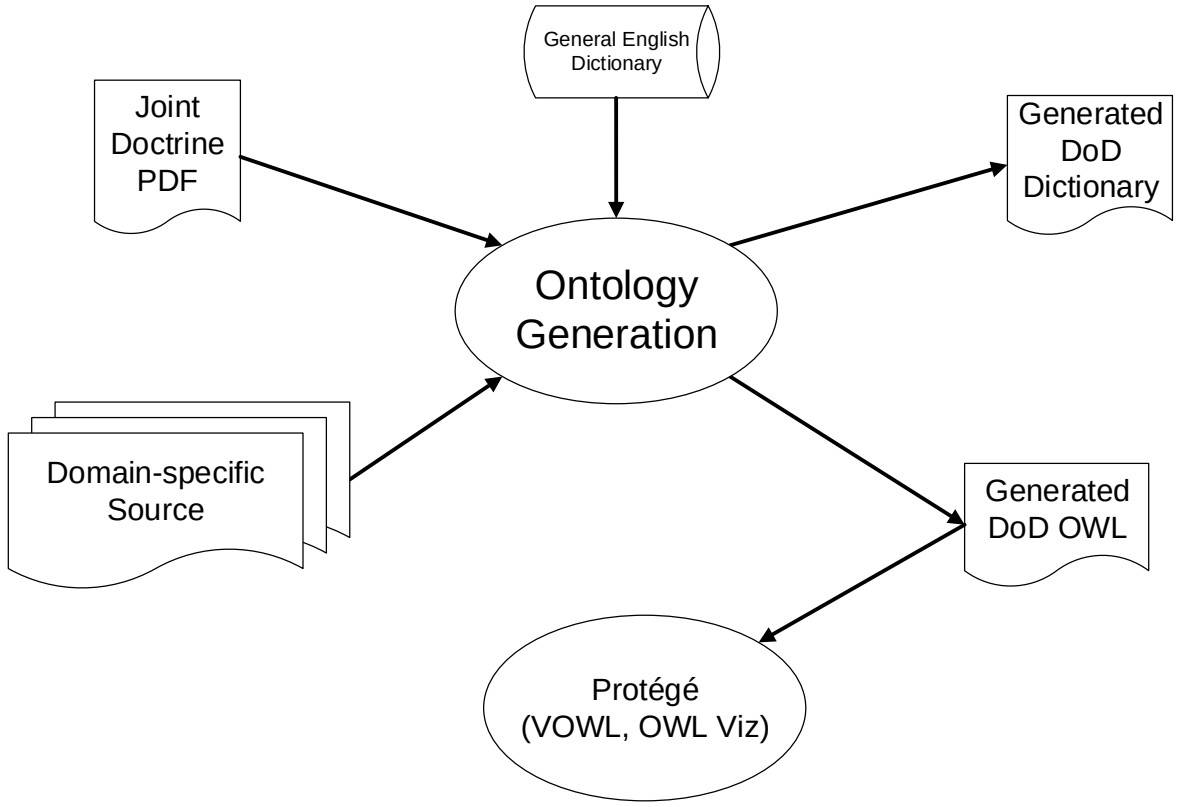
- The Goal of Ontology Generation
- Our Process
- The Technology Used
- Our Results
- Using the Ontology and Dictionary to find Information
- Summary/Discussion
- Who we are

Our Goal

- Generate a substantial portion of an Ontology directly from natural language documents (we chose the Joint Doctrine corpus to demonstrate the approach).
  - http://www.dtic.mil/doctrine/new_pubs/jointpub.htm
  - http://semanticommunity.info/%40api/deki/files/35718/DoD.zip?origin=mt-web
- Use the generated Ontology and NLU software to identify related content in other natural language sources.

