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

Overview

- Background
- Data Integration
- Evolving Cloud Platform
- Unified Integrated Data Representation
  - Hundreds of dynamic data sources, both structured and unstructured
  - Lossless, ad hoc data integration
- Semantic Enhancement
  - Richer cross-source analytic capability

Importance of Intelligence Data Integration

- 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

Evolving Cloud Platform

- Drivers
  - Scale limitations
  - Security - Fine Grain Access Control
- IC/Army Cloud Platform
  - Hadoop
  - BigTable: Cloudbase (Accumulo)
  - Solr
- Future Cloud Platform
  - Prism: Semantic Adapter for Cloudbase
    - Adjudicating Triple Store
    - Authorization control to Jena model

Data Integration

- Data storage models provide structure for data used within information systems
  - Specific definition and format
  - Limited in scope
  - Biased toward implementation strategy of platform
- Semantic models
  - Define data from conceptual view
  - Define meaning of data within context of interrelationships with other data
  - Used for integration of existing data storage models

Data Impedance Mismatch

- Data Structures
  - Granularity/decomposition
- Data Values
  - Codification
  - Format

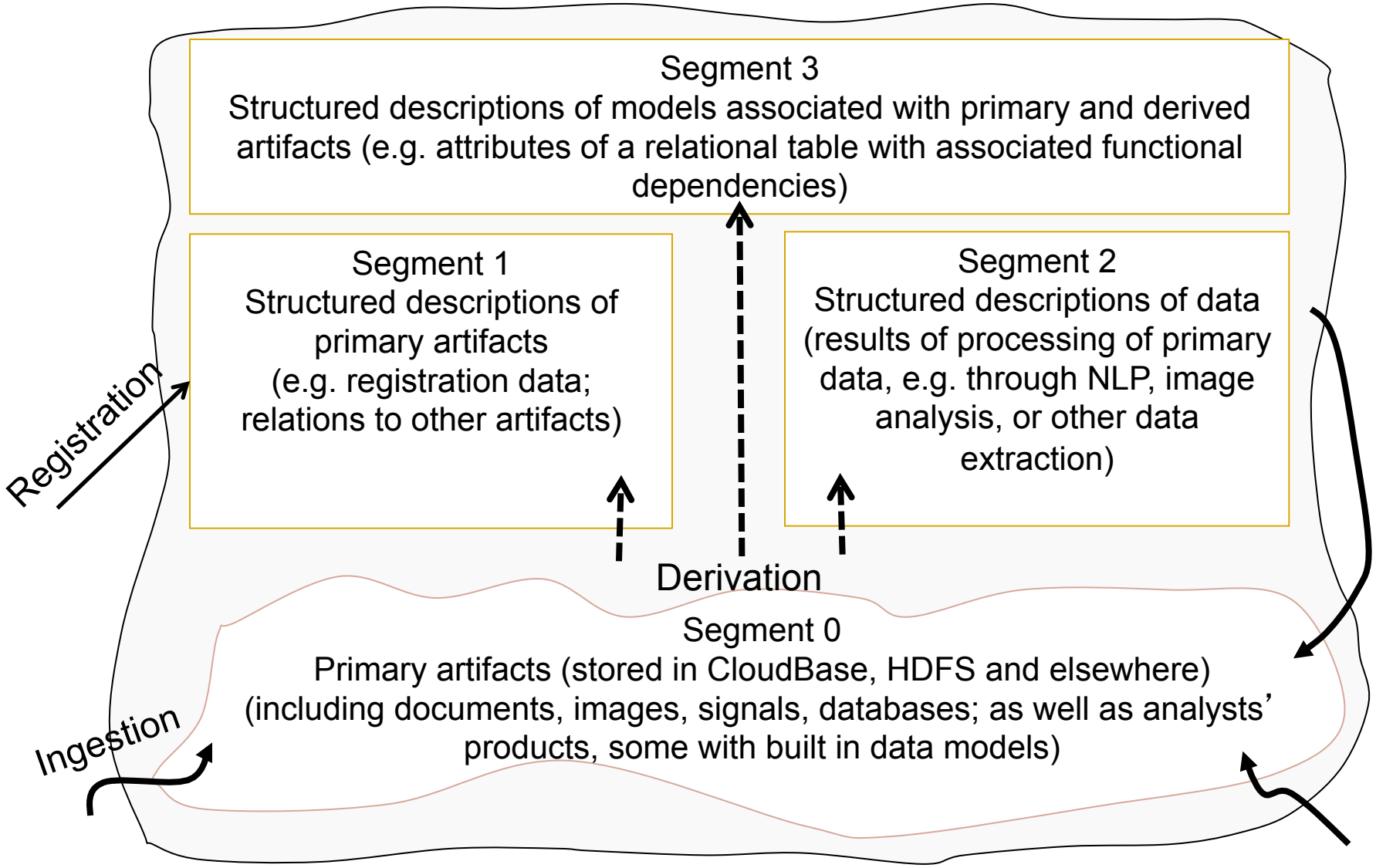
Intelligence Data Integration – Challenges

