Managing Semantic Big Data for Intelligence

Anne-Claire Bourny-Brisset, PhD
DRDC Valcartier – C2I Section

STIDS – 12-15 November 2013
Outline

- Intelligence context
- Information management and integration challenges
- Proposed approach and architecture
- Ontology support
- Enabling technologies
- Future work and conclusions
The problem: Data Variety, Volume, Velocity ...

**Intelligence** is about data: Collection, Processing, Discovery, Retrieval, Exploitation, Analysis, Dissemination

- Increase of sensor data volume (terabytes – petabytes – exabytes)
- Heterogeneity: multiple data formats and standards, mix of structured and unstructured
- Need to quickly acquire and process intelligence information
- Agility is required to be able to incorporate new data sources

**Support to data exploitation**

- Each piece of data represents some part of a situation
- Intelligence data contain entities that must be understood and correlated
Context and objectives

Military Intelligence context
- Increasing amount of data/information stored in stove-piped systems
- Multi-sources: SIGINT, IMINT/GeoINT, HUMINT, OSINT, etc.
- Various formats: sensor data, multimedia (text, images, audio, video)
  - Hard/soft, structured/unstructured
- Information overload

Objectives
- Develop a Multi-Intelligence Data Integration System (MIDIS)
- Build on prior R&D work
  - Domain ontologies, annotation, fact extraction, etc.
- Leverage Semantic and Big Data technologies
- Better support intelligence analysts in fusion & analytical tasks
Approach

- Underlying concepts
  - Dataspase: incorporation of large heterogeneous data
    - co-existence approach (Franklin, Halevy)
  - Unified data representation and integration framework (Yoakum-Stover) exploiting ontologies for semantic enrichment (Salmen, Malyuta, Smith)

- Data flow and processes for data integration
  - Data ingestion mechanism from heterogeneous data sources
  - Semantic enrichment, alignment (data source model, domain ontologies)
  - Ontology support (incremental ontology development)
  - Unified query mechanism
Unified Data Space layered architecture

- **Data-Models Segment 3**
  - Ontologies
  - Data models
  - HBASE

- **Structured Data Segment 2**
  - Concept
  - Predicate
  - Statement
  - Source
  - Source Models
  - HBASE

- **Artefacts Segment 1**
  - Artefact
  - Data
  - Source Models
  - HDFS

- **External Data Sources & Systems**
  - XML
  - GMTI
  - OSINT
  - DB
  - Text Reports

(Adapted from Yoakum-Stover, 2012)
Intelligence Data Integration and Analysis

Big Data Store
- Raw Data (HDFS)
- Structured Data (HBASE)

Transformation Process

KB
- Propositions
- Situation Models
- Hypotheses
- Spatial Features

RDF Store

Reasoning Services

Data management Services
- XML
- CSV
- RDF
- Text

Semantic Alignment

Big Data Search & Analytics

Intelligence Data Integration and Analysis

Semantic Alignment

Big Data Search & Analytics
Ontology support

Taxonomy / Thesaurus
- Terms
- Synonyms
- Generalization/specialization
  vs
- Broader/narrower relations

Terminology
Controlled vocabulary

Ontology
- Concepts
- Attributes
- Relations
- Functions
- Constraints
- Axioms

Communication Metadata

Knowledge organization Categorization

Inference capabilities Automated reasoning
Intelligence ontology(ies)

Role
- Formal reference model for the intelligence domain
- Semantic enrichment, annotation, integration / mapping
- Reasoning / inferencing

Requirements: expressiveness, flexibility, modularity

Development: reuse, incremental extensions

Scope - domains
- Intelligence high-level concepts
  - Physical entities, people/groups, event/activities, feature, information, etc.
- Domain specific models
  - Threat assessment
  - Human geography
  - Terrorism
Semantic enrichment & alignment with ontologies

Aim
- Data annotation and alignment according to ontologies to address data source semantic heterogeneity
- Facilitate unified querying of heterogeneous data
- Facilitate heterogeneous data correlation and fusion

Strategy
- Annotation of structured data sources
  - Establish mapping: Data source term – reference ontology term
- Annotation of unstructured data sources
  - Original source is annotated using terms of ontologies
  - Extraction of metadata, facts and statements (structured data)

Benefits
- Better support of intelligence analysts in the production of intelligence
Domain of interest – Key high-level concepts

Who
Actor / Agent
Person
Organization

What
Event / Action
Materiel
Facility

When / Where
Location

Physical Entities

Descriptive (qualifier/quantifier)
Ontology development - Modularity

(Source Pulvermacher et al, Mitre, 2004)

(Source : Barry Smith - NCOR)
### Upper-level constructs

<table>
<thead>
<tr>
<th>Continuants</th>
<th>Occurrents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Artifact</td>
<td>Process Event / Action</td>
</tr>
<tr>
<td>Agent</td>
<td></td>
</tr>
<tr>
<td>Dependent Entity</td>
<td></td>
</tr>
<tr>
<td>Information Artifact</td>
<td></td>
</tr>
<tr>
<td>Equipment</td>
<td>Military Event</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Social Event</td>
</tr>
<tr>
<td>Facility</td>
<td>Terrorist Event</td>
</tr>
<tr>
<td>Vehicle</td>
<td>...</td>
</tr>
<tr>
<td>Weapon</td>
<td>Planning Process</td>
</tr>
<tr>
<td>Geospatial Site</td>
<td>Operation</td>
</tr>
<tr>
<td>Person</td>
<td></td>
</tr>
<tr>
<td>Organisation Group</td>
<td>Report</td>
</tr>
<tr>
<td>Quality Function</td>
<td>Info Reqt</td>
</tr>
<tr>
<td>Property Role</td>
<td>...</td>
</tr>
<tr>
<td>Plan</td>
<td></td>
</tr>
<tr>
<td>Product</td>
<td></td>
</tr>
<tr>
<td>Report</td>
<td></td>
</tr>
<tr>
<td>Info Reqt</td>
<td></td>
</tr>
<tr>
<td>Reqt</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Leveraging Semantic and Big Data Technologies

- **Semantic Technologies**
  - OWL Ontologies, RDF triples, mapping

- **Big Data Technologies:**
  - Apache Hadoop Framework (Cloudera) – HDFS / HBase
  - Indexing and query mechanisms
    - HDFS, HBase (e.g. Impala)
    - Index tables (permutations of triple patterns) - Sparql query
  - Data Analytics (e.g. Mahout)
    - Data clustering, filtering, profiling

- Integration within a SOA-based Intelligence S&T Integration Platform
Conclusions and future work

- Incremental, flexible approach to data integration
  - Agility, modularity, extensibility
  - Enhanced support to intelligence analysis: data query, correlation, fusion, reasoning
  - Enabler to evolve from single Int production to Multi-Int

- Ontology support
  - Combination of top-down, bottom-up, and horizontal development of ontologies

- Big Data technologies
  - Benefit from distributed processing (volume)
  - Unstructured data (HDFS) – Structured data (HBase) processing
  - Emerging, still immature

- To be investigated further:
  - Data analytics
  - Additional data management services, e.g. Entity resolution
  - Data uncertainty