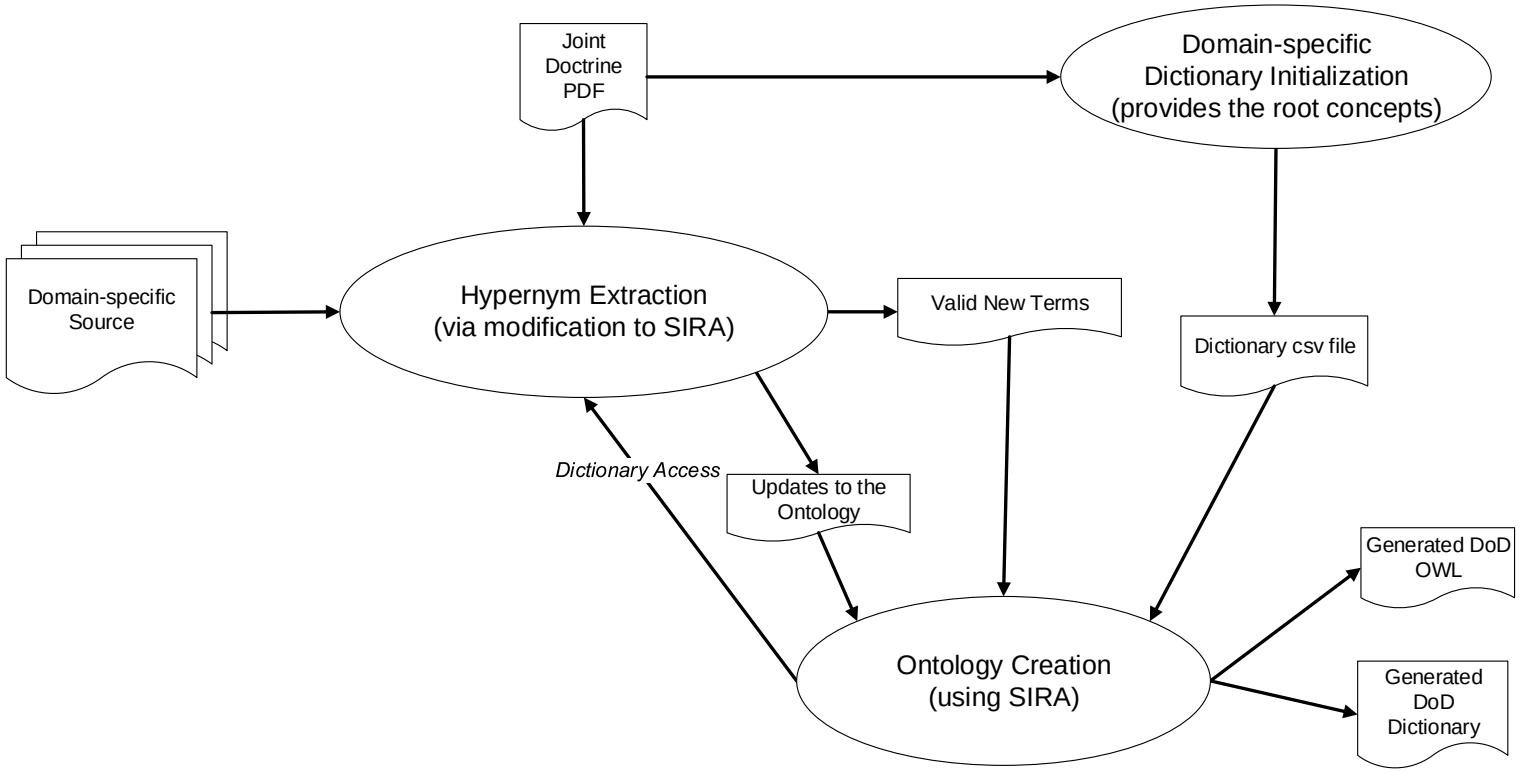
Our Vision for Ontology Generation

- Inputs
  - Given a source file (Joint Doctrine PDF =jp1_02.pdf, Joint Publication 1-02, "Department of Defense Dictionary of Military and Associated Terms")
  - General English Dictionary
- Outputs
  - Dictionary and Ontology
- Validation
  - Other Domain-specific Sources to validate terms (Complete Joint Doctrine Corpus)
- Visualization
  - Load into Protégé and visualize using VOWL and OWL Viz



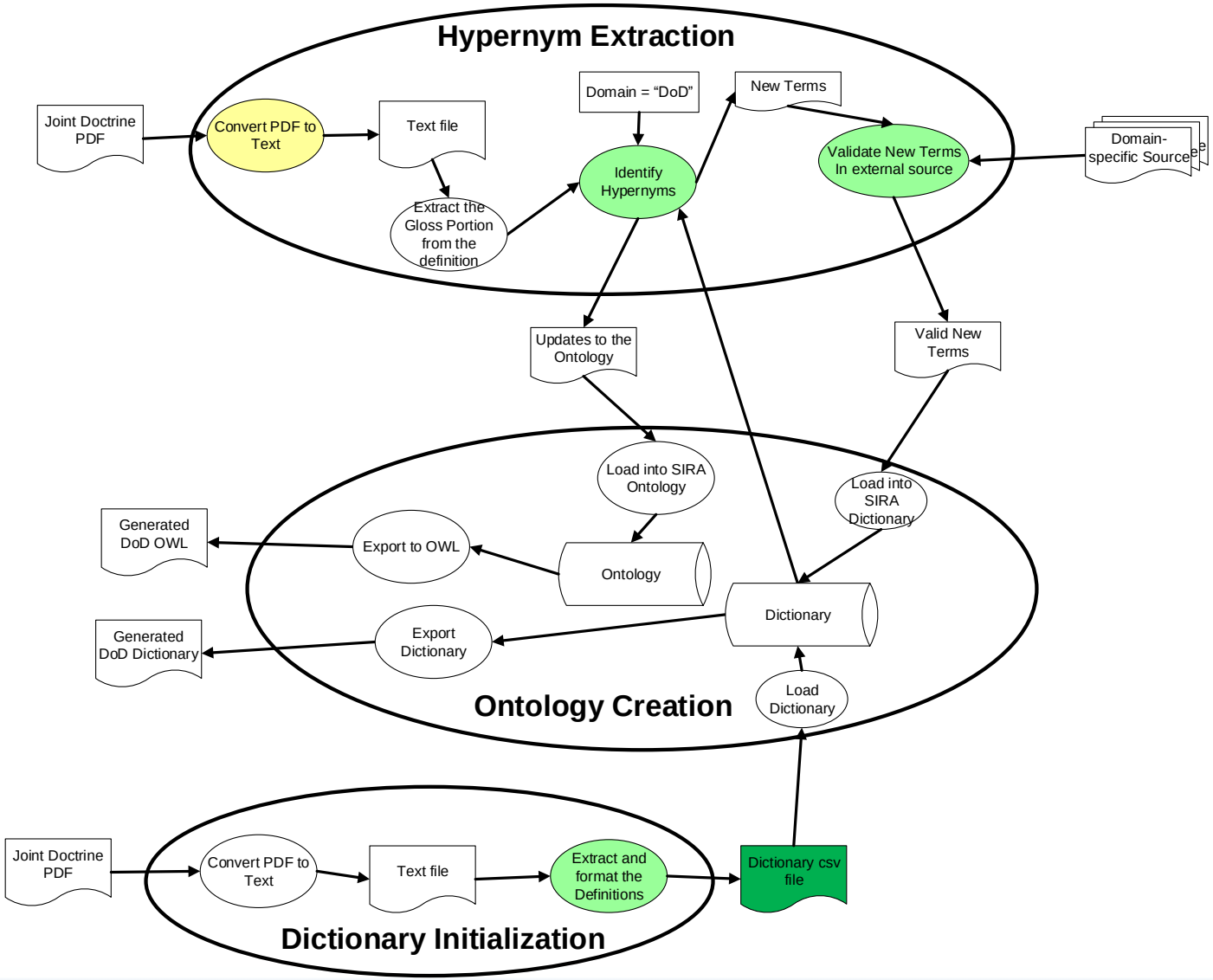
The Process

1. Dictionary Initialization
   - Identifies the domain-specific root concept
2. Hypernym Extraction
   - Finds the modifiers for defined Root terms/concepts and explicit "kind of" statements/enumeration.
   - Ref. "Integrating a Hypernymic Proposition Interpreter into a Semantic Processor for Biomedical Texts" - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1479962/
3. Ontology Creation
   - Provided by SIRA



More Details

1. Dictionary Initialization
   - Some manual cleanup required due to error introduced converting PDF to Text.
2. Hypernym Extraction
   - “Identify Hypernyms” was iterated 4 times to allow for newly defined Root Terms/Concepts to be used.
3. Ontology Creation
   - Ontology created by loading output from Hypernym Extraction.
   - Automatically connects to domain-specific senses in the dictionary.



