Semantic Technologies and CDISC Standards

Frederik Malfait, Information Architect, IMOS Consulting
Scott Bahlavooni, Independent
Part I
Introduction to Semantic Technology
Directed graph of <Subject Predicate Object> triples

Subject: Bob
Predicate: livesIn
Object: Paris

Subject: Paris
Predicate: partOf
Object: France

Subject: Bob
Predicate: age
Object: 32
Ontologies

Fact Base

Bob \(\text{livesIn}\) Paris \(\text{partOf}\) France

Ontology

\(\text{Person}\) \(\text{owl:Class}\)
\(\text{City}\) \(\text{Location}\)
\(\text{Country}\)
In RDF everything is a triple built from resources and literals
- <Subject Predicate Object>
- <Subject Predicate Literal>

Every RDF resource has a Uniform Resource Identifier (URI)
- NCI Thesaurus, term Mild Adverse Event
  http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C41338

Namespaces provide a convenient way to group related resources together
- Define the namespace prefix nci
  xmlns:nci="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#"
- To obtain the Qualified Name (qname)
  nci:C41338
RDFS and OWL define an extensive vocabulary for building ontologies
• Example: `<hasCitizen owl:inverseOf livesIn>`

Meaning is defined by the way new triples are inferred from existing triples
• Stated Triples
  `<Bob livesIn Paris>`
  `<hasCitizen owl:inverseOf livesIn>`
• Inferred Triple
  `<Paris hasCitizen Bob>`

Inference engines provide reasoning capabilities for formal ontologies

Formal ontologies have a precise mathematical foundation based on first-order logic, set theory, and model theory
RDF Representation - Ontology Editors

.png
RDF Representation - Serializations

**RDF/XML**

```xml
<rdf:Description rdf:about="#Bob">
  <rdf:type rdf:resource="#Person"/>
  <age rdf:datatype="http://www.w3.org/2001/XMLSchema#positiveInteger">32</age>
  <livesIn rdf:resource="#Paris"/>
</rdf:Description>
```

**Turtle**

```
:Bob a :Person ;
:age "32"^^<http://www.w3.org/2001/XMLSchema#positiveInteger> ;
```

**N-Triple**

```
```
RDF Query Language - SPARQL

```
SELECT ?subject ?age
WHERE {
  ?subject :age ?age
}
```
RDF Query Language - SPARQL

```
SELECT ?raveForm ?formOrdinal ?formLabel
WHERE {
  ?raveForm rdf:type mms:DataCollectionForm .
  ?raveForm mms:extendsModelComponent/mms:subDataElementDomainOf daco:DED.Diabetes .
}
ORDER BY ?formOrdinal
```

<table>
<thead>
<tr>
<th>raveForm</th>
<th>formOrdinal</th>
<th>formLabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ext.Form.Diabetes.History</td>
<td>50201</td>
<td>Diabetes History</td>
</tr>
<tr>
<td>Ext.Form.Diabetes.HypoglycemicEvents</td>
<td>50202</td>
<td>Hypoglycemic Events</td>
</tr>
<tr>
<td>Ext.Form.Diabetes.HomeGlucometerReadings</td>
<td>50203</td>
<td>Home Glucometer Readings (for hypoglycemic events)</td>
</tr>
<tr>
<td>Ext.Form.Diabetes.AllergicReactionEvents</td>
<td>50204</td>
<td>Allergic Reaction Events</td>
</tr>
<tr>
<td>Ext.Form.Diabetes.PancreatitisEvents</td>
<td>50205</td>
<td>Pancreatitis Events</td>
</tr>
</tbody>
</table>
Resource Description Framework (RDF)
- RDF defines how to express information as a directed graph of resources
- Every resource has a URI and is part of a namespace

RDF Schema (RDFS) and Web Ontology Language (OWL)
- A set of standard predicates to build ontologies (vocabularies, schemas)
- Inference capabilities

SPARQL Protocol and RDF Query Language (SPARQL)
- Language to query a RDF data sets

Simple Knowledge Organization System (SKOS)
- Small RDF vocabulary for building concept models (thesauri, etc.)
Federation and Linked Data

