A Semantics-Based Approach to Schema Matching and Transformation in Network Centric Environments

Authors
Anton DeFrancesco
Bruce McQueary

Network Centricity

- DoD Net-Centric Data Strategy (NCDS) contains the following goals: Visible, Accessible, Understandable, Trustworthy, Interoperable, Responsive, and Institutionalized.
- DoD direction is clear regarding acquisition of new systems and direction for legacy systems: service oriented implementation and net-centric sharing of data are no longer optional.
- Schemas, or schemata, are used to form structure and shape the data and connectivity of the network and services.

XML and Web Services

- XML is the Extensible Markup Language and is used give structure to data. Data is typically represented in human readable form.
- Web Services are defined using Web Service Definition Language (WSDL) which define the overall service contract.
- XML Schema provides a framework for defining data structures and provides guidance to XML form.

Schema Example

```
<xsd:complexType name="BasicPatientStatistics">
  <xsd:sequence>
   <xsd:element name="bloodPressure" >
     <xsd:complexType>
      <xsd:sequence>
       <xsd:element name="systolic" type="xsd:int"/>
       <xsd:element name="diastolic" type="xsd:int"/>
      </xsd:sequence>
     </xsd:complexType>
   </xsd:element>
   <xsd:element name="pulse" type="xsd:int"/>
   <xsd:element name="respitory" type="xsd:int"/>
   <xsd:element name="temperature" type="xsd:int"/>
   <xsd:element name="time-and-date"
     type="xsd:dateTime"/>
  </xsd:sequence>
 </xsd:complexType>
```

Service Integration

- Typically a manual task requiring volumes of documentation and Application Programming Interfaces(API), if they exist.
- Services integrators compare data structures from external services with internal structures to determine suitability for integration.
- This can be problematic due to variations in vocabulary, data organization, and the sheer magnitude of data.

Why service integration?

- Vast amounts of information available to organizations that need to be obtainable in fast, efficient, and secure manner.
- Example is the Haitian relief effort where multiple branches of the military and organizations within each branch are involved.

Schema Comparison

```
<complexType name="PatientRelocation">
 <sequence>
  <element name="identifier" type="RelocationID"/>
  <element name="patient" type="Patient"/>
  <element name="patient-status" type="PatientStatus"/>
  <element name="origination-facility" type="Facility"/>
  <element name="destination-facility" type="Facility"/>
  <element name="authorizing-physician"
      type="Physician"/>
  <element name="receiving-physician"
      type="Physician"/>
  ...
 </sequence>
</complexType>
```

```
<complexType name="PatientTransferData">
 <sequence>
  <element name="transferringPatient" type="PatientData"/>
  <element name="initiatingDoctor" type="MedicalPersonnelData"/>
  <element name="receivingDoctor" type="MedicalPersonnelData"/>
  <element name="originatingHospital" type="HospitalData"/>
  <element name="destinationHospital" type="HospitalData"/>
  <element name="patientCondition"
      type="PatientCondidtionData"/>
  ...
 </sequence>
</complexType>
<complexType name="HospitalData">
 <sequence>
  <element name="name" type="string"/>
  <element name="address" type="AddressData"/>
  ...
 </sequence>
</complexType>
```

Schema Comparison

```
<complexType name="PatientRelocation">
 <sequence>
  <element name="identifier" type="RelocationID"/>
  <element name=”patient” type=”Patient”/>
  <element name=”patient-status” type=”PatientStatus”/>
  <element name="origination-facility" type="Facility"/>
  <element name="destination-facility" type="Facility"/>
  <element name="authorizing-physician"
      type="Physician"/>
  <element name="receiving-physician"
      type="Physician"/>
  …
 </sequence>
</complexType>
```

```
<complexType name="PatientTransferData">
 <sequence>
  <element name="transferringPatient" type="PatientData"/>
  <element name="initiatingDoctor" type="MedicalPersonnelData"/>
  <element name="receivingDoctor" type="MedicalPersonnelData"/>
  <element name="originatingHospital" type="HospitalData"/>
  <element name="destinationHospital" type="HospitalData"/>
  <element name=”patientCondition”
      type=”PatientCondidtionData”/>
  …
 </sequence>
</complexType>
<complexType name=”HospitalData”>
 <sequence>
  <element name=”name” type=”string”/>
  <element name=”address” type=”AddressData”/>
  …
 </sequence>
</complexType>
```

