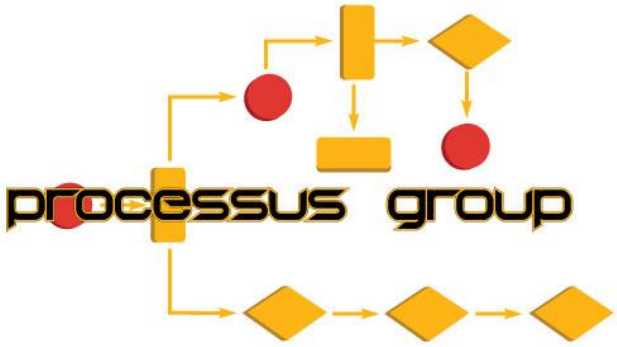
Solving the “Garbage in-Garbage out” Data Issue Through Ontological Knowledge Graphs

Matthew Maher, Rob Orlando
Processus Group and Kord Technologies
310-351-1936
Corona, Ca 92880
maherm@Processusgroup.org
orlandor@processusgroup.org



Abstract

Ontology-Based Knowledge Graphs have emerged as a subfield for Artificial Intelligence, Machine Learning or Artificial Neural Networks by generating knowledge resources. We describe how, ontologies - which provide formal and explicit specifications of conceptualizations for an operational process or domain - play a crucial role in the representation of data in the form of knowledge. Generally, an ontology is specified for a particular domain but we have expanded this concept through our Process Driven Ontologies (PDOs) to focus on the data, information, knowledge, intelligence and understanding (DIKIU) that is shared across domains in the execution of an operational process (i.e. Military Decision Making Process, Targeting process, Intelligence and Cyber Preparation of the Battlefield, etc.). Since intelligent systems' ability to "learn" is reliant on the data extraction from a large diverse data sets, it is imperative that the data is structured in a quality way to provide the patterns for a machine to learn from. If a particular domain has bad data, the intelligent systems will learn bad patterns and create a model that is unfixable. Thus, causing the Garbage in Garbage out (GIGO) dilemma that DoD is facing. By using an ontology to formally and explicitly specify the concepts of a domain adding the semantics to describe how its data relates to each other is a solution for dealing with bad data. Especially if the data the intelligent system relies on is flawed, lacking in semantic and all together unorganized, the use of an ontology model brings meaning to the data, through the creation of knowledge graphs.

The largest problem with the adoption of Artificial Intelligence, Machine Learning or Artificial Neural Networks (AI/ML/ANN) in DoD is in the state of data. The adage of "garbage in garbage out" (GIGO) is slowing the progress of DoD to adopt these advances capabilities. To ease the transition to AI/ML/ANN the intelligent system must have a viable training set of data to facilitate the learning process. For DoD this means accessing very large diverse data sets contained in distributed databases that were created in support of specific domains and do not capture the sharing of knowledge. Not to mention the need to access large segment of the data that takes the form of natural language text or unstructured data shared between staff sections to execute operational processes. To help "clean up" the data issues, we will describe how developing an ontological model brings meaning to the data, through the creation of knowledge graphs. The ontological knowledge graphs are built around the structural complexity of diverse databases by defining the semantic relationships between data stored in databases and information captured through knowledge sharing. Since ontologies represent knowledge or meaning they provide a blueprint for how a semantic-based approach for data modelling and information retrieval will improve the interface between data and machine learning to bring the datasets closer to the users' requirements.

The five core benefits to creating ontological knowledge graphs for "garbage data" is:

1. Learn the structural mapping of the data
2. Define a semantic model of the data
3. Represent domain metadata with related semantics extracted from the relational database schema.
4. Blueprint for fixing the "bad data"
5. Create an authoritative data set for training data associated domain knowledge that intelligent systems use to start the learning process

