Integration of Intelligence Data through Semantic Enhancement

Salmen David
dsalmen@data-tactics.com
Data Tactics Corp.

Malyuta Tatiana
tmalyuta@data-tactics.com
Data Tactics Corp.,
City University of New York

Hansen Alan
alan.hansen1@us.army.mil
Intelligence and Information Warfare Directorate

Cronen Shaun
shaun.cronen@us.army.mil
Intelligence and Information Warfare Directorate

Smith Barry
phismith@buffalo.edu
National Center for Ontological Research,
University at Buffalo

***Abstract*—We describe a strategy for integration of data that is based on the idea of semantic enhancement. The strategy promises a number of benefits: it can be applied incrementally; it creates minimal barriers to the incorporation of new data into the semantically enhanced system; it preserves the existing data (including any existing data-semantics) in their original form (thus all provenance information is retained, and no heavy pre-processing is required); and it embraces the full spectrum of data sources, types, models, and modalities (including text, images, audio, and signals). The result of applying this strategy to a given body of data is an evolving Dataspace that allows the application of a variety of integration and analytic processes to diverse data contents. We conceive semantic enhancement (SE) as a light-weight and flexible process that leverages the richness of the structured contents of the Dataspace without adding storage and processing burdens to what, in the intelligence domain, will be an already storage- and processing-heavy starting point. SE works not by changing the data to which it is applied, but rather by adding an extra semantic layer to this data. We sketch how the semantic enhancement approach can be applied consistently and in cumulative fashion to new data and data-models that enter the Dataspace.**

Keywords: integration, intelligence data, ontology, semantic technology.

I. Introduction

The success of the war fighter and homeland defender in the Net-Centric Warfare environment is largely defined by the ability to quickly acquire and efficiently and accurately process intelligence information from numerous heterogeneous sources of different structure and modality. Traditional data integration approaches fail in the face of the scale, diversity, and heterogeneity of intelligence data sources and data-models because they fail to address one or more of the following requirements:

- Integration must proceed without heavy pre-processing
- Integration must proceed regardless of the data-models used (or not used) in the data sources to be integrated,
- Integration must proceed regardless of the data modality, and without loss or distortion of data, of its associated data semantics, and of data-provenance information,
- Integration must involve the ability to incorporate multiple points of view on the data to be integrated, including different views of the data, for example on the part of different analysts using different analytical tools.

As a first step towards meeting these requirements we introduced in 2009 the Data Representation and Integration Framework (DRIF) [1, 2], which presents minimal barriers to the incorporation of new data into a data resource, thus requiring no heavy pre-processing and no data or data-model conditioning. DRIF embraces the full spectrum of data sources, types, models, and modalities, including text, images, audio, and signals, while supporting a variety of integration and analytic processes and tools. Details are presented below.

The Dataspace store of intelligence data which is the subject of this communication is the result of applying the DRIF to the task of integrating very large heterogeneous primary data artifacts. As the Dataspace has evolved through time, so it has incorporated progressively ever larger quantities of data, and also more specific local implementations and data structures used by data analysts, some of which bring their own data semantics. For the purposes that the Dataspace is intended to serve, it is vital that no restrictions are imposed either on the types of source-artifacts and the associated models and media within the Dataspace, or on the processes by which the Dataspace is populated (whether by loading structured data from a database, by extraction from a text document through some Natural Language Processing application, by automatic analysis of signals, or through inference by a human analyst).

The design of the Dataspace is such that it can incorporate hundreds of millions of unstructured documents and similarly large quantities of images, signals data, and other structured and unstructured primary data artifacts. Each of these artifacts, when it enters the Dataspace, is represented through a set of metadata, including labels specifying image type, MIME type, and so forth, as well as provenance information. Further processing may, for example, associate pixels in an image with the name of a person, or a range of characters in an unstructured text document with the name of a location, or extract a cell from a database table. The DRIF provides a common framework in which the results of all of these processes are represented in a unified way, details of which are provided below. As a result, primary data can be utilized immediately upon entering the Dataspace for a variety of different kinds of search and more sophisticated processing based thereon. DRIF is not, however, a magic bullet; many issues of data integration at the syntactic level will remain, arising for example as a result of data formats which do not match, where we will need to normalize the format into an augmented model that will serve as the target of annotations. This will involve considerable effort to ensure that the needed actions are performed promptly and consistently whenever new data comes in. Here, however, we focus exclusively on those issues which arise at the stage of what we can loosely call the 'representational' aspects of data integration.