Semantic Technology Used

- Software Used
  - Semantic Insights NLU system (SIRA)
    - Enhanced to recognize Hypernymic Propositions
    - Enhanced to generate semantic relationships in the Ontology
  - OWL Output displayed using VOWL and OWL Viz plugins to Protégé
- Background Knowledge used
  - Basic English Dictionary (provided by SIRA)
  - Upper level Ontology Organization to be provided by BFO

Our Process and Findings

1. Dictionary Initialization
   - Convert jp1_02.pdf, Joint Publication 1-02, "Department of Defense Dictionary of Military and Associated Terms" to plain text. [done]
   - Convert the plain text document to dictionary import/export format. [done]
   - Load the definitions into the dictionary under the "DoD" domain. [2365 terms]
2. Hypernym Extraction
   - Generate the concept hierarchy by identifying the hypernymic propositions in the definition gloss and add the new terms to the dictionary under the "DoD" domain. Done in four iterations. [>10K concepts]
   - Validation: Search reference corpus to find the newly identified hypernyms. Using keyword search, did not find all the hypernym terms in the Joint Doctrine Corpus. But many were found in on-line Military source documents. – still need to automate.
3. Ontology Creation
   - Load the generated Concept Hierarchy in SIRA [done]
   - Using SIRA, identify the top-most 65 concepts without generalizations that have the most specializations. [done]
4. Visualization
   - Export the generated ontology from SIRA to Protégé, for visualization. [done]
5. Analyze a Target Document for related content
   - Using the SIRA NLU engine, generated DoD Ontology, and a natural language query, *harvest the content from a document of interest* and render the results as a report with bibliography. [done]
6. Next Steps
   - Generate the Ontology for the relationships between the concepts based on definitions in Joint Doctrine, jp1_02.pdf.
   - Generate a graph database (i.e. triplestore) searchable by SPARQL queries.

Are the new Specialized Concepts Valid?

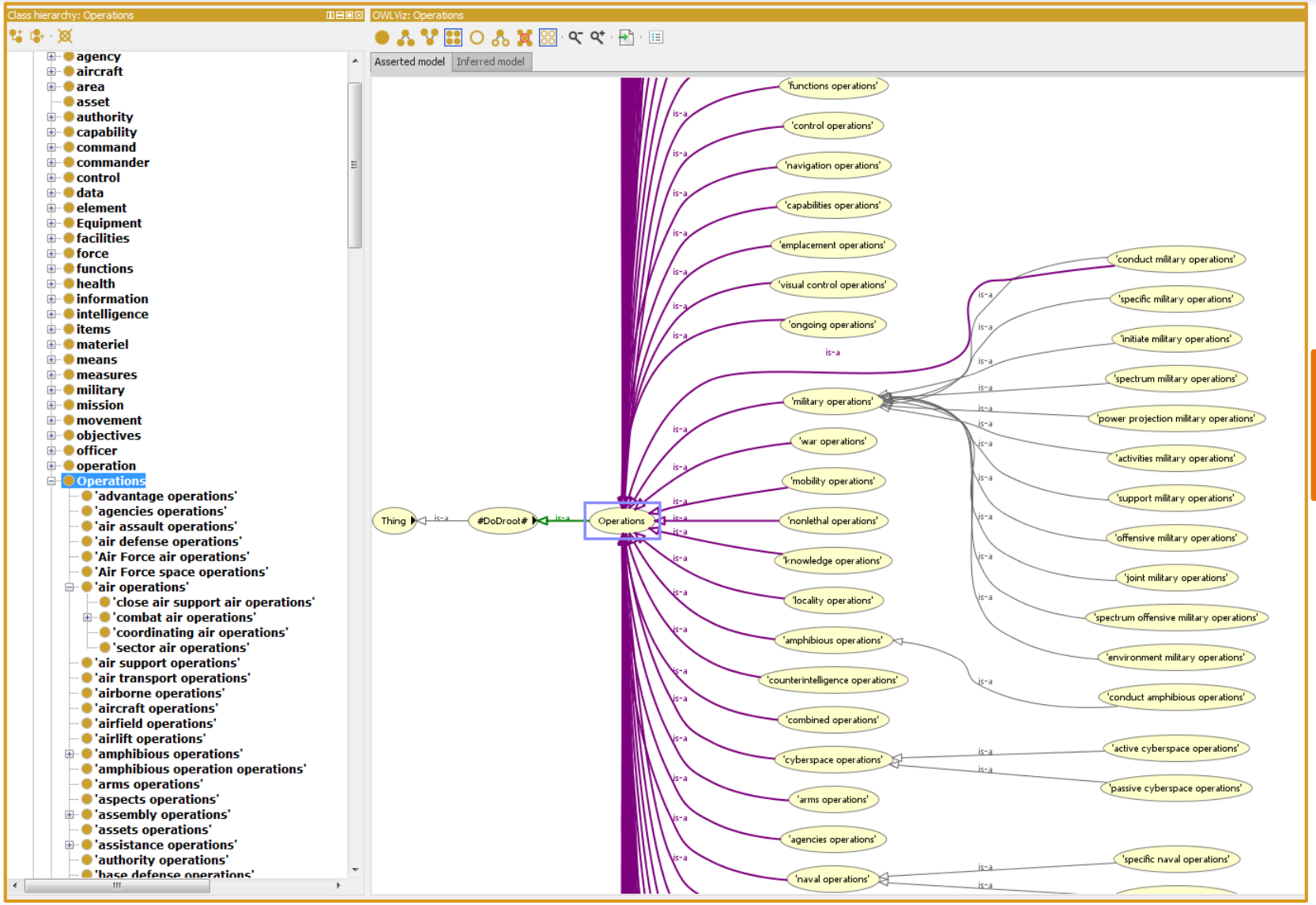
- New term/concept specializations are generated from the Hypernymic Propositions in text.
- Validity Check approaches
  1. Explicit:
     - Is the new term/concept already in use in other domain-specific documents?
     - Process: keyword search domain-specific sources.
  2. Implicit:
     - Is the new term/concept already described or referenced in other domain-specific documents without using the term itself.
     - Process: Look for text mentioning the base concept and describing additional characteristics consistent with the new hypernym.

