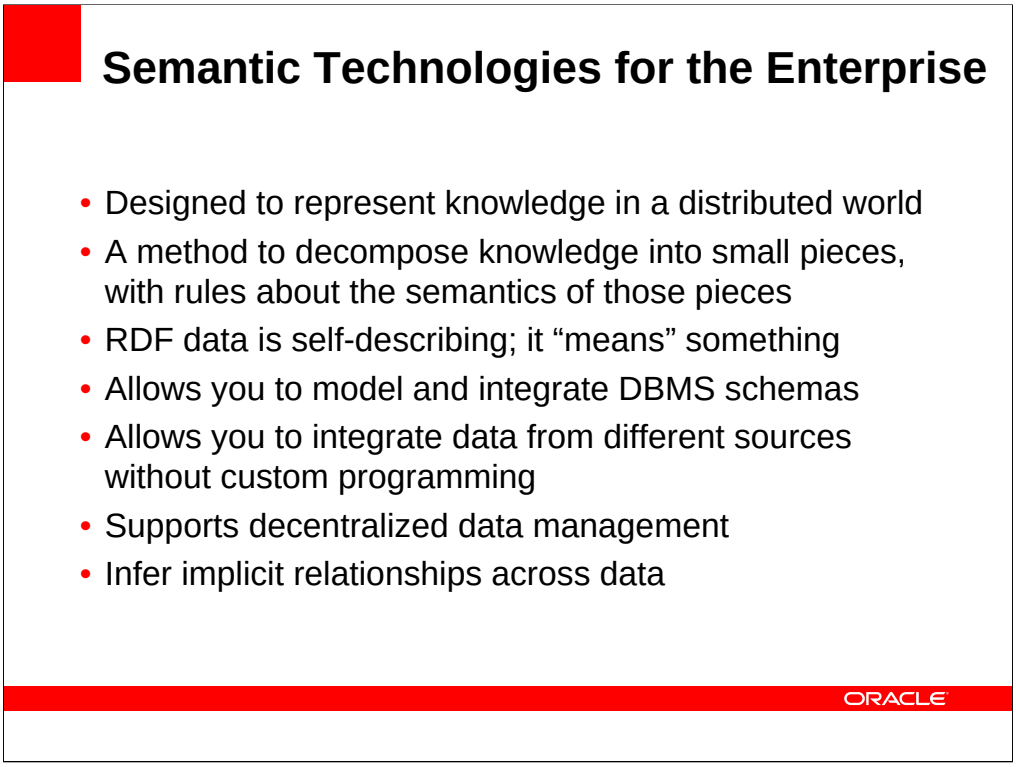
ORACLE

Oracle Database Semantic Technologies Overview

Oracle Semantic Technologies Agenda

- Semantic technologies for the enterprise
- Why use Oracle Database as a semantic data store
- Customer examples
- Features Overview

ORACLE



This presentation on Oracle Spatial 11g semantic technologies assumes some knowledge of basic principles of semantics.

Semantic technologies provide a metadata repository to access other data

•Semantics abstracts unstructured content by extracting meaning from entities in underlying data and structuring it so it can be queried in a meaningful way

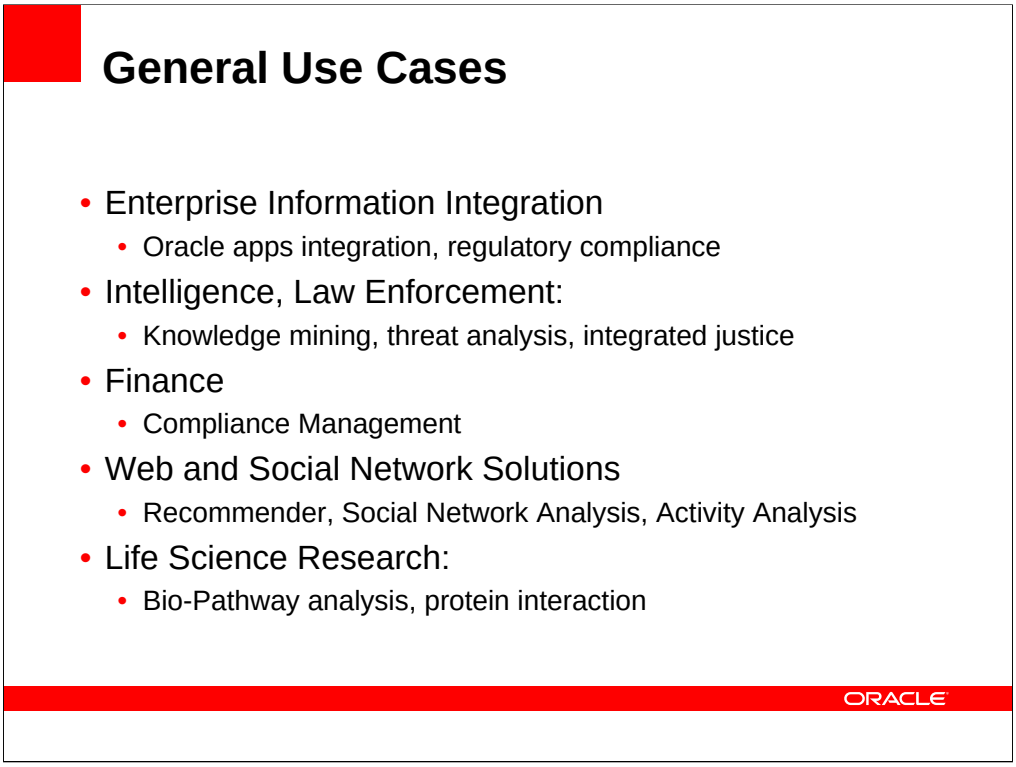
•RDF (Resource Description Framework) is the standard for encoding this metadata into a flexible triples data model (subject- object –predicate)

