ORACLE

Oracle Database 11g Semantics Technical Talk

Oracle Semantic Technologies Agenda

- Introduction
- Oracle Database 11g Semantic capabilities
  - Load
  - Inference
  - Query
  - Jena Adaptor
  - Support for W3C standards
  - Performance benchmark
  - Partners

ORACLE

Oracle Database Semantic Data Store

- A feature of Oracle Spatial 11g Option for Oracle Database 11g Enterprise Edition
  - Requires Partitioning and Advanced Compression Options
- An open and <u>persisted</u> RDF data model and <u>analysis platform</u> for semantic applications
- An RDF Data Model with inferencing (RDFS, OWL and user-defined rules)
- Performs SQL-based access to triples and inferred data
- Combines SQL query of relational data with RDF graphs and ontologies
- Supports large graphs (billion+ triples)
- Easily extensible by 3rd party tools/apps

ORACLE

Only Oracle Database 11g…

Has an open, persistent, analytic semantic data management platform

- Scalability – Trillions of triples
- Availability – tens of thousands of users
- Security – protect sensitive business data
- Performance – timely load, query & inference
- Accessibility – to enterprise applications
- Manageability – leverage IT resources
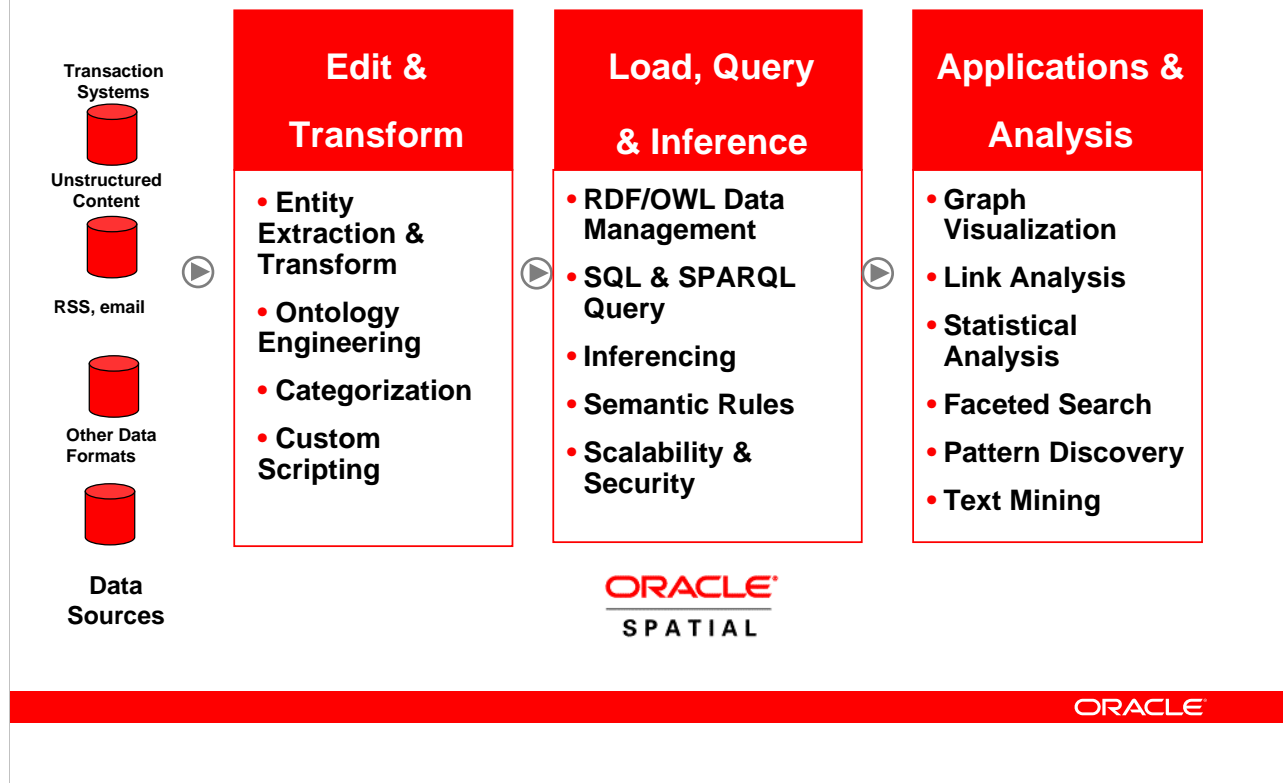
= Oracle Database Strengths