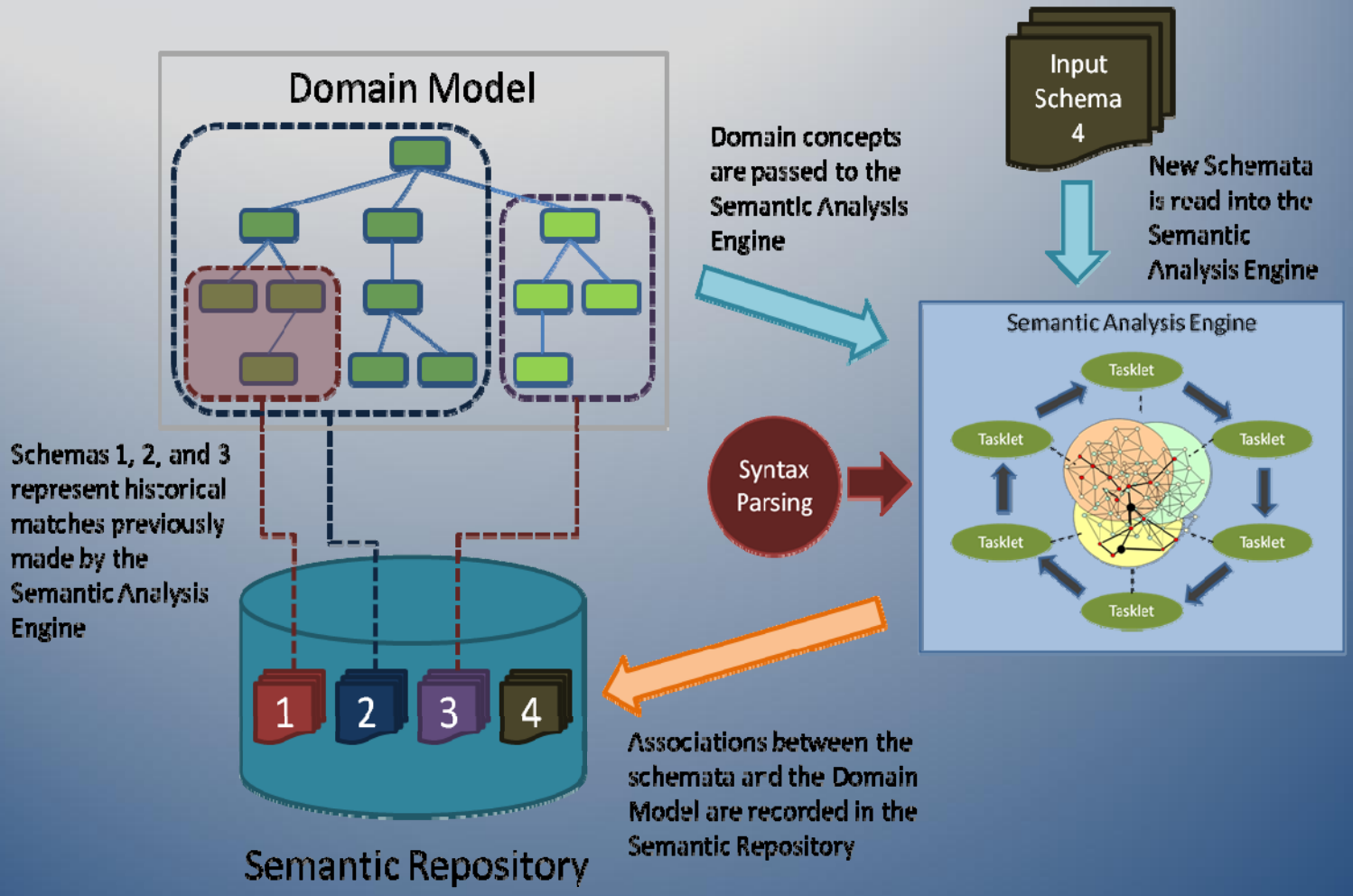
Syntactic vs. Semantic Analysis?

- Syntactic Analysis - Direct use of text, syntax, and statistics to determine matches.
  - Provides small amount of correlative power.
  - Coupled with a synonym set can provide reasonable functionality.
- Semantic Analysis – Leverages text, syntax, and domain models to determine meaning of data.
  - Utilizes structure of data to add validity to conclusions.
  - Provides ability to compare based on deeper meaning of data.

Semantic Matching and Transformation Service (SMTS)

- Provides services to compare two WSDL files against a Domain Model. Common elements are identified and stored for future comparisons.
- Syntactical and Semantic methodologies are used to identify similarities and conjecture the meaning of the elements within the WSDL.
- SMTS uses a continuous refinement method for narrowing down the scope.