Some primary artifacts within the Dataspace already incorporate useable semantic content – for instance a structured database which incorporates meaningful column headers, or a message with a structured payload incorporating meaningful tags. But such content is *ad hoc*. It is tied to specific local implementations and typically falls short of what is needed to secure semantic interoperability of the implementations involved because of the absence of a common formally coherent approach to semantics and of a common governance process.

Moreover, full semantic integration is in any case prevented by the needs of openness of the Dataspace to ever new sorts of primary data and analytically derived data. It is to compensate for this problem that we have developed our strategy for semantic enhancement. We start out from the assumption that semantic data enrichment can be achieved only incrementally, through the step-by-step creation of ontology modules that are designed in coordinated fashion to work well both with each other and with specific bodies of Dataspace content. The vision is a lightweight, flexible approach comprising an *extra ontology layer* that leverages the contents of the Dataspace without adding storage and processing weight to what is an already storage- and processing-heavy resource. We discuss the details of semantic enhancement in section IV. First, however, we introduce the DRIF and the Dataspace to which the SE strategy will be applied.

II. Data Representation and Integration Framework

Our starting point is a body of U.S. Department of Defense (DoD) intelligence data within what we are here calling the Dataspace. The implementation in the specific context upon which we focus here is engineered around cloud computing paradigms and is primarily based upon open-source cloud software stack components. This cloud computing foundation leverages advantages of linear scaling and parallel distributed computation when faced with the reality of ever increasing data volumes and integration processing. All the work described is either deployed or in the final stages of testing prior to deployment.

The Dataspace is built using the Data Representation and Integration Framework (DRIF), which has been designed to represent large quantities of data in a form that is useful to the end user both for direct inspection and for the application of various kinds of analytics. Representations of source data artifacts and their contents within the Dataspace are of two forms, which we call *primary* and *derived*, respectively. The Dataspace is divided into corresponding segments (see Figure 1) in a way that supports a comprehensive approach to integration that allows accommodation of the multiple views of the primary and derived data and of the associated data-semantics and metadata which arise for example as a result of the workings of multiple different sorts of analytical tools.

A. Approach to Integration

Our approach to integrating intelligence data starts with source artifacts consisting of primary data across a variety of representation modalities. This primary data is weakly integrated in the sense that indexes are provided to support simple (string-based) data search across all primary artifacts.

Some primary data comes with its own native structure, and further structure will typically be added thorough analytical processing. The second integration step addresses the need for the unified storage of this structured data to support more complex structured search across both primary and derived artifacts.

Importantly, we here embrace the diversity of domain-specific data-models employed throughout the Intelligence Community while at the same time reaping benefits from an approach that is data-model agnostic. This is because the unified representation provided by the DRIF allows analytic processing of data in highly diverse primary artifacts associated with different native data-models to be used as targets of cross-artifact analytics. For example, and most simply, it is possible to perform unrestricted string search across structured artifacts of highly different sorts. Examples of more sophisticated analytics include computer-aided data-model harmonization, for example by allowing significant overlap of sets of values of attributes from different databases to be flagged by the analytic process as a potential indication that the attributes have the same meaning, thereby making it possible for the relevant portions of the two databases to be enriched through fusion.

B. Dataspace Organization

The organization of the Dataspace is schematically illustrated in Fig. 1.

Segment 0 is a store of primary artifacts, including documents, images, signals, and analysts' work products vetted for re-use as input for further processing. The physical implementation of Segment 0 may be such that all data is stored internally; or it may be distributed, so that source artifact data may for example be either contained in the cloud store or stored externally to the Dataspace and referenced in the cloud store. Primary data vary widely by nature; they may have different structures (for example of a relational database), or they may be unstructured (for example, free text, audio or video files), and they may be of different modalities (for example they may be cells of a relational database, audio sequences, assertions of an analyst).