ORACLE

You may be evaluating different technologies necessary to develop a semantic solution. These advantages are available using Oracle Database. They are not available using in-memory stores or specialty RDF stores.

Oracle Customer Examples

- Enterprise Information Integration
  - Hutchinson 3G Austria
- Large Public Dataset for Data Integration
  - Uniprot dataset at the Swiss Institute of Bioinformatics
- Data Integration
  - Yale University
  - Stanford University
  - University of Cincinnati
- Bio-surveillance
  - University of Texas at Houston
- Re-use of Legacy Data
  - Pharmaceutical companies

ORACLE

Semantic Data Management Workflow

Data Sources

- Transaction Systems
- Unstructured Content
- RSS, email
- Other Data Formats

Edit & Transform

- Entity Extraction & Transform
- Ontology Engineering
- Categorization
- Custom Scripting

Load, Query & Inference

- RDF/OWL Data Management
- SQL & SPARQL Query
- Inferencing
- Semantic Rules
- Scalability & Security

Applications & Analysis

- Graph Visualization
- Link Analysis
- Statistical Analysis
- Faceted Search
- Pattern Discovery
- Text Mining

ORACLE SPATIAL

ORACLE

This is a canonical semantics workflow. Data is coming from structured and unstructured sources. A common ontology is used in order to define patterns and relationships across them to relate concepts, terms and map across schemas.

- ETL entities from unstructured & structured data sources OS, CMS, Web, databases
- Categorize using an ontology
- load into RDF as triples, infer new relationships using the ontology and query
- Apply search analytics and decision making tools

Oracle Semantic Technologies in Oracle Spatial provide the technology to load, inference and query. Oracle Partners provide tools for functions in the 1st & 3rd columns as well as a UI for Oracle Semantic Technologies.

Oracle Spatial 11g Semantic Capabilities
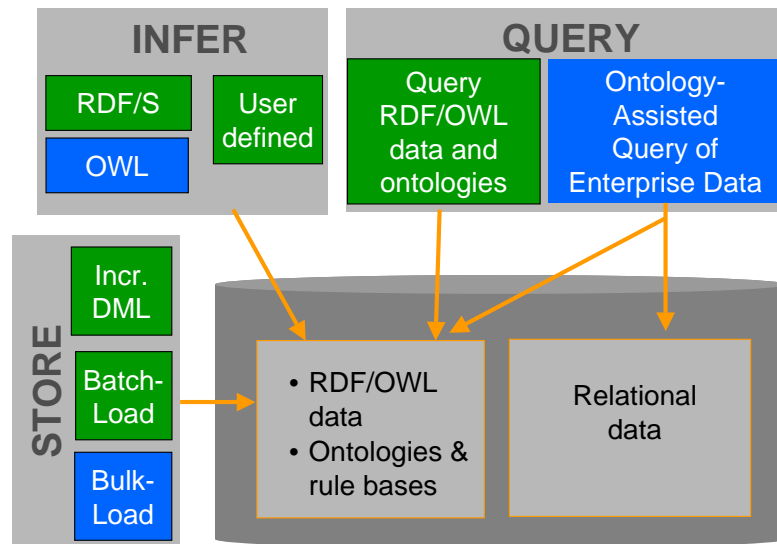
Oracle Spatial 11g provides an open, persistent, analytic semantic data management platform with scaleable storage, persistent inference and robust Semantic and SQL query capability.