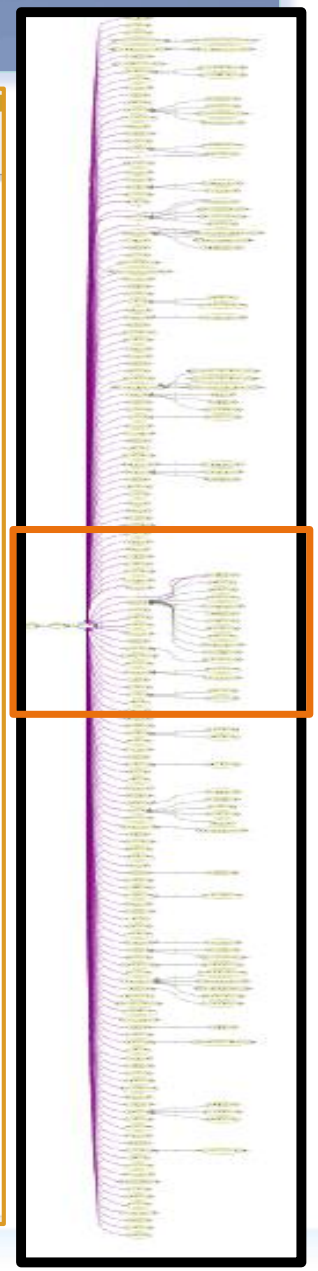
Derived Joint Doctrine Top Concepts

- Listed here are the concepts that have been identified as the top-most concepts (the ~65 concepts without generalizations that have the most specializations).
- In this sense, they are “fundamental” to the domain concept set; this is also validated by manual assessment.
- An abstract root concept #DoDroot# was created to identify them as part of the Joint Doctrine domain.

The “Operations” Concept

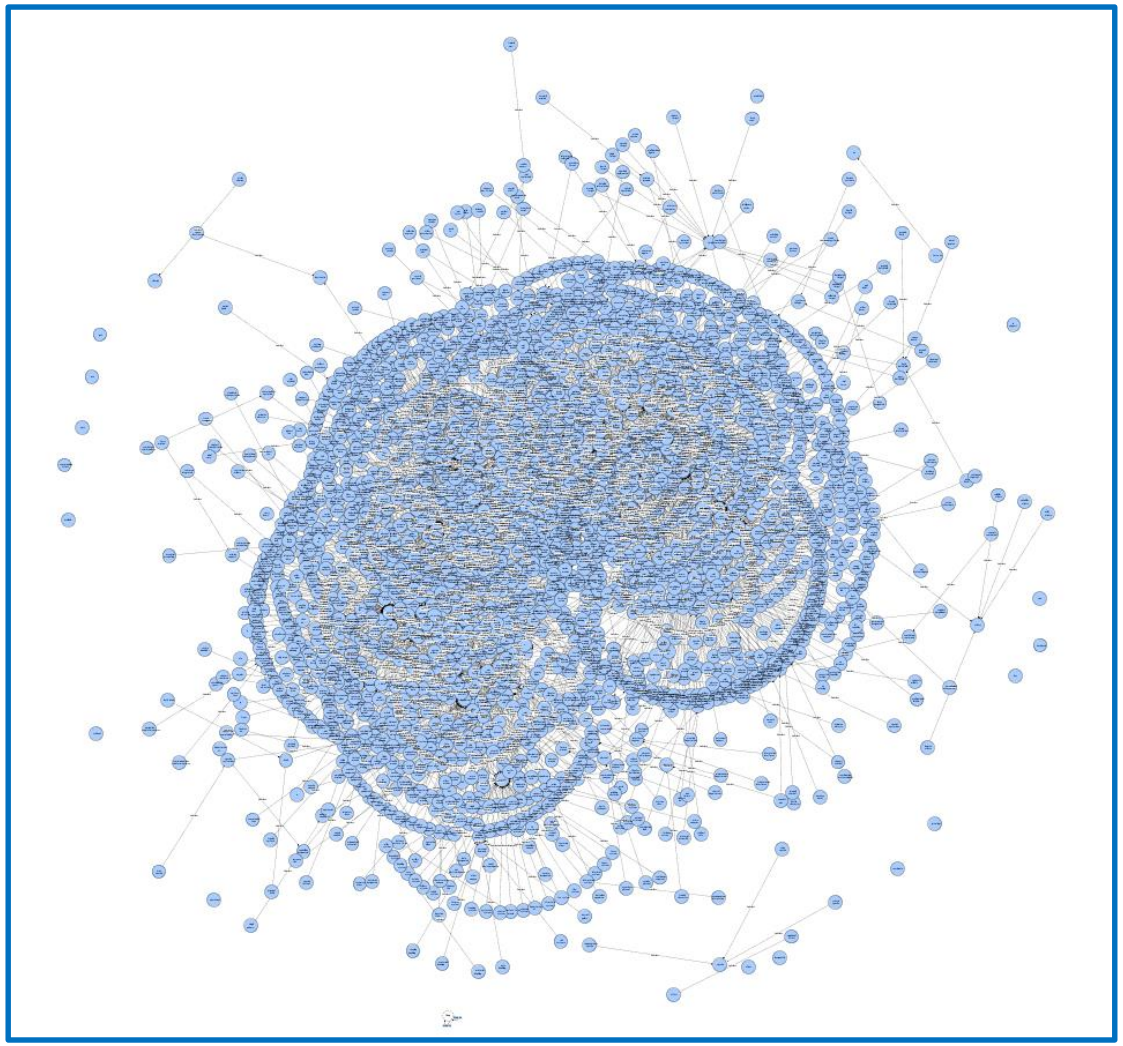
- Protégé and OWL Viz were utilized to produce these two visualizations.
- On the immediate right is a segment of the concept hierarchy for “Operations” with associated generated specializations.
- The diagram on the far right shows this segment in the context of the complete set of associated generated specializations.
- Note the rare instance of multiple inheritance, with “conduct military operations” specializing both “Operations” and “military operations”.