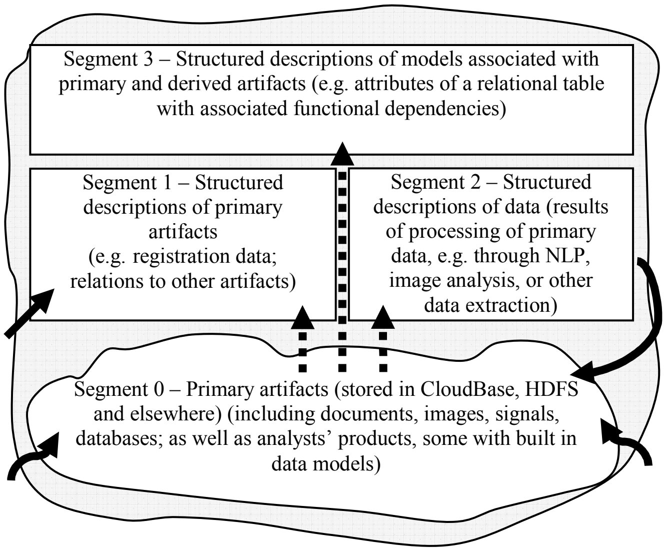

Segment 3 – Structured descriptions of models associated with primary and derived artifacts (e.g. attributes of a relational table with associated functional dependencies)
Segment 1 – Structured descriptions of primary artifacts (e.g. registration data; relations to other artifacts)
Segment 2 – Structured descriptions of data (results of processing of primary data, e.g. through NLP, image analysis, or other data extraction)
Segment 0 – Primary artifacts (stored in CloudBase, HDFS and elsewhere) (including documents, images, signals, databases; as well as analysts' products, some with built in data models)

Figure 1. Organization of the Dataspace. Solid line: registration processes; curved solid lines: processes that ingest artifacts into the Dataspace, including feeding back into the Dataspace analysts' products – results of the Dataspace processing ; dashed lines: derivation processes.

Segment 1 includes primary artifact registration data as well as specifications of relations between artifacts (for example, nesting of an image within a document, or attachment of one document to another). Segment 1 will include also data pertaining to the way each derived artifact of Segments 2 and 3 is derived from primary artifact(s) in Segment 0.

Segment 2 stores the structured data that is either already present in primary artifacts or derived therefrom through analytic processing resting on data-models represented in Segment 3.

Segment 3 stores the descriptions of the data-models used in Segment 2. These data-models may include database schemas, message formats, or XML schemas. The data-models themselves are primary artifacts and are thus stored in Segment 0 and registered in Segment 1.

The Dataspace is evolving continuously not only because of new primary data ingested from the outside, but also because new artifacts are being created, for example, through analysts' reports based on processing of existing data. These artifacts themselves have a status of new primary artifacts.

C. Segments as Abstractions Over the Artifacts

Each of Segments 1-3 is an abstraction over the corpus of primary data artifacts (Segment 0) and supports analytics of a particular type:

- Segment 1 is a high-level view of the entire artifact corpus including the relations between the artifacts, but with no reference to their internal contents.
- Segment 2 is a collection of detailed views of the internal contents of the artifacts at the level of individual data items.
- Segment 3 describes the data-models which support the two sets of views just mentioned as well as synoptic views (ultimately including SE-based views) of the type which can foster harmonization.

D. Where Models and Primary Data Come Together

We believe that the principal contribution of the Dataspace endeavor is to resolve certain problems of storage and thus of representation, enrichment, and evolution of large bodies of data. The goal is to provide room for both primary data and the multiple results of processing these data by different analysts or analytic methods. To achieve this we introduced in [3] a strategy for description of data that is designed to enable true data integration across a constantly evolving and highly heterogeneous resource comprehending extremely large volumes of data. As already recognized at the very beginning of contemporary high-level research in biomedical ontology [4], this end can be achieved only if data are exposed in a way that is independent of their original intended use. This must involve some means to represent original data-models at a level of abstraction that is higher than that of primary data. We accordingly propose an abstract data-model based on five core elements: sign, concept, term, predicate, and statement, which we believe is sufficient to represent any data-model in these terms.

Sign: A *sign* g_i is a string that is the abstracted proxy within the dataspace for one or more chunks of data used in some primary artifact with the intention of referring to some individual entity (e.g. person, location, organization, object, event). Examples include: a sign of the type proper name that is associated with an expression (for example 'he' or 'Dr. Watkins' occurring in a document; a label annotating an area in a pixel array as forming an image of some building; a label annotating a fragment of an audio stream or other signal as recording some explosion event. Each sign is associated with one or more physical extents within those primary artifacts with which it is associated, which we call *mentions* (the latter are what are elsewhere called *tokens*). The collection $G = \{g_i\}$ comprehends all signs extracted from primary data artifacts and changes with the incorporation of new artifacts.