•Storage model, loading, and management for data represented in RDF/OWL

•SQL-based query of RDF/OWL data

•Ontology-assisted query of Relational data

•Native inference engine to infer new relationships from RDF/OWL data

Semantic Data Management Tasks

- Create a Semantic Network

- Create a RDF/OWL model

- Load RDF/OWL data

- Optionally create user-defined rulebases

- Infer new RDF/OWL data

- Query RDF/OWL data

- Perform Ontology-assisted Query

ORACLE

The major activities associated with building / querying a Semantic Store in Oracle Database 11g include the following…
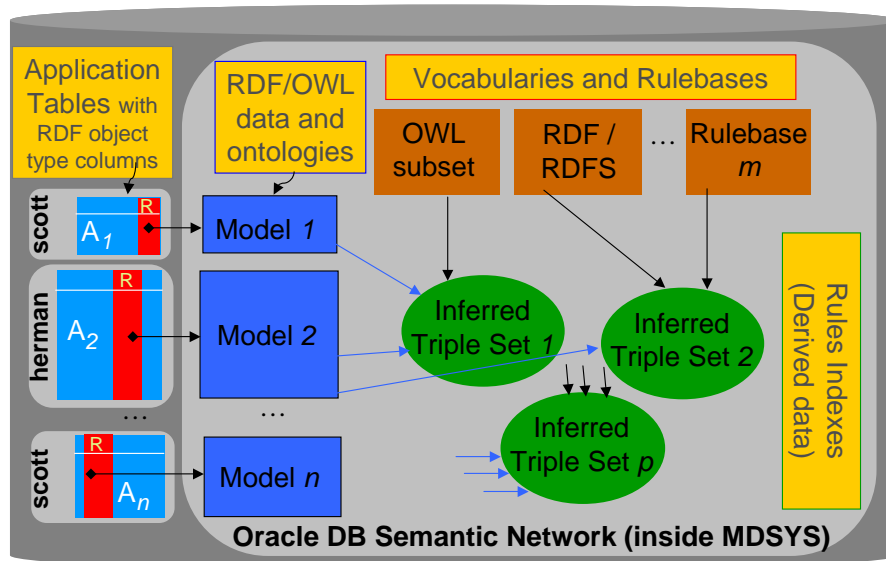
Creating a Semantic Network and Model

- Creating a Semantic Network
  - Generates all RDF% tables and views under user MDSYS

- Creating a RDF/OWL model
  - Create a table with a column of type SDO_RDF_TRIPLE_S
  - Optionally choose a tablespace for the model

ORACLE

9

Entities in the Oracle Semantic Network

Application Tables with RDF object type columns

scott — A$_1$ (R)

herman — A$_2$ (R)

...

scott — A$_n$ (R)

RDF/OWL data and ontologies

Model *1*

Model *2*

...

Model *n*

Vocabularies and Rulebases

OWL subset

RDF / RDFS

... Rulebase *m*

Inferred Triple Set *1*

Inferred Triple Set *2*

Inferred Triple Set *p*

Rules Indexes (Derived data)

Oracle DB Semantic Network (inside MDSYS)

ORACLE

Access Control on a Semantic Network

- Models
  - A view owned by MDSYS is created
  - Creator has SELECT privilege with GRANT option
  - DML on an RDF object type column requires privileges on the associated application table
- Rulebases
  - A view owned by MDSYS is created
  - The creator has SELECT & DML privilege w/ GRANT
- Rules Indexes for Inferred Triple Sets
  - A view owned by MDSYS is created
  - Creator must have SELECT privilege on underlying model and rulebase views
  - The creator has SELECT privilege with GRANT

ORACLE

Loading RDF/OWL Data

- Oracle Database is a scalable native graph data store

    - Oracle Database 11g stores up to 8 exabytes

- Semantic data stored optimally in relational tables

- Load Options: Bulk, Batch, and DML INSERT

- Single management environment for all your data

ORACLE

Oracle Spatial provides a native semantic data store in Oracle Database based on the W3C RDF standard to store semantic models.