- Semantic models in RDF format are easy to federate
- Federation of Data = union of triples from both graphs
- Linked Data = referencing between graphs

Federated Graph

http://example.org/people#

Bob

+ livesIn

  + Paris

    + age
      
        + 32

http://example.org/geo#

France

Paris

+ owl:sameAs

+ partOf
In addition, we provide localized versions of DBpedia in 111 languages. All these versions together describe 20.8 million things, out of which 10.5 million overlap (are interlinked) with concepts from the English DBpedia. The full DBpedia data set features labels and abstracts for 10.3 million unique things in up to 111 different languages; 8.0 million links to images and 24.4 million HTML links to external web pages; 27.2 million data links into external RDF data sets, 55.8 million links to Wikipedia categories, and 8.2 million YAGO categories. The dataset consists of 1.89 billion pieces of information (RDF triples) out of which 400 million were extracted from the English edition of Wikipedia, 1.46 billion were extracted from other language editions, and about 27 million are data links to external RDF data sets. The Datasets page provides more information about the overall structure of the dataset. Dataset Statistics provides detailed statistics about 22 of the 111 localized versions.
Linked Open Data (LOD) Cloud

Richard Cyganiak
Anja Jentzsch
http://lod-cloud.net/
Semantic Technology Application Areas

- **Semantic Web**
  - Move the web beyond a web of documents, add semantics
  - Large scale web data integration

- **Knowledge Management**
  - Vocabularies, taxonomies, thesauri, nomenclature, dictionaries e.g. SNOMED CT, NCI Thesaurus, HGNC, etc.
  - Domain ontologies, i.e. formal knowledge representation and reasoning capabilities for a certain domain of interest

- **Roche Data Standards Office**
  - Data standards and metadata driven workflow
  - Metadata registries, e.g. based on ISO 11179
  - Data integration and master data management
Part II
Metadata Management at Hoffmann-La Roche
Design and manage end-to-end clinical trial data standards
  • from protocol to submission
  • with supporting business processes

Drive adoption of current and future industry standards

Provide lifecycle management and governance for all data standards

Contribute to external standards bodies
  • CDISC
  • TransCelerate
  • Phuse/FDA
Roche Global Data Standards Repository

- Single point of access for end-to-end data standards
  - Protocol and Study Design
  - Data Collection, Data Tabulation, Data Analysis
  - Regulatory Submission

- Enable consistent metadata management
  - Operational and conceptual metadata
  - Versioning and life cycle management

- Follow industry standards when available and fit for purpose
  - CDISC standards
  - ISO 11179 Metadata Registry (MDR) standards
  - W3C semantic standards (RDF, OWL, SKOS)
GDSR Data Standards Content

- External content
  - SDTM 1.2, SDTMIG 3.1.2, and amendments
  - CDISC Controlled Terminology
  - NCI Thesaurus

- Integrated data standards, Roche and Genentech
  - Data collection: Safety and every Roche TA (2000 data elements)
  - Data tabulation: SDTM Sponsor Extensions

- Value level metadata
  - Lab measurements (2000)
  - Preferred units and unit conversions
  - Questionnaires (150)
GDSR ISO 11179 Part 6

- **Steward**
  - OrganizationalRole
    - Registrar
      - ReferenceDocument
        - documentIdentifier : string
        - documentLanguage : string
        - documentTitle : string
        - documentType : string
      - AdministeredItem
        - administrationRecord : AdministrationRecord
        - context : Context
        - priorVersion : AdministeredItem
        - reference : ReferenceDocument
        - supersedes : AdministeredItem
      - Context
    - AdministrationRecord
      - registrar : Registrar
      - registrationStatus : RegistrationStatus
      - steward : Steward
      - submitter : Submitter
      - administrativeNote : string
      - changeDescription : string
      - creatorDate : date
      - effectiveDate : date
      - explanatoryComment : string
      - lastChangeDate : date
      - unresolvedIssue : string
      - untilDate : date
      - version : positiveInteger
  - Submission
    - RegistrationStatus
      - LifecycleRegistrationStatus
GDSR System Components