To fix the "bad data" is very complex and very expensive. To truly address the data contained in large data sets you must first add structure, ensure the data is quarriable, analyze the data, provide visualization, harmonizing/normalizing and correlating the data. The use of traditional data management tools and processing applications is not an efficient way to address the problem since they

will not provide the semantics and interrelationships mapping based on domain knowledge that the data needs to organize the large and complex data sets. Because of this we have developed the process for using ontologies to bridge the gap by understanding the similarities and differences of data represented in a database verse an ontological model. Ontological knowledge graphs and relational data models have differences and similarities, ontologies are better at defining semantically rich data that is quarriable, than database schemas. An ontology is better at specifying domain knowledge of a domain of interest. Ontologies are also based on logic, specifically formal language description logic, which provides a readable constructing defining interrelationships across category definitions for deciding subset and superset relationships of data. We have been investigating efficient ways to process these large data sets to get benefits such as: search performance, creation of reference data, creation of knowledge graphs for AI/ML/ANN and enable reasoning. Our solution is to build ontological knowledge graphs for large data sets for two purposes, to clean up the data set and to develop the knowledge representation of operational data for how intelligent systems to learn from knowledge.

Our process is to

(1) define a semantic model of data,
(2) specify domain knowledge based on doctrine and experience
(3) define links between different types of semantic knowledge.

To simplify this issue, think of the GIGO data problem like a hoarding situation. A house that is considered a hoarding situation is just a giant unorganized container of stuff. The stuff in the house has little or no organization. In the overwhelming piles it is near impossible to discern the difference between valuables and rubbish. Much like large data stores. The data continues to "pile up" without the appropriate metadata or relationships identified in the data. To clean up the hoarding, you have two options destroy everything and start over, or to systematical remove the garbage and organize it into the appropriate classification. Choosing the second option, you essential clean the hoard by sorting and describing the stuff as recyclables, valuables, keep sakes and rubbish. So instead of destroying everything and starting over you can salvage the stuff (data) you need, label it (Metadata) and determine its use (add the semantics). The same options are present with "garbage data". The organization can spend significant resources in starting over or developing an ontology that creates a layer above the data base that describes the data by defining links between ontology concepts and relational data. Hence, how ontologies can be used to discover and describe information from large data sets across DoD and provide operational meaning to the data. Using ontological knowledge graphs for the specification of datas' meta data as a foundation to efficiently discover useful information for analysis and organization of the data sets. The ontology development will not only identify the structural complexity of complex databases but also the semantic relationships between data stored in databases.

This process was successful in the analysis of data that was required to plan and configure a tactical radio and waveform. When we started the process the radio/waveform required over 6000 data elements configured to make it operational. After our initial findings our model provided analysis that only 1154 of the data elements were required to support the operation of the radio/waveform. This reduction allowed the developer to templatize or automate over 4800 parameters. Once we reduce the data down to 1154, we started to define the interrelationships of the data. For example, we showed how one of the power setting had a direct relationship to 12 other data settings. This allowed us to build the rules in the command and control systems that if this setting is changed these other 12 data elements were also automated. Once we finished defining the interrelationships the number of parameters that actually had to be configured was reduced to less than 200. Reducing the cognitive

burden on both the soldier and the system.

In this example we broke down the existing configuration data that is stored in its database and aligned it to ontology properties. The constructs of a database model such as tables, columns, datatypes and constraints were mapped to the ontology as classes, instances, properties and constraints. Thus, creating a radio waveform domain ontology that represents metadata with related semantics extracted from the relational database schema of the waveform.

By representing data as knowledge in a semantics-driven ontology we make a knowledge graph that is built on semantics. Semantics that are the basis for creating new inferences from the data which would otherwise go unseen in a database. Our ontologies generate new knowledge while the database is lying dormant, waiting to be queried by a system that doesn't understand the data. Our ontology knowledge graph aren't like any other database; it provides new insights, which can be used to infer new relationships, semantics and metadata for large datasets and is based on operational processes. We define the data, who/what creates, consumes, analyzes, modifies and combines the data into information. Ontology Knowledge graph is a tool for managing the semantics, metadata association and interrelationships of data organized with the operational work flows through every organization and provide the domain data analytics to clean up data ensuring it can be used for leverages Artificial Intelligence, Machine Learning or Artificial Neural Networks (AI/ML/ANN).