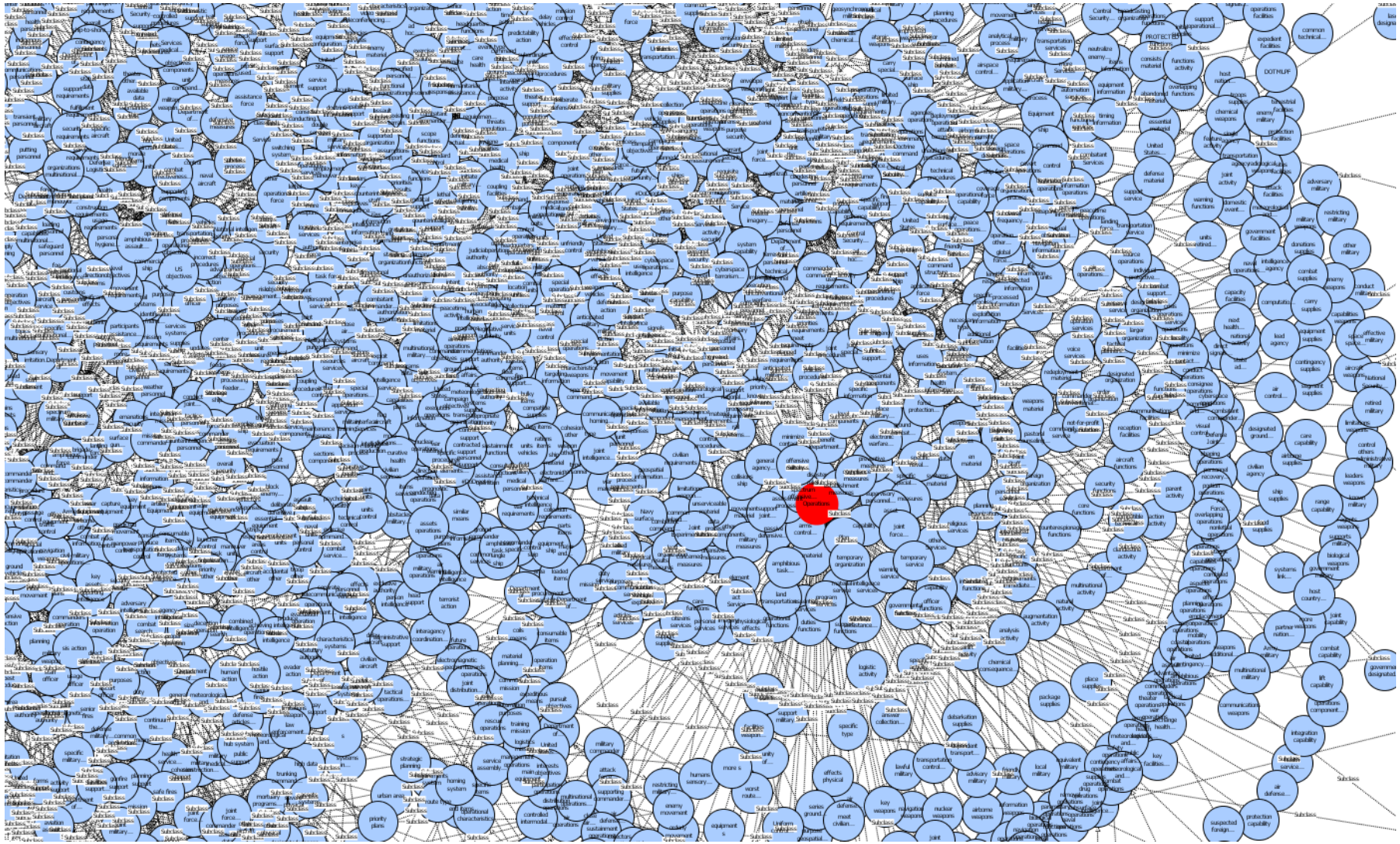
Generated DoD Ontology at a glance

- Protégé and VOWL were utilized to produce the visualization on the right.
- Within this visualization, macro-level “rings” emerged, representing those concepts with the greatest “density” of specializations.
- This is helpful in identifying concepts that are of paramount importance in the domain, as well as relationships among these important concepts.
- “Outliers” are peripheral in the conceptual sense, as well as visually.

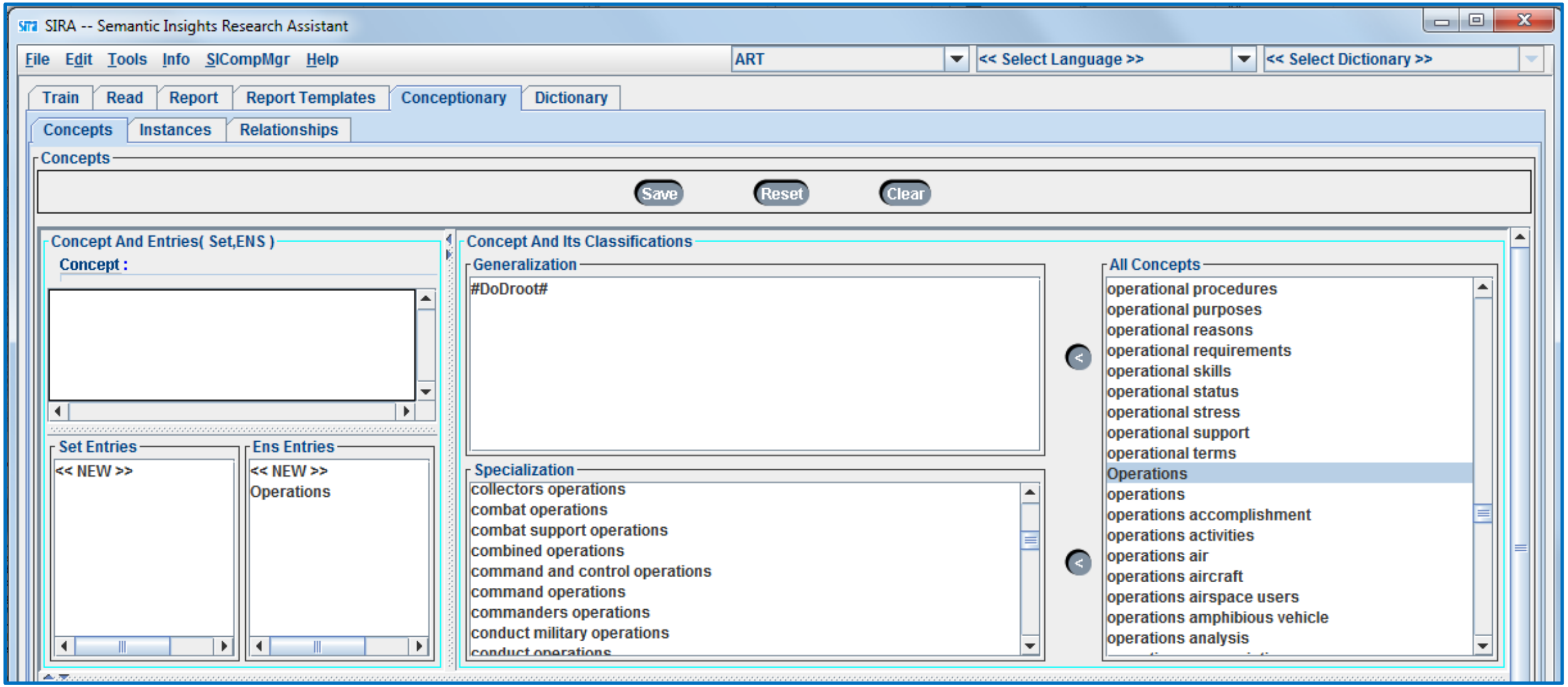


Locating “Operations”