Concept: A *concept* c_i is (for the purposes of this exposition) a string that is used in the Dataspace to represent some general category or grouping. The purpose of concept strings is to represent and allow reuse of classifications native to primary artifacts. Concepts are taken from data-models registered in Segment 1. Examples of concepts are: the classes of an ontology such as UCore SL, the tag set in an XML Schema Document (XSD), and the attribute or table names in a relational database. The collection $C = \{c_i\}$ comprehends all concepts within the Dataspace and changes as new data-models are incorporated.

Term: A *term,* t_{ij}, is an ordered pair of strings $<g_i,c_j>$, where $g_i \in G$ and $c_j \in C$. Each term results from a process of contextual disambiguation of a *sign*, a process which associates a *sign* with a *concept*, as in <123-45-6789, SSN>. The collection $T =$

$\{t_{ij}\}$ comprehends all terms identified by analytic processing of primary artifacts.

Predicate: A *predicate* (by which we mean here always: *binary relational predicate*) p_i is a string that is used to connect terms in accordance with domain and range constraints. Predicates are used in the formation of *statements* (as described below). Examples of predicates are: hasSSN, hasLocation, hasBirthDate. Predicates are derived from data-models registered in Segment 1, for example from table column headings or from XML tags. The collection $P = \{p_i\}$ comprehends all predicates within the Dataspace and changes as new data-models are added.

Statement: A *statement* s_i is an ordered triple consisting of a subject, a predicate, and an object. The collection $S = \{s_i\}$ of statements is recursively defined. At the lowest level, statements are ordered triples consisting of a term, a predicate, and a second term. In higher-level statements, subjects and objects may be lower-level statements. Examples: <[Bruno, PersonName] hasSSN [123-45-6789, SSN]>

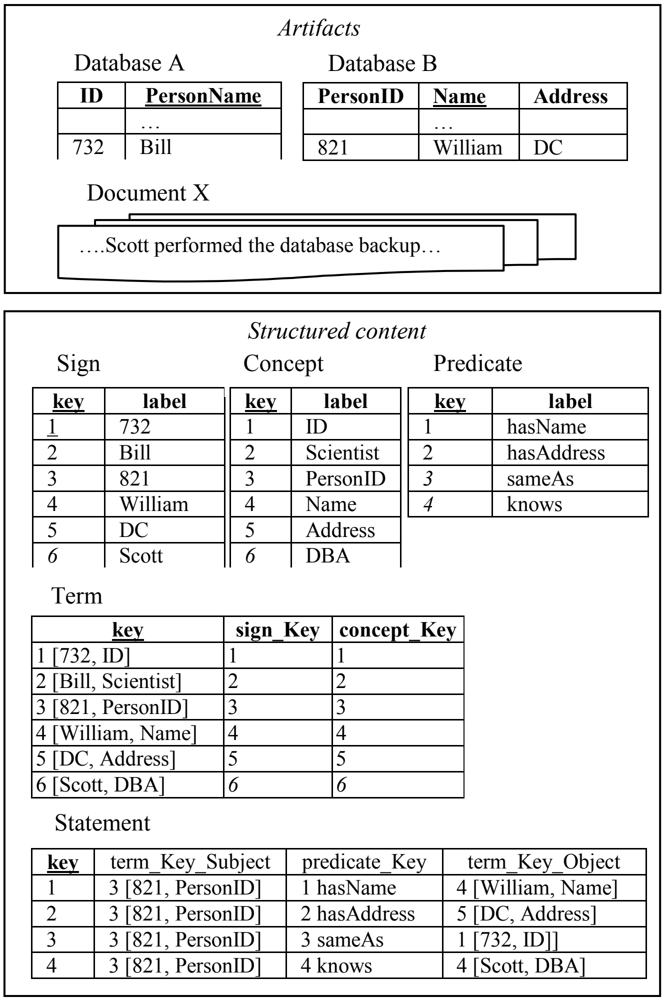
The five primitives of the DRIF (sign, concept, predicate, term, and statement) define a data reference model which, by effectively decoupling data from data-models, can represent any sort of data-model at the level that is useful for integration.

Fig. 2 schematically illustrates the representation of structured data in accordance with the DRIF for three sample primary artifacts, two of them relational databases, the third an unstructured document. The example also shows how data-semantics come to be added to the Dataspace in *ad hoc* fashion – here, because an analyst decides to to introduce a new Concept DBA (meaning: database administrator). Additional Statements establishing relationships between Terms using Predicates SameAs and Knows are also included in the Figure.