- Integration must involve the ability to proceed:
  - Without heavy pre-processing
  - Regardless of the data-models used (or not used) in source data
  - Without loss or distortion of data, of its associated data semantics, and of 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

Unified Integrated Data Representation

- Must present minimal barriers to the incorporation of new data sources
- Require no heavy pre-processing and no data or data-model conditioning
- Embrace 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

Unified Integrated Data Representation



Data Description

- Where models and data come together
- Data integration across a constantly evolving and highly heterogeneous resource comprehending extremely large volumes of data
- Defining features
  - Data are exposed in a way that is independent of its original intended use
  - Original data-models are represented at a level of abstraction that is higher than that of primary data
- Applies 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

Current State

- Currently we have mostly syntactic integration
  - Of hundreds of millions of unstructured documents, representations of image and signals data, and other unstructured and structured primary data artifacts
  - Of results of various analytic processing of these artifacts
- Data can be utilized immediately upon ingestion for search and other analytic processing
- At best: ad hoc semantic integration
  - Tied to specific local implementations
  - Typically falls short of what is needed to secure semantic interoperability

Semantic Enhancement

- Previously focused on the representational aspects of data and on the basic types of data integration that such representation provides
- The current phase focuses on Semantic Enhancement (SE)
- SE is a strategy that is being implemented to improve our handling of the enormous heterogeneity of repository content
- SE 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

Goals of SE

- Semantic data enrichment is achieved incrementally, through the step-by-step creation of ontology modules
- A lightweight, flexible approach comprising an *extra ontology layer* that leverages the contents of the repository without adding storage and processing weight
- Simple yet efficient harmonization
  - Takes place not by changing the data-semantics to which it is applied, but rather by adding an extra *semantic* layer to it
  - To support comprehensive and relevant cross-model data analytics
  - Long-lasting solution that can be applied *consistently* and in *cumulative* fashion to new models entering the repository

Goals of SE (cont.)

- Semantic enhancement will
  - Be efficiently and in a unified fashion used in search (and eventually in automated reasoning) and analytics
  - Provide views of the repository with different level of detail
  - Can serve not merely as a tool of harmonization of the data-models internal to the repository but also in a way that allows integration with other, external data resources

Solution – Basic Principles

- Incrementally growing framework of ontologies (management of ontologies is performed by trained professionals)
- Based on a small upper level ontology (ULO)
  - Defines the common *architecture* for the lower-level content- or domain-specific ontologies and is managed by a common governing body spanning all communities of interest
  - Basic Formal Ontology (BFO), which has been implemented in more than 100 similar projects
- With a small number of mid-level ontologies (MLOs)
  - Represent *multi-* and cross-*domain contents* and are LLO containers
  - *Bridge* the ULO and LLOs
  - Are defined by users
- With *data-semantics* contained in Low-Level Ontologies (LLOs)
  - Represent the *specific narrow homogeneous domain* contents