- Content Management
- Content Publishing
- Metadata Repository
- Single Point of Access
Part III
Semantic Models for CDISC Standards
SDTMIG 3.1.2 Walk-Through
The Events class captures planned protocol milestones such as randomization and study completion, and occurrences, conditions, or incidents independent of planned study evaluations occurring during the trial (e.g., adverse events) or prior to the trial (e.g., medical history).
Resource Form

Annotations

Incoming References

- sdtmts:parentCodelist
  - sdtm:C66768.C17998
  - sdtm:C66768.C48275
  - sdtm:C66768.C49494
- sdtmgs:datasetColumn
  - Column.AE.AEOUT

Other Properties

- sdtmts:cdiscDefinition
  - A condition or event that is attributed to the adverse event and is the result or conclusion of the adverse event. (NCI)
- sdtmts:cdiscSubmissionValue
  - OUT
- sdtmts:cdiscSynonyms
  - Outcome of Event
- sdtmts:cdiscCodelistName
  - Outcome of Event
- sdtmts:cdiscIsExtensibleCodelist
  - false
- sdtmts:nciCode
  - C66768
- sdtmts:nciPreferredTerm
  - CDISC SDTM Adverse Event Outcome Terminology
- rdf:type
  - sdtmts:Codelist

URL: http://gdsr.roche.com/cdisc/sdtm-terminology#C66768
Resource Form

**URL:** http://gdsr.roche.com/cdisc/sdtm-terminology#C66768.C49494

- **Annotations**
- **Incoming References**

**Other Properties**

- **sdmtscdiscDefinition**
  - One of the possible results of an adverse event outcome that indicates that the event has not improved or recuperated. (NCI)

- **sdmtscdiscSubmissionValue**
  - NOT RECOVERED/NOT RESOLVED

- **sdmtscnciCode**
  - C49494

- **sdmtscnciPreferredTerm**
  - Not Recovered or Not Resolved

- **sdmtsparentCodelist**
  - sdtmt:C66768

- **rdf:type**
  - sdtmts:CodelistElement
A Permissible variable should be used in a domain as appropriate when collected or derived. Except where restricted by specific domain assumptions, any SDTM Timing and Identifier variables, and any Qualifier variables from the same general observation class are permissible for use in a domain based on that general observation class. The Sponsor can decide whether a Permissible variable should be included as a column when all values for that variable are null. The sponsor does not have the discretion to not submit permissible variables when they contain data.
Part IV
RDF Model Driven REST Services
True resource oriented service model using HTTP GET <URI> requests

GET response is model driven and fully configurable in RDF describing
  • which resource properties and relationships to return
  • how and where to retrieve the response data from RDF
  • how to format the response
  • how to optionally transform an XML response using XSLT and XSL-FO

Requests can ask for
  • data formats XML, JSON, CSV
  • schema formats XSD, Java JAXB ready classes

Model driven REST services are specified in an ontology based on the idea of a resource facet describing a logical view of a resource