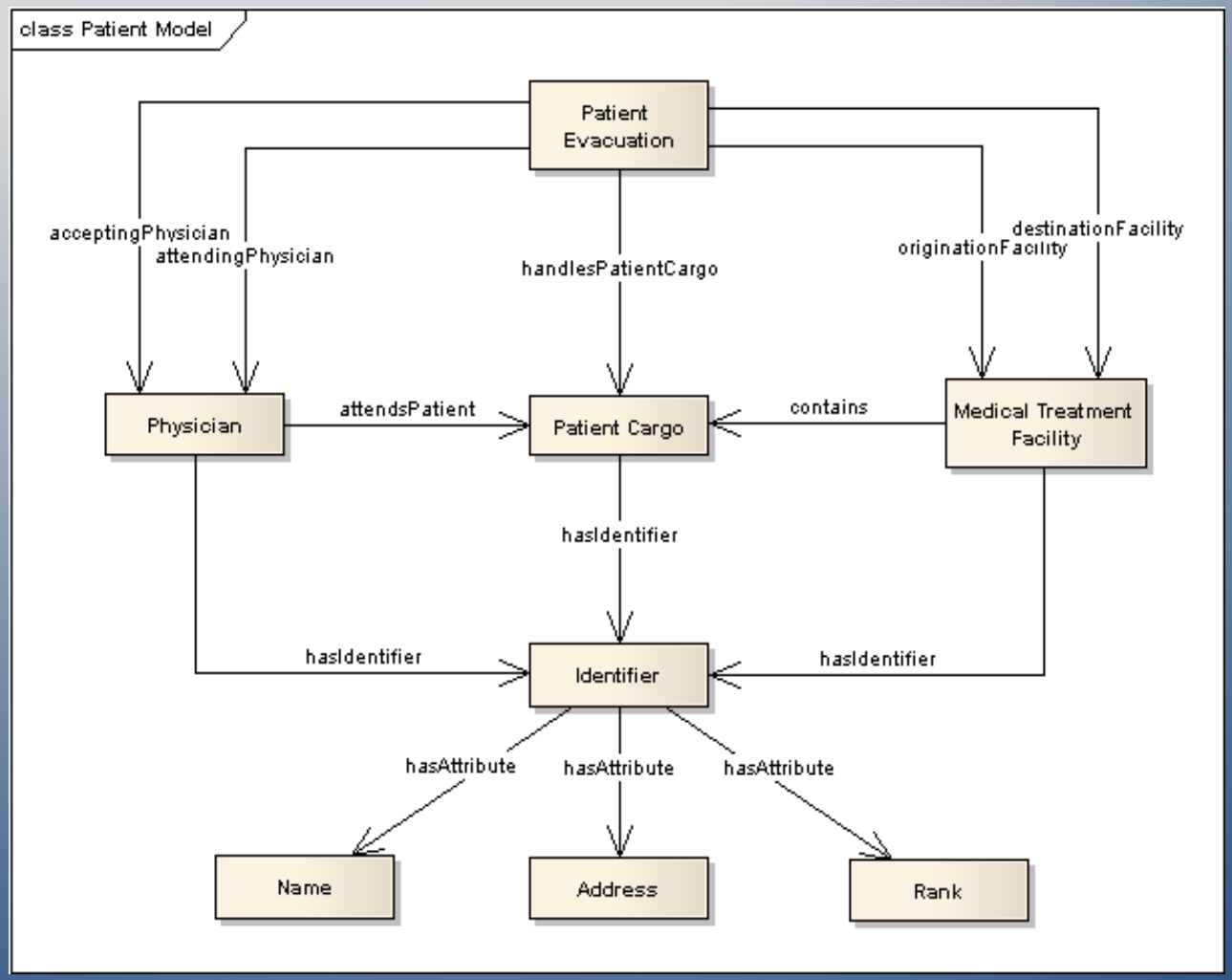
SMTS Architecture



Domain Model

- Ontological construct that is represented in OWL.
- Domain Axioms – small units of "truth" – are encoded into the domain. An example would be that a patient relocation requires a starting location and a destination location.
- Axioms capture information surrounding the domain and detail the artifacts, processes, and relationships needed to communicate the concepts of the domain space.