Testing done to date on over a billion triples.

RDF Triples

Hand_Fracture — :subClassOf → Arm_Fracture
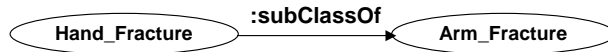
- RDF triple represents a *directed* *labeled* edge

<http://purl.uniprot.org/uniprot/Q4U9M9>

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

<http://purl.uniprot.org/core/Protein>

- RDF triple *set* represents a directed labeled graph

```
# uniprot:... rdf:type :Protein
<http://purl.uniprot.org/uniprot/Q4U9M9> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.uniprot.org/core/Protein> .
<http://purl.uniprot.org/uniprot/P15711> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.uniprot.org/core/Protein> .
<http://purl.uniprot.org/uniprot/Q43495> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.uniprot.org/core/Protein> .
<http://purl.uniprot.org/uniprot/P18646> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.uniprot.org/core/Protein> .
<http://purl.uniprot.org/uniprot/P13813> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.uniprot.org/core/Protein> .
# other attributes
<http://purl.uniprot.org/uniprot/Q4U9M9> <http://www.w3.org/2000/01/rdf-schema#seeAlso> <http://purl.uniprot.org/interpro/IPR007480> .
```

ORACLE

A graph is a collection of triples that Oracle refers to as a model.

Alternatives for Loading RDF/OWL Data

- Bulk-load
  - Fastest, but skips triples with literals > 4k bytes
  - Load data into a staging table (using SQL*Loader from a file or Named Pipe containing N-Triple formatted data)
  - Invoke PL/SQL API to do bulk load from the staging table

- Batch-load
  - Fast and can handle long literals
  - Invoke Java-based API to load from file containing N-Triple formatted data

- DML INSERT on RDF/OWL data
  - INSERT on app table (for loading small amounts of data)
  - UPDATE of app table's SDO_RDF_TRIPLE_S col
  - DELETE of row(s) from app table

ORACLE

Storage Capabilities

- Value point equivalence (via Canonicalization)
- Set property
- Graph level access control
- Fidelity (preserving user-specified form of data)
- Ancillary information for each triple
- Strict parse with error reporting (optional for bulk-load)
- Hash collision detection and resolution
- Long literal values > 4000 bytes (Not for bulk-load)

ORACLE

Inferencing RDF Data

- Native inference engine in Oracle Database for
  - RDF, RDFS, OWL subset, and user-defined rules
  - Rules are stored in rulebases in Oracle Database
- New relationships (triples) are entailed (inferred) from an RDF/OWL graph by applying rules to a model
  - Uses Forward Chaining
  - Minimizes on-the-fly computation for faster queries
- Automatic identification of new relationships (triples)

  Ex: hand_fracture :subClassOf arm_fracture,
      arm_fracture   :subClassOf upper_extremity_fracture
      => hand_fracture :subClassOf upper_extremity_fracture

ORACLE

• Oracle Database is the first commercial relational database to offer native inference capability.

• It uses a forward-chaining mechanism that infers new relationships from the existing model and store them persistently in the database.

• Persistent storage means relationships can be precomputed and inferenced at a convenient time for later querying. This is an important capability for large triple stores.