Resources may have multiple facets and apply facet composition
GET <SDTM Variable URI>

```xml
<variable
    xmlns="http://gdsr.roche.com/ws/data-tabulation/facet.dt.variable.default"
    uri="http://gdsr.roche.com/pd-biometrics/sdtm-ext#ColumnExtension.VS.DOMAIN">
    <variableOrdinal>2</variableOrdinal>
    <name>DOMAIN</name>
    <label>Domain Abbreviation</label>
    <type>Char</type>
    <length>2</length>
    <controlledTermsOrFormat>VS</controlledTermsOrFormat>
    <role>Identifier Variable</role>
    <cdiscNotes>Two-character abbreviation for the domain.</cdiscNotes>
    <sponsorNotes>Set to 'VS'.</sponsorNotes>
    <compliance>Required Variable</compliance>
    <reference>SDTM 2.2.4, SDTMIG 4.1.2.2, SDTMIG Appendix C2</reference>
</variable>
```
GET &lt;SDTM Domain URI&gt;

```xml
<domain xmlns="http://gdsr.roche.com/ws/data-tabulation/facet.dt.domain.default"
        uri="http://gdsr.roche.com/cdisc/sdtmig-3-1-2#Table.VS"
        label="Vital Signs">
  <domainOrdinal>19</domainOrdinal>
  <observationClass>Findings Observation Class</observationClass>
  <datasetName>VS</datasetName>
  <domainCode>VS</domainCode>
  <domainLabel>Vital Signs</domainLabel>
  <variableList>
    <variable uri="http://gdsr.roche.com/pd-biometrics/sdtm-ext#ColumnExtension.VS.STUDYID">
      <variableOrdinal>1</variableOrdinal>
      <name>STUDYID</name>
      <label>Study Identifier</label>
      <type>Char</type>
      <length>8</length>
      <role>Identifier Variable</role>
      <cdiscNotes>Unique identifier for a study.</cdiscNotes>
      <sponsorNotes>Set to STUDYID in SDTM.DM domain.</sponsorNotes>
      <compliance>Required Variable</compliance>
      <reference>SDTM 2.2.4</reference>
    </variable>
    <variable uri="http://gdsr.roche.com/pd-biometrics/sdtm-ext#ColumnExtension.VS.DOMAIN">
      <variableOrdinal>2</variableOrdinal>
      <name>DOMAIN</name>
      <label>Domain Abbreviation</label>
      <type>Char</type>
      <length>2</length>
      <controlledTermsOrFormat>VS</controlledTermsOrFormat>
      <role>Identifier Variable</role>
      <cdiscNotes>Two-character abbreviation for the domain.</cdiscNotes>
    </variable>
  </variableList>
</domain>
```
GET <Measurement URI>

```xml
<measurement
    xmlns="http://gdsr.roche.com/ws/lab-metadata/facet.measurement.default"
    uri="http://gdsr.roche.com/pd-biometrics/lab-metadata#Analyte.ALBUMIN">
  <label>Albumin</label>
  <code>ALBUM</code>
  <name>ALBUM</name>
  <lbtestcd>ALB</lbtestcd>
  <lbtest>Albumin</lbtest>
  <lbspec>BLOOD</lbspec>
  <hasNoUnits>false</hasNoUnits>
  <SI-Unit>g/L</SI-Unit>
  <USCV-Unit>g/dL</USCV-Unit>
</measurement>
```
```xml
<measurementCategory
   xmlns="http://gdsr.roche.com/ws/lab-metadata/facet.measurementCategory.default"
   uri="http://gdsr.roche.com/pd-biometrics/lab-metadata#MeasurementCategory.Chemistry">
  <label>Chemistry</label>
  <domain>LB</domain>
  <lbcat>CHEMISTRY</lbcat>
  <measurementList>
    <measurement uri="http://gdsr.roche.com/pd-biometrics/lab-metadata#Analyte.A1AGLP"
                  xmlns="http://gdsr.roche.com/ws/lab-metadata/facet.measurementCategory.default">
      <label>Alpha-1 Acid Glycoprotein</label>
      <code>A1AGLP</code>
      <name>A1AGLP</name>
      <lbtestcd>A1AGLP</lbtestcd>
      <lbtest>Alpha-1 Acid Glycoprotein</lbtest>
      <lbspec>BLOOD</lbspec>
      <hasNoUnits>false</hasNoUnits>
      <SI-Unit>g/L</SI-Unit>
      <USCV-Unit>mg/ml</USCV-Unit>
      <sponsorNotes>AAG is a common synonym for this analyte.</sponsorNotes>
    </measurement>
    <measurement uri="http://gdsr.roche.com/pd-biometrics/lab-metadata#Analyte.A1_ANTITRYSIN_SER"
                  xmlns="http://gdsr.roche.com/ws/lab-metadata/facet.measurementCategory.default">
      <label>Alpha-1 Antitrypsin, Serum Conc</label>
      <code>A1ANTRYS</code>
      <name>A1ANTRYS</name>
      <lbtestcd>A1ANTRYP</lbtestcd>
      <lbtest>Alpha-1 Antitrypsin</lbtest>
      <lbspec>BLOOD</lbspec>
      <hasNoUnits>false</hasNoUnits>
      <SI-Unit>g/L</SI-Unit>
      <USCV-Unit>mg/dl</USCV-Unit>
      <sponsorNotes>LEGACY - DO NOT USE</sponsorNotes>
    </measurement>
  </measurementList>
</measurementCategory>
```
GET <Code List URI>

```xml
<stdmCodelist
    xmlns="http://gdsr.roche.com/ws/data-tabulation/facet.dt.stdmCodelist.default"
    url="http://gdsr.roche.com/cdisc/stdm-terminology#C101834">
  <code>C101834</code>
  <label>Normal Abnormal Response</label>
  <extensible>false</extensible>
  <cdiscSubmissionValue>NORMAL</cdiscSubmissionValue>
  <cdiscSynonyms>Normal Abnormal Response</cdiscSynonyms>
  <cdiscDefinition>
    A terminology codelist to describe a normal or abnormal response.
  </cdiscDefinition>
  <nciPreferredTerm>CDISC SDTM Normal Abnormal Response Terminology</nciPreferredTerm>
  <valueList>
    <value url="http://gdsr.roche.com/cdisc/stdm-terminology#C101834.C14165">
      <code>C14165</code>
      <submissionValue>NORMAL</submissionValue>
    </value>
    <value url="http://gdsr.roche.com/cdisc/stdm-terminology#C101834.C17998">
      <code>C17998</code>
      <submissionValue>UNKNOWN</submissionValue>
    </value>
    <value url="http://gdsr.roche.com/cdisc/stdm-terminology#C101834.C25401">
      <code>C25401</code>
      <submissionValue>ABNORMAL</submissionValue>
    </value>
    <value url="http://gdsr.roche.com/cdisc/stdm-terminology#C101834.C48658">
      <code>C48658</code>
      <submissionValue>INDETERMINATE</submissionValue>
    </value>
  </valueList>
</stdmCodelist>
```
REST Based Workflow Automation