The reader familiar with the Resource Description Framework (RDF/RDFS) may wonder what is different here. RDF employs a similar level of abstraction, but it is a language, while what we are offering here is a specific, albeit still highly abstract, data-model. This data-model could of course be specified very easily using the RDF language; but it could be specified also using relational database or some other storage technology. Our choice of data-model was motivated further by the fact that our implementation and security requirements dictated the use of a specific type of cloud storage solution [5, 6] that is both highly scalable and offers highly granular security access controls.

III. Semantic Enhancement

The DRIF focuses on the representational aspects of the Dataspace and on the basic types of data integration that such representation provides. In what follows we describe the current phase of evolution of DRIF, the phase of Semantic Enhancement (SE). SE, as we conceive it, is a light-weight and flexible solution that leverages the richness of the native source data and of any local semantics associated with these data without adding storage and processing weight. The SE strategy is compliant with and complements the DRIF.



Artifacts

Database A

ID	PersonName
	…
732	Bill

Database B

PersonID	Name	Address
	…	
821	William	DC

Document X

….Scott performed the database backup…

Structured content

Sign

key	label
1	732
2	Bill
3	821
4	William
5	DC
6	Scott

Concept

key	label
1	ID
2	Scientist
3	PersonID
4	Name
5	Address
6	DBA

Predicate

key	label
1	hasName
2	hasAddress
3	sameAs
4	knows

Term

key	sign_Key	concept_Key
1 [732, ID]	1	1
2 [Bill, Scientist]	2	2
3 [821, PersonID]	3	3
4 [William, Name]	4	4
5 [DC, Address]	5	5
6 [Scott, DBA]	*6*	*6*

Statement

key	term_Key_Subject	predicate_Key	term_Key_Object
1	3 [821, PersonID]	1 hasName	4 [William, Name]
2	3 [821, PersonID]	2 hasAddress	5 [DC, Address]
3	3 [821, PersonID]	3 sameAs	1 [732, ID]]
4	3 [821, PersonID]	4 knows	4 [Scott, DBA]

Figure 2. Simplified example of structured content derived from 3 primary artifacts.

A. *Goals of Semantic Enhancement*

SE is a strategy that is currently being implemented to improve our handling of the enormous heterogeneity of Dataspace content. It is centered on building a flexible and extensible framework of hierarchically organized, controlled structured vocabularies – called 'ontologies' – covering different areas of relevance to intelligence analysis. The framework will be constructed in part by reusing already existing resources, in part through collaboration with other defense and military organizations in the creation of new ontology modules. The ontologies will be used in an incremental process of annotation (or 'tagging') of those concepts and predicates already identified in data-models within the Dataspace along the lines described in our discussion of Segment 3 above. The latter amount to what we referred to above as '*ad hoc* semantics'. Because the salient data-models derive from so many heterogeneous sources, they use a multiplicity of partially overlapping and partially conflicting vocabularies, which it is the task of SE to reconcile by associating co-referring concepts and predicates (strings) employed within distinct data-models in the Dataspace to single nodes within the external SE ontologies.

To function in the needed way, annotations must be cumulative, in the sense that our strategy will ensure that tags created by different annotators will be consistent with each other. The value of annotations must also be preserved when the SE ontologies change, for example through refinements created to reflect advances in knowledge, and to this end the ontologies must be subject to strict versioning policies.

Finally, the SE framework must be implemented in such a way that it can serve not merely as a tool of harmonization of the data-models internal to the Dataspace but also in a way that allows integration with other, external data resources wherever common ontologies are used for annotation.

To address these constraints is by no means a simple matter. When data value codifications do not match – for example when we have 1,2,3 in one data source, R, G, B in another data source, and RED, GREEN, BLUE in our Color ontology, then annotation for each source to hierarchy values can be very labor intensive and require significant SME effort.

B. Sample Benefits of Semantic Enhancement

We can see the sorts of benefits that SE will provide already at the level of search, where problems arise because of the multiple different ways of describing data within the Dataspace. Problems that need to be confronted include:

1. The need to find data items identified by means of terms which are *narrower* or *broader* in meaning than the terms analysts will standardly use when searching;
2. The need to find data items in documents that are formulated using a language or technical jargon with which analysts are unfamiliar.