Domain Model - Example

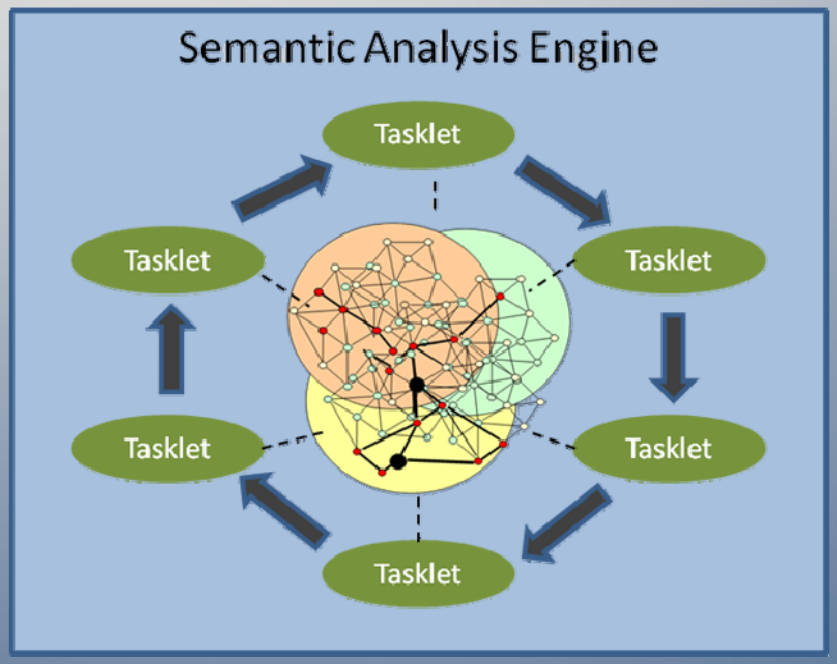


Semantic Analysis Engine (SAE)

- Processing engine that executes small execution units called Tasklets.
- The SAE continuously iterates through all Tasklets until a stopping point is reached.
- Each Tasklet represents an algorithm that operates on the incoming schemata and model. At the end of each Tasklet execution a numeric value is returned and plugged into an algorithm. Once all Tasklets are run the algorithm determines if another cycle of operation is required.

Semantic Analysis Engine

Syntax Parsing



The engine cycles through tasklets until a final state is reached and all viable associations are established.

asklets

- The engine uses word synonyms, hypernyms (is kind of), and meronyms (has part) in an effort to identify the closest meaning of a word to the context of its use.
- Capability matching is employed to isolate elements with similar characteristics but using terminology that may not have any direct associations.
- Comparisons are done against the Semantic Model in an effort to identify any direct correlations between the incoming data and the Semantic Model. If a direct correlation applies, strength is also added to supporting associations surrounding this correlation.
- OWL directives within the Semantic Model are applied to the incoming data to provide additional characteristics for analysis. (e.g. Are some data properties considered transitive?)

Web Ontology Language (OWL)

- The Domain Model leverages OWL's rich set of knowledge constructs – classes, properties, and property characteristics.
- One Tasklet places data within a temporary model and reasons across the data to allow additional domain rules to be applied. This Tasklet utilizes Semantic Web Rule Language (SWRL) to apply rules to the data.
- SWRL allows rules such as the transitive declaration:

$$\forall\chi\forall\gamma\forall\alpha(P(\chi,\gamma)\wedge P(\gamma,\alpha)\rightarrow P(\chi,\alpha))$$

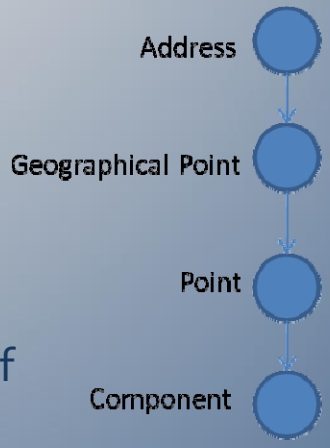
An example of which is:

$$\forall\chi\forall\gamma\forall\alpha(\text{hasAttribute}(\chi,\gamma)\wedge\text{hasAttribute}(\gamma,\alpha)\rightarrow\text{hasAttribute}(\chi,\alpha))$$

exical Analysis

WordNet is used to expand the variation of words.

- Synonyms are used as direct replacements for encountered words or phrases.
- Hypernyms provide a base meaning for a word. An example is Address, which has a hypernym of Geographical Point.



Stemming is used to find the root of a word by analyzing it and removing prefixes and suffixes that can alter it. Slowly can be stemmed to Slow.

apabilities

- Capabilities are an entities inherent trait or ability.
- All domain concepts have zero or more capabilities associated with them.
- Once an association between a WSDL element and a domain concept becomes strong enough, the capabilities associated with the domain concept are now associated with the WSDL element.
- Examples can be DiagnosticCapability and TherapeuticCapability. A doctor may have both while an MRI may only have the DiagnosticCapability.