- Representation of a protocol schedule of assessments
- Creation of a CRF view of data collection standards in PDF format
  - Operational View
  - Annotated CRF Submission View
- Creation of a Medidata Rave upload file (ALS)
- Creation of a non-CRF File Format Specification
- SDTM transformation specification and code generator
- Integration with Master Data and Document Management systems
Part V
Community Work
What we would like to see...

CDISC Ownership

URL: http://rdf.cdisc.org/sdtmig-3-1-2/std#Model.SDTMIG-3-1-2

- Annotions
- Other Properties
  - mms:contextDescription
    - The Study Data Tabulation Model implementation Guide is a CDISC defined guide for the implementation of SDTM providing a detailed specification of the SDTM domains.
  - mms:contextLabel
    - Study Data Tabulation Model Implementation Guide (SDTMIG) Version 3.1.2
  - mms:contextName
    - sdtmig-3-1-2
  - rdf:type
    - mms:Model
- Incoming References
  - mms:context
    - EventsObservationClass
    - FindingsAbout
    - FindingsObservation
    - InterventionsObservationClass
    - RelationshipDataset
    - SpecialPurposeDomain
    - TrialDesignModel

Form | Browser | Graph | Source Code
Collaborations

- **cdisc2rdf**
  - Precompetitive collaboration to publish CDISC Standards in RDF
  - SDTM 1.2, IG 3.1.2, CT published January 2013
  - Discontinued and folded into FDA/PhUSE CSS work

- **National Cancer Institute**
  - Publishing CDISC Controlled Terminology in RDF since June 2013

- **FDA/Phuse CSS 2013-2014**
  - Semantic Technology project within the Emerging Technologies WG

- **CDISC and TransCelerate**
  - Contribute to the SHARE architecture and modeling effort
  - Define requirements for RDF Export of SHARE content
Phase I - April to August 2013
- Represent existing CDISC standards in RDF
  - CDASH 1.1
  - SDTM 1.2, SDTMIG 3.1.2, SDTM 1.3, SDTM IG 3.1.3
  - SEND IG 3.0
  - ADaM 2.1, ADaMIG 1.0
- Including domain assumptions
- Integrated with Controlled Terminology

Consolidation - September 2013

Phase II - October 2013 to February 2014
- Protocol Representation, Schedule of Assessments, Trial Design
- Conformance Checks
- Represent clinical trial data in RDF