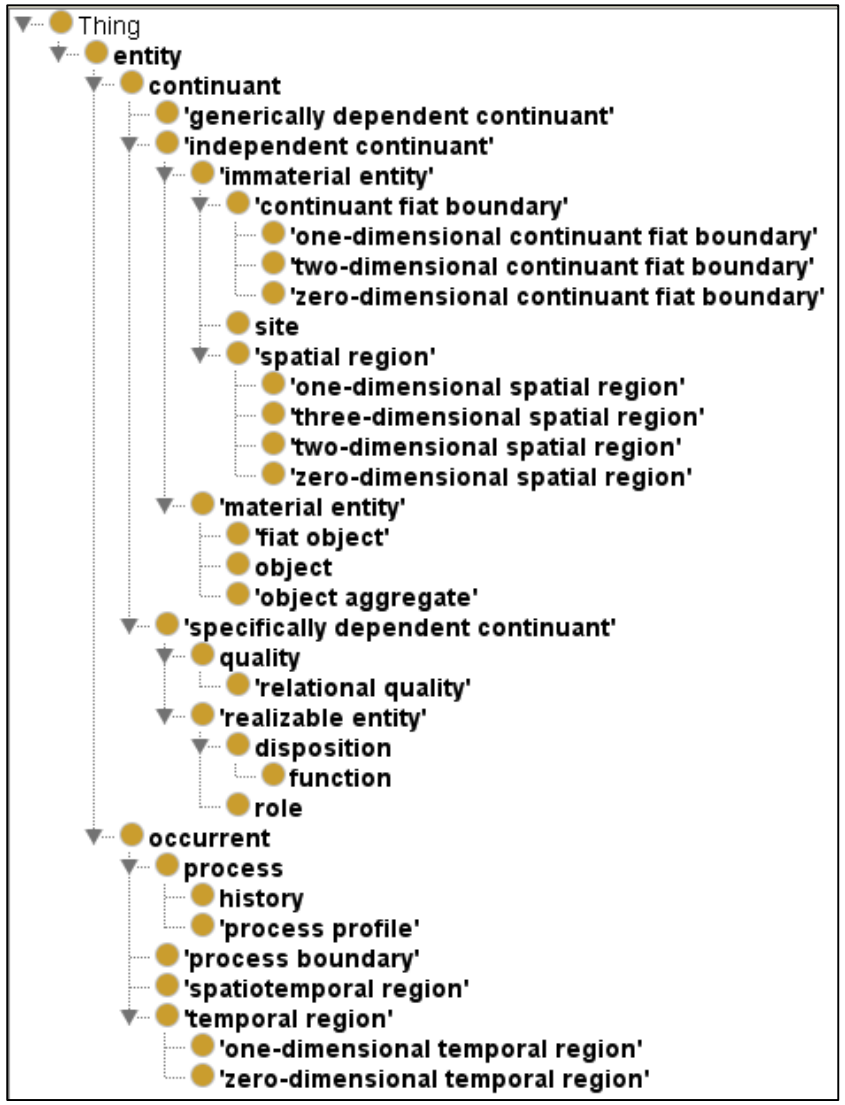
- As an example of a macro-level “ring”; note the concept of “Operations” as the center.
- The concept of “Operations” appears to be of paramount importance in the domain.