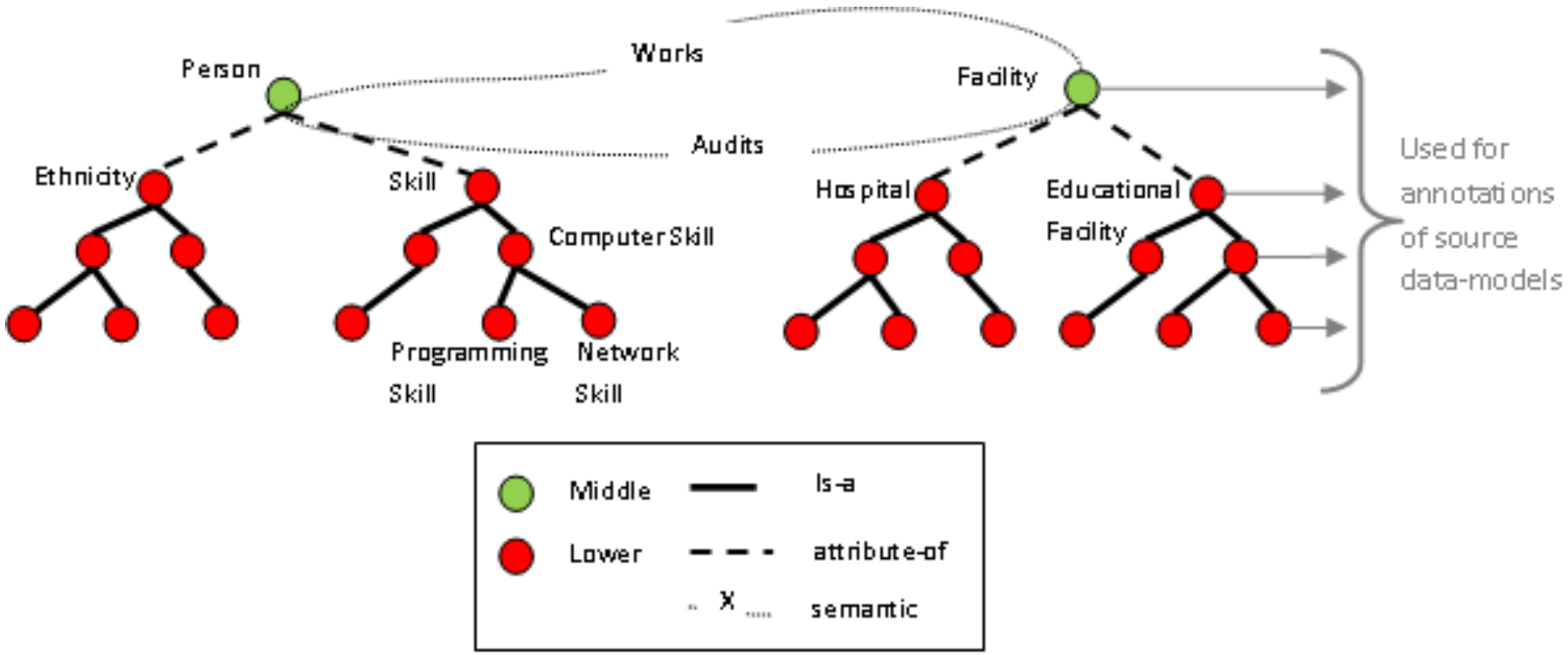
Examples of MLOs and LLOs

- **MLO Cross-Domains**
  - Geospatial
  - Biometrics
  - Person
  - Provenance and Trust
  - Organization
  - Signals and Sensors
  - Equipment
  - Facility
- **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

SE Implementation Strategy

- Step 1 – Identify a subset of topic areas where data integration is a priority for analytics and find existing ontologies which may potentially be reused
- Step 2 – Formulate a list of MLOs that would be needed to annotate the data in corresponding areas
- Step 3 – 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
- Step 4 - Use these ontologies to annotate the data-models in corresponding portions of the repository

Organization of SE



Relations Within the SE

- MLO:
  - An MLO represents an entity with multi- or cross-domain contents, e.g. Person. It is constructed from the LLOs
  - One of the MLOs will be the Ontology of Relations such as: Owns, WorksFor, Audits, etc. between Person and Facility MLOs
- LLO:
  - Each LLO represents semantics of a particular domain, e.g. Education/Skills
  - The semantics of a LLO is organized primarily using the ULO relationship *is-a*. Super-class – sub-class hierarchies will be used for annotating, and, respectively, for search
  - Each LLO represents an independent semantic content – LLOs do not share nodes
  - Other relationships between the nodes of the LLOs can be added to be used in data quality control and advanced analytics

Benefits of the Approach

- Ensures that the repository evolves in a cumulative fashion as it accommodates new kinds of data
- Provides a more consistent, homogeneous, and well-articulated representation of structured content which originates in multiple internally inconsistent and heterogeneous models
- The use of the selected MLOs and LLOs brings integration with other government initiatives and brings this endeavor closer to the federally mandated net-centric data strategy
- It will create an integrated repository that is effectively searchable and that provides content to which more powerful analytics can be applied
- Quick ROI
  - Represents a "pay-as-you-go" approach – investments can be made only in specific areas according to identified need
  - Management and exploitation of the repository will become more cost-effective

Conclusions and Potential Risks

- It is important that the annotations are built using a common, controlled vocabulary
- The common set of ontologies must be used in a consistent fashion by all of those engaged in the task of annotation. This will require dedicated, joint training of those charged with the task of annotation
- Will require the establishment of an ontology governance process The upper level ontology will be managed by a small governing board spanning all communities of interest; the lower level ontologies will be managed by groups of domain experts, each of which will involve at least one member from the governing board
- Will bring ROI only if people use it. Needs dissemination across the community and buy-in from high authorities

Acknowledgements

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

Back-Up

DRIF Abstract Data-Model

- Sign - a string that is the abstracted proxy within the repository for one or more chunks of data used in some primary artifact
- Concept – a string that is used in the repository to represent some general category or grouping. Is used represent and allow reuse of classifications native to primary artifacts
- Term – an ordered pair (Sign, Concept). Results from a process of contextual disambiguation of a sign, a process which associates a sign with a concept.
- Predicate (by which we mean here binary relational predicate) – a string that is used to connect terms in accordance with domain and range constraints
- Statement –an ordered triple consisting of a subject, a predicate, and an object

Additional Notes on Annotating

- Original data and data-semantics are included in the repository without loss and or distortion, and there is no need to represent all details of original storage data structures in the Semantic Enhancement
- The vocabulary not necessarily has to cover all semantics of the repository – semantics that are unlikely to be used in search or are not important for integration can be not included in the Enhancement. These semantics will still be available in the source data-models and can be accessed when drilling down to them
- A complex ontology is not needed – a common and shared vocabulary is sufficient for virtual semantic integration and search/ analytics
- The approach is very flexible, and investments can be made in specific areas according to need (pay-as-you-go)
- The approach is tunable – if the chosen annotations of a particular subset of a source data-model are too general for data analyses, the respective LLOs can be further developed and source models re-annotated

Example of Search