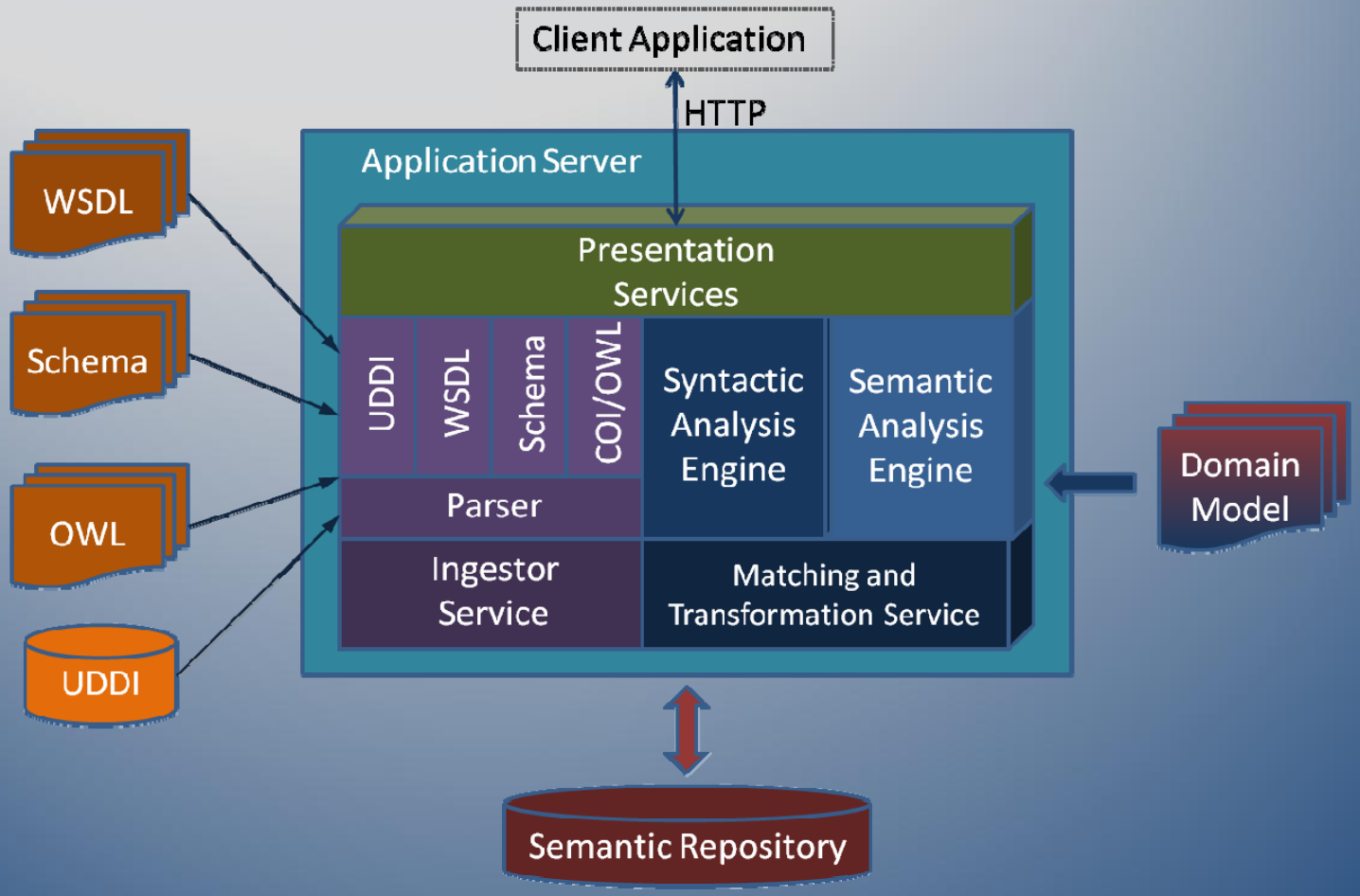
emantic Repository

- The Semantic Repository is a database of all processed semantic relationships.
- Created to alleviate long term cost of processing. As the Domain Model and input schema get larger, the processing grows dramatically. The storage of associations allows the user to later retrieve information without having to reprocess the data.

chema Comparison

- Once both schema have been analyzed against the Domain Model, they are then compared against each other using Capabilities as the basis of the comparison.
- By using Capabilities as the crux of the comparison, the SMTS is able to apply an additional level of reasoning to the comparison.
- Schema element are matched and ranked based on the strength of their matched Capabilities.

rchitecture



onclusion

- The manual application of schema matching and transformation is costly and time consuming.
- As more systems adopt the NCDS, the number of Web Services will also expand, compounding the cost and work associated with schema matching.
- The SMTS alleviates much of the mundane analysis required in schema matching and allows integrators to focus on a narrowed set of integration points.