Vocabularies: RDFS and OWL

- RDFS (RDF Schema)
  - Structuring of resources and properties
    - rdfs:class → Class of resources
    - rdfs:subClassOf → hierarchy of classes
    - rdfs:subPropertyOf → hierarchy of properties

- OWL (Web Ontology Language)
  - Builds on RDFS …
    - Property Characteristics: transitivity, symmetry, functional, inverse functional, inverse
    - Class construction via set operations and property restrictions
  - Separate layers have been defined balancing *expressibility vs. implementability*: OWL Lite, OWL DL, OWL Full

ORACLE

OWL Subsets Supported

Three subsets to meet most needs*

- RDFS++
  - RDFS plus owl:sameAs and owl:InverseFunctionalProperty
- OWLSIF
  - Based on Dr. Horst's pD* vocabulary[1]
- OWLPrime
  - owl:TransitiveProperty, SymmetricProperty, FunctionalProperty, InverseFunctionalProperty
  - owl:inverseOf
  - sameAs, differentFrom
  - owl:disjointWith, complementOf
  - owl:hasValue, allValuesFrom, someValuesFrom
  - owl:equivalentClass, equivalentProperty

* Jointly determined with domain experts, customers and partners

1 Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary

11g OWL Inference PL/SQL API

- **SEM_APIS.CREATE_ENTAILMENT**(
  - Index_name
  - sem_models('GraphTBox', 'GraphABox', …),
  - sem_rulebases('OWLPrime'),
  - passes,
  - Inf_components,
  - Options
)
  - Use "PROOF=T" to generate inference proof

Typical Usage:
- First load RDF/OWL data
- Call create_entailment to generate inferred graph
- Query both original graph and inferred data

Inferred graph contains only new triples! Saves time & resources

- **SEM_APIS.VALIDATE_ENTAILMENT**(
  - sem_models(('GraphTBox', 'GraphABox', …)
  - sem_rulebases('OWLPrime'),
  - Criteria,
  - Max_conflicts,
  - Options
)

Typical Usage:
- First load RDF/OWL data
- Call create_entailment to generate inferred graph
- Call validate_entailment to find inconsistencies

Advanced Options

Give users more control over inference process

- Selective inference (component based)
  - Allows more focused inference.
  - E.g. give me only the subClassOf hierarchy.

- Set number of passes
  - Normally, inference continue till no further new triples found
  - Users can set the number of inference passes to see if what they are interested has already been inferred
  - E.g. I want to know whether this person has more than 10 friends

- Set tablespaces used, parallel index build

- Change statistics collection scheme

Adding User-Defined Rules

- Supporting semantics beyond OWLPrime
- User-defined rules (filter is supported)

Antecedents		Consequents
?x :parentOf ?y . ?z :brotherOf ?x .	➜	?z :uncleOf ?y

- E.g. to support core semantics of owl:intersectionOf

```
<owl:Class rdf:ID="FemaleAstronaut">
 <rdfs:label>chair</rdfs:label>
 <owl:intersectionOf rdf:parseType="Collection">
   <owl:Class rdf:about="#Female" />
   <owl:Class rdf:about="#Astronaut" />
 </owl:intersectionOf>
</owl:Class>
```

1. ➜ :FemaleAstronaut rdfs:subClassOf
         :Female
2. ➜ :FemaleAstronaut rdfs:subClassOf
         :Astronaut
3. ?x rdf:type :Female .
    ?x rdf:type :Astronaut . ➜
            x rdf:type :FemaleAstronaut

Integrating with complete DL reasoners

- TBox tends to be small in size
  - Generate a class subsumption tree using complete DL reasoners (like Pellet, KAON2, Fact++, Racer, etc)
- ABox can be arbitrarily large
  - Use Oracle OWL to infer new knowledge based on the class subsumption tree from TBox
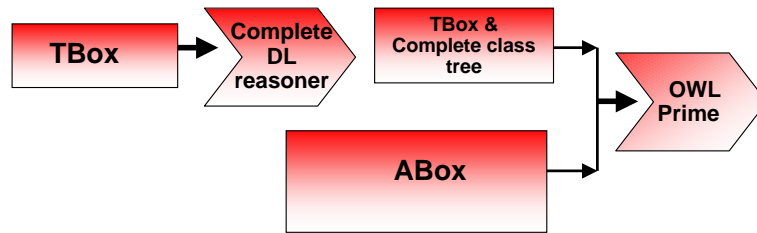
Querying RDF/OWL Data