To provide some very simple examples: we know that a given package 'has been shipped with a red label', but the documents that we have pertaining to this package use only the word 'vermillion'; or we need to find references to a package identified as 'containing furniture', but the documents we have refer only to 'chairs'; or we need to find a given package suspected of containing crack cocaine, but the audio recordings we have at our disposal relating to this package refer only to 'bobo' or 'botray' or 'boubou'. If we are restricted to string search, our queries would not return the needed results. Hence, we need a framework which expands string search by capturing type and subtype information, and also incorporates synonym information. These needs are targeted along two dimensions; first, through the fact that all SE ontologies will be organized around a central backbone subtype (or *is_a*) hierarchy; and second through the progressive incorporation in all nodes of the SE ontologies of links to relevant synonyms derived through the annotations which will link ontology nodes to the rich collection of corresponding concepts and predicates in other areas of the Dataspace.

C. The Strategy for Semantic Enhancement

Our strategy is designed to achieve its goals not by changing the Dataspace, but rather by adding an extra *semantic layer* thereto. The strategy is thus similar to that underlying the Universal Core (UCore), which arose out of the National Information Sharing Strategy supported by multiple U.S. Federal Government Departments, by the intelligence community, and by a number of other national and international organizations [7, 8]. Here, a small controlled vocabulary was provided for multi-community use to associate simple summary tags to message payloads for purposes of data search and integration.

Reflecting the extreme diversity of intelligence data, multiple subject-matter expert communities will be contributing to the SE. For the strategy to work and provide useful and efficient integration, these multiple distributed teams must use the SE approach in a consistent fashion. Previous efforts to create a broad-based, multi-community ontological approach to data integration in defense and intelligence domains have failed because the incompatible, and often over-simplistic, views of reality incorporated into legacy databases and data-models led to incompatible development of ontologies in ways that precluded interoperability. Many advocates of semantic approaches to data integration have still failed to appreciate the tremendous challenges, both technical and human, created by the entrenched predisposition on the part of ontology developers to create ontologies each on the basis of their own potentially idiosyncratic data representations.

The solution which we advocate is modeled on the successful semantic annotation approach pioneered in the field of bioinformatics by the Gene Ontology [9]. This approach is now being pursued systematically within the framework of the OBO Foundry [10, 11], which starts out from the idea that the most effective way to ensure mutual consistency of ontologies created by multiple independent groups over time and to ensure that these ontologies are maintained in such a way as to keep pace with advances in knowledge is to organize ontologies as a collection of modules with discrete (non-overlapping) subject-matters maintained by subject-matter experts, according to a strategy outlined in [12]. To ensure consistency, these ontologies should be created as extensions of more generic higher level ontologies, subject to common rules for example concerning the treatment of definitions, and they should be based on a small common upper-level ontology (ULO), whose domain and content neutral. For example, it will include relations such as *is-a* (for subtype), *member-of*, *part-of*, and so on. As initial ULO we choose the Basic Formal Ontology (BFO) [13], which has been implemented in more than 100 similar projects, and which serves as the basis of the already mentioned UCore Semantic Layer [8].

The ULO will be associated with a small number of Mid-Level Ontologies (MLOs) defined by downward population from the ULO. The MLOs will serve in turn as bridge to a number of Low-Level Ontologies (LLO), which will specify narrow content domains. Each MLO represents cross-domain entities, such as Person or Information, and will be constructed in tandem with the LLOs which it subsumes in order to ensure the mutual consistency and interoperability of the subsumed LLOs. The MLOs and LLOs must in turn be associated with the resources of a relation ontology, providing for the representation of content-specific relations such as Owns, WorksFor, Audits, and so on.

Initial due diligence efforts in our strategy of semantic enhancement requires us to identify an initial collection of authoritative codifications at Mid- and Lower Levels – along roughly the lines depicted in Table 1 – and to begin the process of formalizing them within the BFO common upper-level ontological framework. In some areas ontologies will need to be created *de novo*, since no adequate authoritative codifications will exist.

Examples of MLO cross-domains

- Geospatial
- Biometrics
- Person
- Provenance and Trust
- Organization
- Signals and Sensors
- Equipment
- Facility

Examples of LLO domains

Subsumed by Geospatial

- Geospatial Feature
- Country

Subsumed by Biometrics

- Fingerprint
- Iris

Subsumed by Person

- Employment Data
- Criminal Data
- Medical Data
- Ethnicity and Tribe
- Skill

Subsumed by Provenance and Trust

- Data Quality
- Access Permissions
- Data Source
- Evidence

Table 1: Sample Ontologies within the SE Structure

D. *Implementation of the SE Strategy*

We can now outline the steps which are involved in realizing this strategy in the specific context of the Dataspace, where we already have data structured using the DRIF.

First Step: Review the contents of the Dataspace, specifically that concepts and predicates in Segment 1, and identify a subset of topic areas where data integration is a priority for analytics.