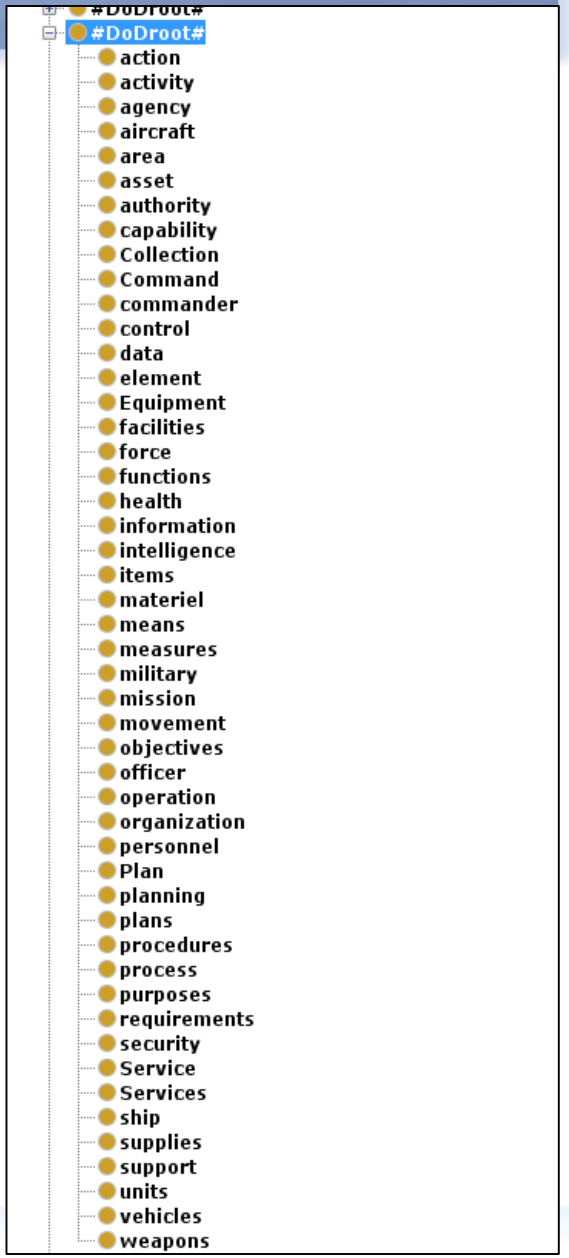
SIRA Ontology View



Mapping the fundamental Joint Doctrine Concepts to BFO

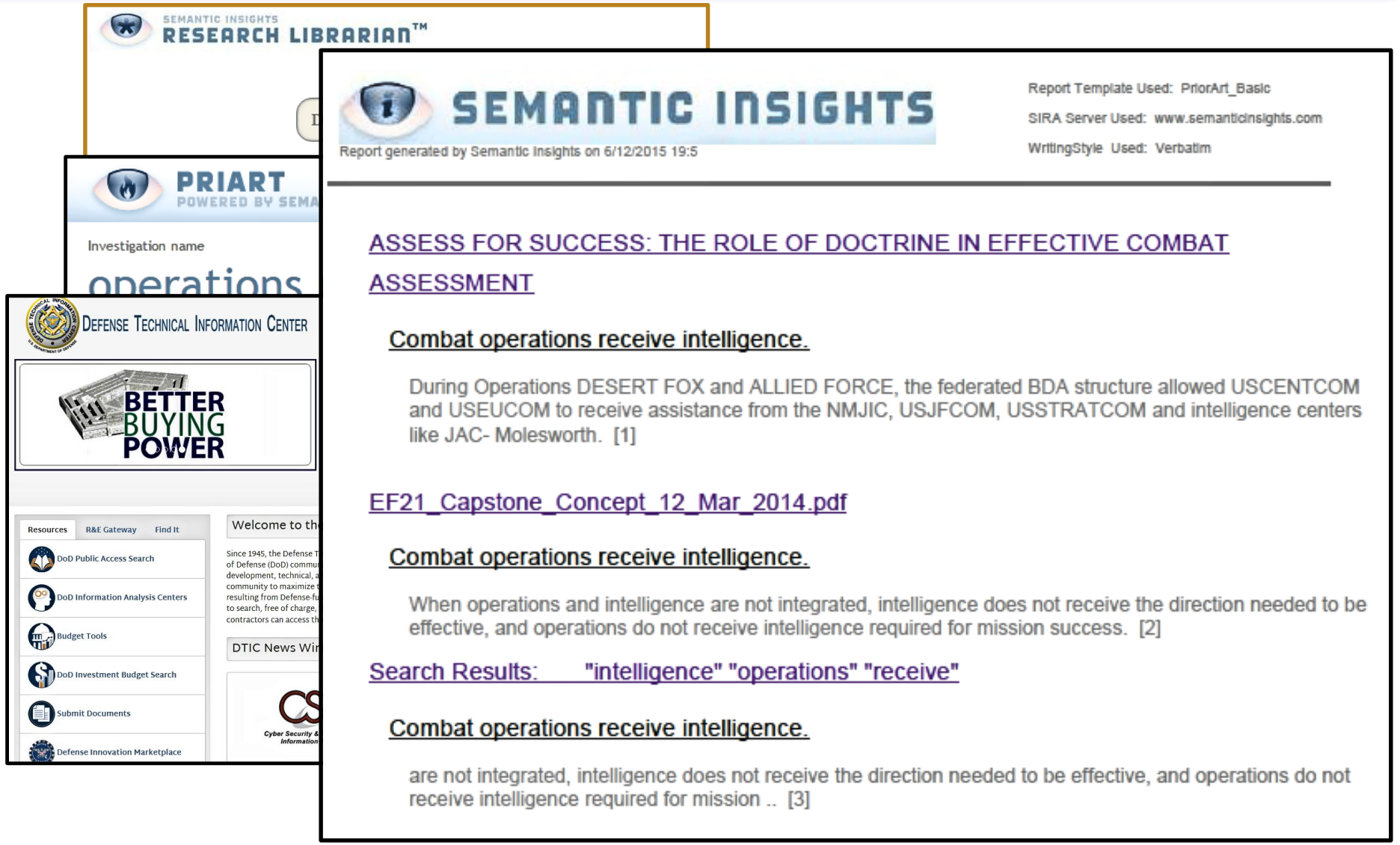


- On the left is the class hierarchy that comprises the Basic Formal Ontology (BFO). This can be utilized as a high-level ontology, to "wrap around" the generated domain ontology.
- On the right is a list of the "root" classes of the generated domain ontology.
- Integrating the two hierarchies entails design activity that results in a domain class being designated as a specialization of a BFO base class.
- For example,
  - *area* isa *two-dimensional-spatial-region*
  - *aircraft* isa *object*
  - *movement* isa *process*
  - *health* isa *quality*



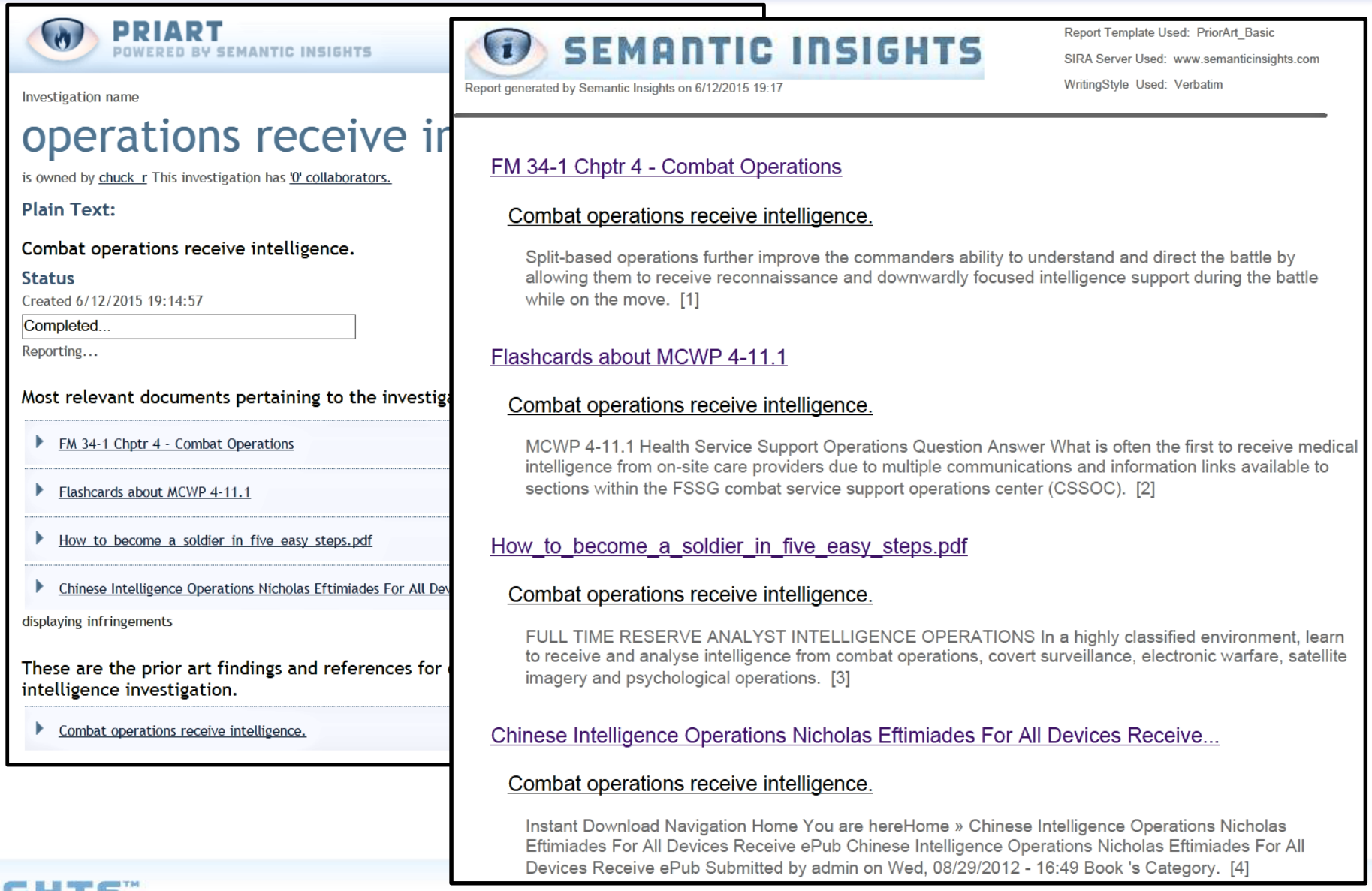
Using the Ontology and Dictionary to find Information