•RDF can also be used to model and integrate relational data

•So RDF can be used to describe relationships across a variety of data sources w/o custom programming

•RDF allows you to manage metadata models centrally w/o bringing all of the data into one place

•RDF Schema & OWL takes an RDF store to the next level with the ability to create inferences or new knowledge from RDF models using a range of W3C RDF & OWL rules. Oracle Database expands this capability with support user-defined rules.



Use cases – there are a whole range of applications for semantics

•Enterprise integration across different applications such as CRM, BPM, CMS – perhaps querying to create a regulatory compliance report

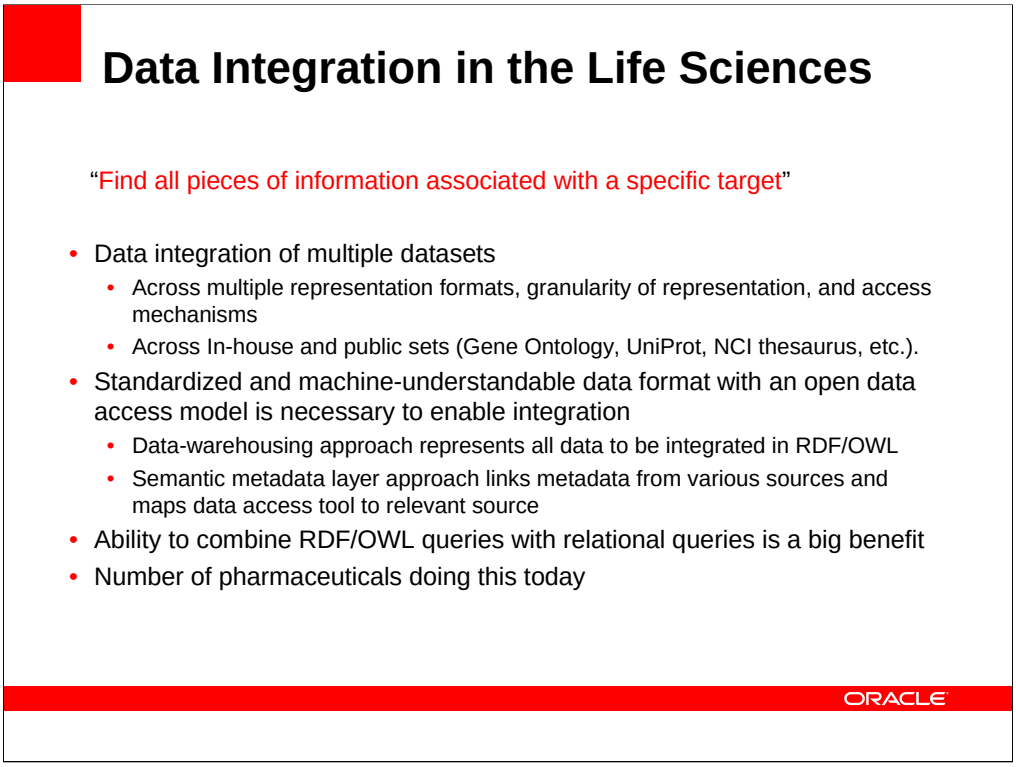
•Governance/compliance/risk – inference and audit tracking, enforce segregation of duties e.g., check's approver can't also write it.

•Intelligence and law enforcement – large social networks –find relationships between people and places and events - such as searching DMV, RMV, police, FBI to create a consolidated report and find out whether there are any pending warrants on an individual before granting bail – Metatomix does this

•Finance – similar to integrated justice – used by large banks for compliance, governance and risk management

•Social networks – create relationship of similarity between people or books

•LS research – doing 100's of thousands of experiments – difficult to find interactions and associations with historical data



Let's talk about the importance of data integration in Life Science (LS) and how LS communities are using semantic technologies for data integration.

A common question in LS is "Show me everything about a certain drug target". This type of query necessitates integrating a wide range of public and commercial data sources in different formats, which would be hard to do if you didn't have a consistent data format, data access model, query language and way of defining concepts and relationships across different types of data sources.