Second Step: Formulate a list of MLOs that would be needed to annotate the data in corresponding areas. As far as possible identify existing ontologies which may potentially be reused for this purpose, and build initial versions of new ontologies where needed.

Third Step: Identify a specific subset of the content of the source data-models, and identify LLOs that will capture this subset in a semantically coherent fashion, ensuring that each LLO is subsumed by some MLO. Subject matter experts should be recruited to take charge of creation and maintenance of the LLOs and MLOs and of their use in annotations. In this way we can create a cadre of SMEs with expertise in annotation and in supporting semantic enhancement.

In realizing the above we need to maximize as far as possible the *reuse* of ontologies which are already being used by relevant communities. This is because the strategy will be successful only to the degree that a critical mass of potential users are able to be convinced of its utility and thus incentivized to engage in advancing it further for example by extending it new types of data and by disseminating the resource to new groups of analysts. Reusing already existing ontologies will not merely provide a core of familiar terms which analysts can use for search purposes, it will also increase the degree to which we can integrate into the Dataspace data that has already been annotated in consistent fashion by external bodies.

Fourth Step: When once a stable, initial set of ontologies has been created, we use these ontologies to annotate the data-models in corresponding portions of the Dataspace. As should by now be clear, the entire strategy is an incremental one, based on a principle of low hanging fruit: the idea is not to import the above ontologiesas a whole; rather we examine the existing Dataspace resources and identify expressions therein for which counterparts in the ontologies already exist or can easily be added. In constructing the ontologies these expressions will be provided with a common logical architecture and a common set of relations defined through the ULO top level and in terms of which logical definitions for terms in the ontologies can then be formulated. The result can be used as a basis for the application of general-purpose tools, including standard OWL reasoners FaCT++, RACER, or Pellet, which can be used to check ontologies in the SE resource for mutual consistency.

Stage 4 of the SE process consists in associating each set of equivalent data source concepts with a single common MLO or LLO expression (which will be added at the appropriate level within the SE ontology structure where not already present). Further types of integration are thereafter brought about automatically. Whenever any Dataspace resource becomes linked to one of our chosen ontologies in a way that can be used to generate corresponding annotations, it thereby becomes linked to all the other Dataspace (and external) resources that have already been annotated with the same SE ontologies. This creates a snowball effect, whereby each new annotation increases the value of existing annotations [9], and provides further incentives for the use of the SE ontologies by new groups of users.

E. Organization of the SE Ontologies

Fig. 3 illustrates the organization of the SE ontology space. Each LLO represents the reality of a particular narrowly defined domain, for example in an area such as Education and Skills.

An MLO is a container of LLOs. Since we will be developing LLOs in step-by-step fashion to address what are at any given time the most urgent needs of Dataspace users, there will be data which cannot as yet be annotated with the full granularity of detail which the annotator requires. The strategy is to use such cases to advance the further development of the ontology resource base, again following the model tested in the bioinformatics domain [9]. For example, an analyst may want to use the SE resources to extract and disambiguate data from a particular document. For different reasons the analyst may not be able to use the most detailed semantics and will use a more general one. LLO taxonomies will also be used by analytics to produce results of different level of detail: from a fine-grained view of narrow areas within the Dataspace to coarse grained pictures of larger domains.

Because original data and data-semantics are in every case preserved without loss or distortion in the Dataspace as it exists prior to Semantic Enhancement, there is no need to represent all details of original storage data structures in the SE stage. This means that complex ontologies are not needed – a common and shared vocabulary is sufficient for virtual semantic integration and search/analytics, while underlying details are maintained by the authors of specific primary artifacts. Similarly, the collection of SE ontologies does not need to cover all of the *ad hoc* local semantics within the Dataspace – content that is unlikely to be used in search or is not important for integration can be excluded from the Enhancement step, since it will still be available in the source data-models and can be accessed when drilling down to the appropriate level.

The SE approach is highly flexible. It represents a "pay-as-you-go" approach in the sense that investments can be made only in specific areas according to identified need. It is also tunable in the sense that, if a given body of annotations for a particular subset of a source data-model is too general for data analyst purposes, then the respective LLOs can be further developed as needed.