- The Generated DoD Ontology and Dictionary was loaded into SIRA
- Currently testing various queries (e.g. “Combat operations receives intelligence.”)
- Semantic Insights Research Librarian™ and was used to find related content in DTIC (http://www.dtic.mil/dtic/)



Finding documents in the wild…

Auto-generated multiple keyword search queries rapidly gather a large set of documents for automated reading.



PRIART
POWERED BY SEMANTIC INSIGHTS

Investigation name

operations receive ir

is owned by chuck_r This investigation has '0' collaborators.

Plain Text:

Combat operations receive intelligence.

Status

Created 6/12/2015 19:14:57

Completed...

Reporting...

Most relevant documents pertaining to the investig

- FM 34-1 Chptr 4 - Combat Operations
- Flashcards about MCWP 4-11.1
- How_to_become_a_soldier_in_five_easy_steps.pdf
- Chinese Intelligence Operations Nicholas Eftimiades For All Dev

displaying infringements

These are the prior art findings and references for intelligence investigation.

- Combat operations receive intelligence.

SEMANTIC INSIGHTS

Report generated by Semantic Insights on 6/12/2015 19:17

Report Template Used: PriorArt_Basic
SIRA Server Used: www.semanticinsights.com
WritingStyle Used: Verbatim

FM 34-1 Chptr 4 - Combat Operations

Combat operations receive intelligence.

Split-based operations further improve the commanders ability to understand and direct the battle by allowing them to receive reconnaissance and downwardly focused intelligence support during the battle while on the move. [1]

Flashcards about MCWP 4-11.1

Combat operations receive intelligence.

MCWP 4-11.1 Health Service Support Operations Question Answer What is often the first to receive medical intelligence from on-site care providers due to multiple communications and information links available to sections within the FSSG combat service support operations center (CSSOC). [2]

How_to_become_a_soldier_in_five_easy_steps.pdf

Combat operations receive intelligence.

FULL TIME RESERVE ANALYST INTELLIGENCE OPERATIONS In a highly classified environment, learn to receive and analyse intelligence from combat operations, covert surveillance, electronic warfare, satellite imagery and psychological operations. [3]

Chinese Intelligence Operations Nicholas Eftimiades For All Devices Receive...

Combat operations receive intelligence.

Instant Download Navigation Home You are hereHome » Chinese Intelligence Operations Nicholas Eftimiades For All Devices Receive ePub Chinese Intelligence Operations Nicholas Eftimiades For All Devices Receive ePub Submitted by admin on Wed, 08/29/2012 - 16:49 Book 's Category. [4]

Summary for Discussion

1. An increasing level of automated generation of initial Ontology appears feasible.
2. We experienced a 3:1 increase in defined terms over the dictionary definitions explicitly given in jp1_02.pdf, Joint Publication 1-02, "Department of Defense Dictionary of Military and Associated Terms".
3. We stopped the generation at an Ontology specialization depth of 4.
4. As the size of the Ontology increases we encountered limitations in the Owl Visualizers we tested. Changes were also required to SIRA to handle the increase in combinatorics
5. The generated Ontology and Dictionary provided domain knowledge used by SIRA to improve reading accuracy.
   - This was evidenced by an increase in recognizing multi-word terms in generated SIRA Reader and increase in finding generalization and specialization terms.
   - More work is required to quantify this improvement.
6. Generating the Ontology relationships also appears feasible.
   - We have working code to identify the relationship patterns in sentences involving terms from the generated "DoD" ontology. Some of these relationships represent Verbs, some represent Prepositions, some represent gerund clauses, and some are adjectival (i.e., indicate some characteristic of a concept). Work to add these to Ontology is underway.

Who we are

- Semantic Insights is the R&D division of Trigent Software, Inc. www.trigent.com
- We focus on developing semantics-based information products that produce high-value results serving the needs of general users requiring little or no training.
- Visit us at www.semanticinsights.com

Chuck Rehberg

As CTO at Trigent Software and Chief Scientist at Semantic Insights, Chuck Rehberg has developed patented high performance rules engine technology and advanced natural language understanding technologies that empower a new generation of semantic research solutions.

Chuck has more than thirty years in the high-tech industry, developing leading-edge solutions in the areas of Artificial Intelligence, Semantic Technologies, analysis and large -scale configuration software.

For more information, please visit us at www.semanticinsights.com or send email to chuck_r@semanticinsights.com.

Dave Decker

Dave Decker is a Research Fellow at Semantic Insights, developing advanced semantic and natural language understanding technologies.

Dave has more than thirty years experience at the intersection of artificial intelligence and product development. In a variety of industries, he has led innovative R&D with machine learning algorithms including neural networks, often in the context of data science and analytics.

For more information, please visit us at www.semanticinsights.com or send email to dave_d@semanticinsights.com.

SEMANTIC INSIGHTS™

A Division of Trigent Software, Inc.