- Choice of SQL or SPARQL
- SPARQL-like graph queries embedded in SQL
  - Enhance SPARQL queries w/ SQL processing such as join, order by, group by, filter conditions
  - Combine queries on semantic, relational and unstructured data
  - Ex: "find me all fractures related to upper_extremity_fracture that occurred in patients between ages 5 and 10"
- Jena plug-in for Oracle Database includes the full SPARQL API
- Oracle planning native support for SPARQL

ORACLE

Oracle provides The ability to:

•access semantic data through SQL and SPARQL

•do relational and graph queries in the same SQL statement - The graph model isn't treated as a standalone application – it is stored with and can be queried in combination with other data in Oracle Database.

•Use SPARQL -like capability via the Jena plug-in so application developers and partners can build applications on top of Jena and query Oracle Database with SPARQL.

SPARQL-like capability is not full SPARQL because the standard wasn't finalized at the time of Oracle Database 11g release. SPARQL support in the database is planned for a future release.

Querying RDF/OWL Data

- Matches RDF/OWL graph patterns with stored data
- Returns a table of results
- Supports SQL operators/functions to process results
- Supports Table Function Rewrite
- No staging when combined with relational queries

```
SELECT ...
FROM ...,
TABLE (SEM_MATCH invocation
) t, ...
WHERE ...
```

```
SEM_MATCH (
    '(?x :partOf :NorthAmerica)',        -- pattern: all persons
    SEM_Models('gmap'),                  --  RDF/OWL models
    SEM_Rulebases('OWLPRIME'),           -- rulebases
    SEM_Aliases(…)                       -- aliases
    null                                 -- no filter condition
)
```

ORACLE

Table Columns returned by SEM_MATCH

Each returned row contains one (or more) of the following columns for each variable ?x in the graph-pattern:

Column Name	Type	Description
x	varchar2	Value matched with ?x
x$rdfVTYP	varchar2	Value TYPe: URI, Literal, or Blank Node
x$rdfLTYP	varchar2	Literal TYPe: e.g., xsd:integer
x$rdfCLOB	CLOB	CLOB value matched with ?x
x$rdfLANG	varchar2	LANGuage tag: e.g., "en-us"

Projection Optimization: Only the columns referred to by the containing query are returned.
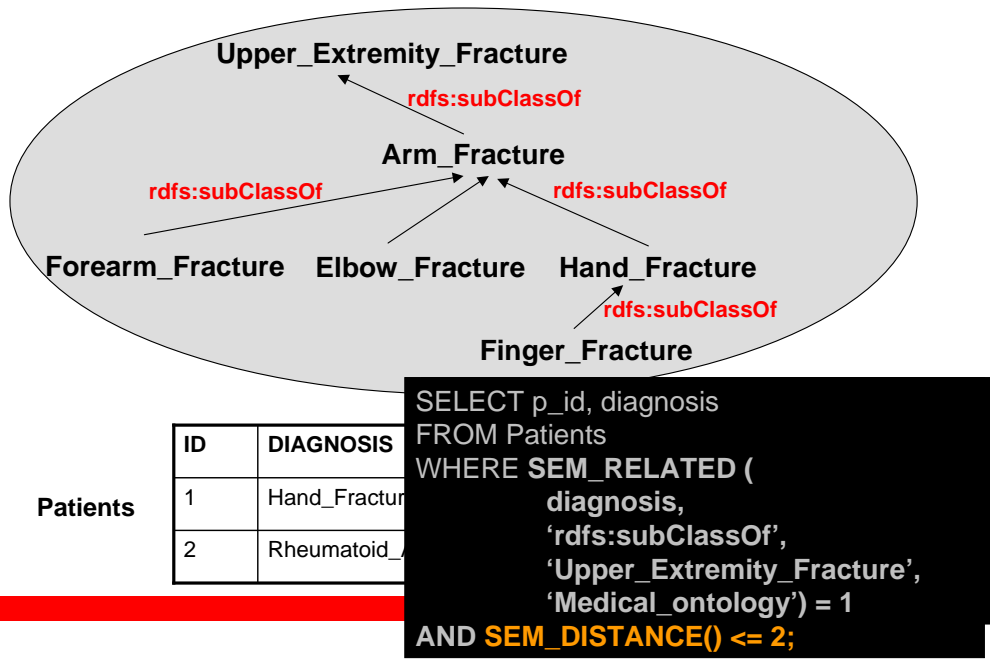
ORACLE