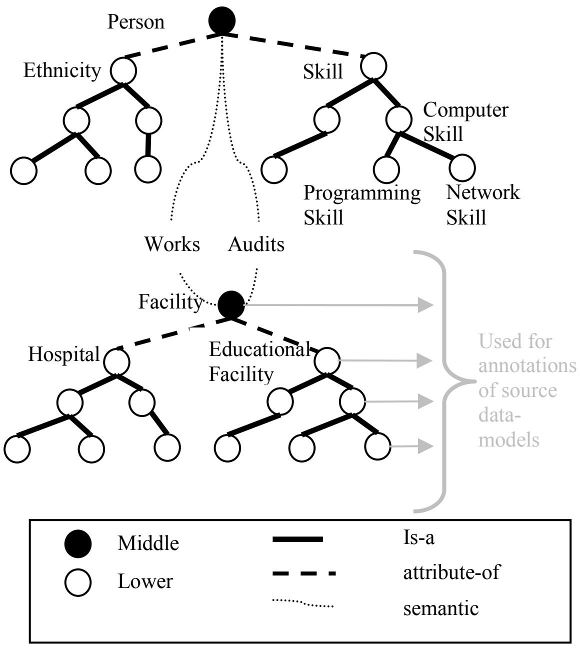


Figure 3. Simplified Example of an SE Ontology Structure.

IV. CONCLUSION

Together, the DRIF and SE provide what we believe is a workable data-integration solution. The DRIF is a highly flexible framework, with few constraints and including an RDF-style decomposed representation of structured data which allows the collection of data resources without loss or distortion in a way that achieves syntactic integration and preserves the local semantics of primary sources and of analytics software. SE provides semantic integration in a light-weight yet incrementally extendible fashion, and in a way that can foster global integration without adding storage and processing weight to already storage- and processing-heavy Dataspace.

The SE approach provides a strategy to allow the Dataspace to be understood as evolving *cumulatively* as it accommodates new kinds of data. It provides a more *consistent, homogeneous,* and *well-articulated* representation of structured content that originates in multiple internally inconsistent and heterogeneous models. And while it involves considerable initial SME investment in ontology creation and annotation, we believe that it will allow the management and exploitation of the Dataspace to become more *cost-effective* over time.

In addition, the use of the selected MLOs and LLOs brings integration with other government initiatives and brings the Dataspace endeavor closer to the federally mandated net-centric data strategy; it also makes the integrated Dataspace more effectively searchable and provides an expanding body of content to which more powerful analytics can be applied in the future.

Acknowledgments: This work was funded by US Army CERDEC I2WD. The authors thank Mr. Kesny Parent, DCGS-A Branch Chief, for continued support.

REFERENCES

[1] S. Yoakum-Stover, T. Malyuta, N. Antunes, "A Data Integration Framework with Full Spectrum Fusion Capabilities", Presented at the Sensor and Information Fusion Symposium, Las Vegas, NV, Aug 3-7, 2009.

[2] A. Hansen, D. Salmen, T. Malyuta, and N. Antunes. "An Evolving Integrated Dataspace on the Cloud." Presented at the Sensor and Information Fusion Symposium, Las Vegas, NV, July 26-29, 2010.

[3] S. Yoakum-Stover, T. Malyuta, "Unified Integration Architecture for Intelligence Data", Proceedings of DAMA International Europe Conference, London, UK, 2008.

[4] Rosse, C. and Mejino, J. L. V. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics 36, 2003, 478-500.

[5] R6 Cloudbase documentation and source code.

[6] Hadoop. http://hadoop.apache.org.

[7] http://ncor.us/ucore-sl.

[8] B. Smith, L. Vizenor and J. Schoening, "Universal Core Semantic Layer", Ontology for the Intelligence Community, Proceedings of the Third OIC Conference, George Mason University, Fairfax, VA, October 2009, CEUR Workshop Proceedings, vol. 555.

[9] D. Hill, et al., "Gene Ontology Annotations: What they mean and where they come from", BMC Bioinformatics, 2008; 9(Suppl 5): S2.

[10] B. Smith, et al., "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration", Nature Biotechnology, 25 (11), November 2007, 1251-1255.

[11] B. Smith, W. Ceusters "Ontological Realism: A methodology for coordinated evolution of scientific ontologis", Applied Ontology 5 (2010) 139-188, http://x.co/adRJ.

[12] W. Ceusters, B. Smith, J. M. Fielding, "LinkSuite^{TM}: formally robust ontology-based data and information integration," in Database Integration in Life Sciences, Berlin, Springer, 2004. http://ontology.buffalo.edu/bio/LinkSuite.pdf

[13] Basic Formal Ontology. http://www.ifomis.org/bfo/.