Ontology-Assisted SQL Query

- Traditionally the relationship between two terms is only checked in a <u>syntactic</u> manner

- Need a way to check <u>semantic</u> relationships by consulting an ontology

- Introduces two operators

  - SEM_RELATED (<col>,<pred>, <ontologyTerm>,
    <ontologyName> [,<invoc_id>])
  - SEM_DISTANCE (<invoc_id>) ← <u>Ancillary Oper.</u>

ORACLE

26

Example: Query using Semantic Operators

Upper_Extremity_Fracture

rdfs:subClassOf

Arm_Fracture

rdfs:subClassOf   **rdfs:subClassOf**

Forearm_Fracture   **Elbow_Fracture**   **Hand_Fracture**

rdfs:subClassOf

Finger_Fracture

Patients

ID	DIAGNOSIS
1	Hand_Fractur
2	Rheumatoid_A

```
SELECT p_id, diagnosis
FROM Patients
WHERE SEM_RELATED (
         diagnosis,
         'rdfs:subClassOf',
         'Upper_Extremity_Fracture',
         'Medical_ontology') = 1
AND SEM_DISTANCE() <= 2;
```

This is an example of inference, querying on upper extremity fracture returns hand fractures even though there is no direct relationship.

Jena Adaptor for Oracle Database

- Provides SPARQL access to Oracle Database
- Integrates 3ʳᵈ party tools w/ Oracle Database
- Facilitates integration with external reasoners
  - e.g. PelletInfGraph can be created on top of GraphOracleSem
- Supports Bulk, Batch, DML data load with long literals
- Implements Jena Graph/Model/BulkUpdateHandler APIs
- Scaleable and integrated
  - Data not cached in memory
  - SPARQL queries processed as SQL in Oracle Database
  - A SPARQL query with only conjunctive patterns is a single SEM_MATCH query
- Download from OTN, supported by Oracle

ORACLE

Oracle Database provides query access through SQL and SPARQL (SPARQL Protocol and RDF Query Language).

SPARQL queries to Oracle Database today are supported through the Jena Adaptor for Oracle Database that can be downloaded from OTN.

Oracle is planning to provide full SPARQL support in a future release.

Oracle Jena Adaptor Example

```
String queryString =
    " PREFIX foaf: <http://xmlns.com/foaf/0.1/> " +
    "SELECT ?person WHERE { ?person foaf:firstName \"Julie\" .
 } ";

QueryExecution qexec = QueryExecutionFactory.create(
    QueryFactory.create(queryString), modelOracleSem);
```

ORACLE

Lehigh University benchmark (LUBM)

- Facilitates evaluation of semantic data store products

- LUBM 8000 benchmark used

- 1.106 billion triples = ~262gb (1.068 billion w/o duplicates)

- Required 156gb of storage (~157 bytes per triple)
  - including storage for the application table and the indexes

- This represents a 40% reduction in storage!!

  - Oracle Advanced Compression Option used

ORACLE

LUBM 8000 Settings

- Hardware
  - CPU → Single-CPU P4 (3.0GHz with Hyper Threading)
  - Memory → 4GB
  - Hard Disks →Two 500GB 7200rpm SATA 3.0G
- OS: Red Hat Enterprise Linux (32-bit)
- DBMS
  - Oracle Database Enterprise Server 11g Release 1
  - Settings
    - db_block_size=8192
    - pga_aggregate_target=2000M
    - sga_target=1800M
    - Db_file_multiblock_read_count=128
    - Filesystemio_options='SETALL'
 - Temp tablespace was allocated on a separate hard disk

ORACLE

LUBM 8000 Performance Summary

- Bulk-Load
  - 1.1 billion triples (LUBM8000)
    - Time to load staging table: 2 ½ to 11 ½ hrs
    - Time using Bulk-load API: about 31 hrs
    - Storage: data 41 GB, indexes 94 GB, app table 22 GB
- Inference using OWLPrime
  - 1.068 billion triples (LUBM8000 minus the duplicate triples)
    - Inferred triples: 521.7 million
    - Time: 56.7 hrs
- Query on entailed data
  - 133 million triples (LUBM1000 minus dups) plus 60 million inferred
    - Most queries < 5 sec, rest 1 to 5 min, and one took longer

ORACLE

W3C Semantic Technology Stack

http://www.w3.org/2007/03/layerCake.svg
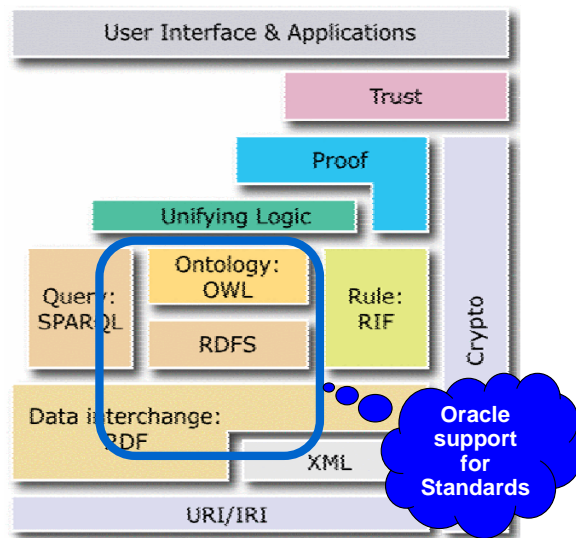
Commitment to W3C Semantic Standards

- Our implementation entirely based on W3C standards (RDF, RDFS, OWL)
  - SPARQL support through Jena
- Members of following W3C Web Semantic Activities:
  - W3C Data Access Working Group (DAWG)
  - W3C OWL Working group
  - W3C Semantic Web Education & Outreach (SWEO)
  - W3C Health Care & Life Sciences Interest Group (HCLS)
  - W3C Multimedia Semantics Incubator group
  - W3C Semantic Web Rules Language (SWRL)

ORACLE 34

Oracle Spatial semantics implementation conforms with W3C standards for storage, schema and rules.

Oracle Partners In Action

- TopBraid Suite - ontology management and visualization

- 360° Solutions - semantic platform for data integration & faceted search

- OntoBroker - high order inference and reasoning

ORACLE

Oracle partners add value to Oracle Database semantic infrastructure technology with market leading tools and applications:

Summary

Semantic Technology support in the database

- Store RDF/OWL data and ontologies
- Inference new RDF/OWL triples via native inferencing
- Query RDF/OWL data and ontologies
- Ontology-Assisted Query of relational data

Whitepapers, documentation, sample code, downloads:

   oracle.com/technology/tech/semantic_technologies

ORACLE

Oracle has the leading and only commercial semantic relational database in the industry with…

•Native Storage of RDF and OWL

•Native Inference using W3C standards

•Query of semantic data using SQL extensions and SPARQL

•And an innovative Ontology-Assisted Query of relational data

For More Information

search.oracle.com

`semantic technologies` 🔍

or

oracle.com

ORACLE