The W3C standards for semantics provide a metadata repository that enables this kind of integration with:

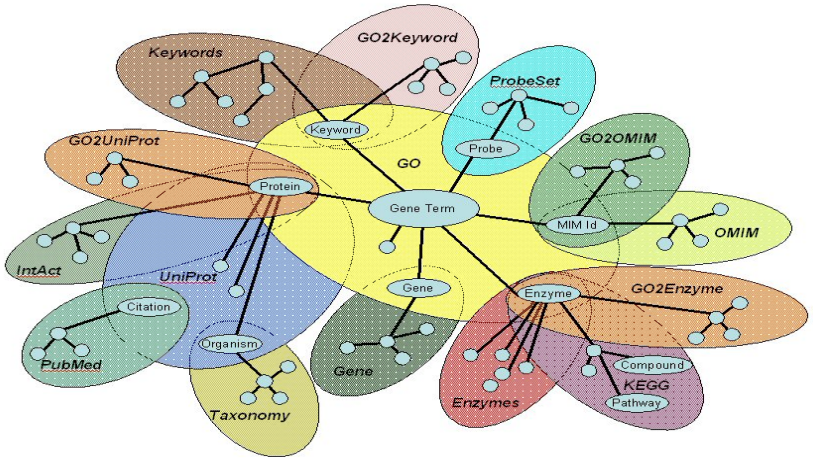
•A standard way to store semantic metadata using the machine-understandable subject-object-relationship triple data format, and the RDF open graph model

•A standard way to define concepts and inference new relationships in an ontology, and query across different models using RSFS, OWL and SPARQL

Oracle Database is uniquely suited to provide data management for semantic metadata repositories with scaleable storage, persistent inference and a rich query capability that includes the ability to make mixed semantic and SQL queries.

Oracle partners add value on Oracle Database with market leading tools and applications such as TopQuadrant's TopBraid Suite for ontology management and visualization, Metatomix's Semantic Platform for data integration and faceted search, and Ontoprise's OntoBroker for high order inference or reasoning.

Use Case: Integrated Bioinformatics Data



ORACLE

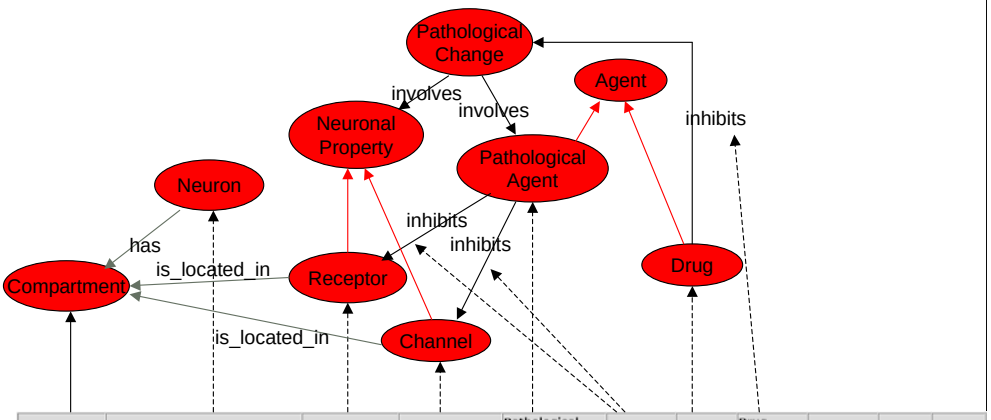
Source: Siderean Software

This is an example of integrating many graph models from public domain and commercial proprietary sources

Objective is to query and find patterns across these models that might exist in a drug discovery environment

An ontology maps concepts in models to concepts in individual databases

Relational to Ontological Mapping



Compartment	Cell: NeuronDB	Receptor	Channel	Pathological Agent (PA)	PA Action	Drug	Drug Action	Stage	Note	Detail
Soma	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66750
Dendrite	CA1 pyramidal neuron		I A	beta Amyloid	Inhibits			Early	View	66240
	Olfactory bulb mitral cell	GabaA						Early	View	66750
Unspecified	Oocyte		I L high threshold	beta Amyloid	Inhibits			Early	View	66252
								Early	View	66753
	CA1 pyramidal neuron			beta Amyloid	Inhibits			Early	View	66758
	CA1 pyramidal neuron	NMDA	I Calcium	beta Amyloid	Inhibits		Inhibits		View	66250

Courtesy: SenseLab, Yale University

This illustrates relational to ontological mapping of experiment results in the form of spreadsheets to predefined ontology

•Researchers can query the model using SPARQL and SQL to search for experiment results and find relationships between tests that may have been done at different times or from different groups

•This is a way to integrate across spreadsheets

•TopQuadrant TopBraid suite is very good at building and managing and updating ontologies with Oracle Database

•Note: Red lines show class – subclass relationships, black lines show properties

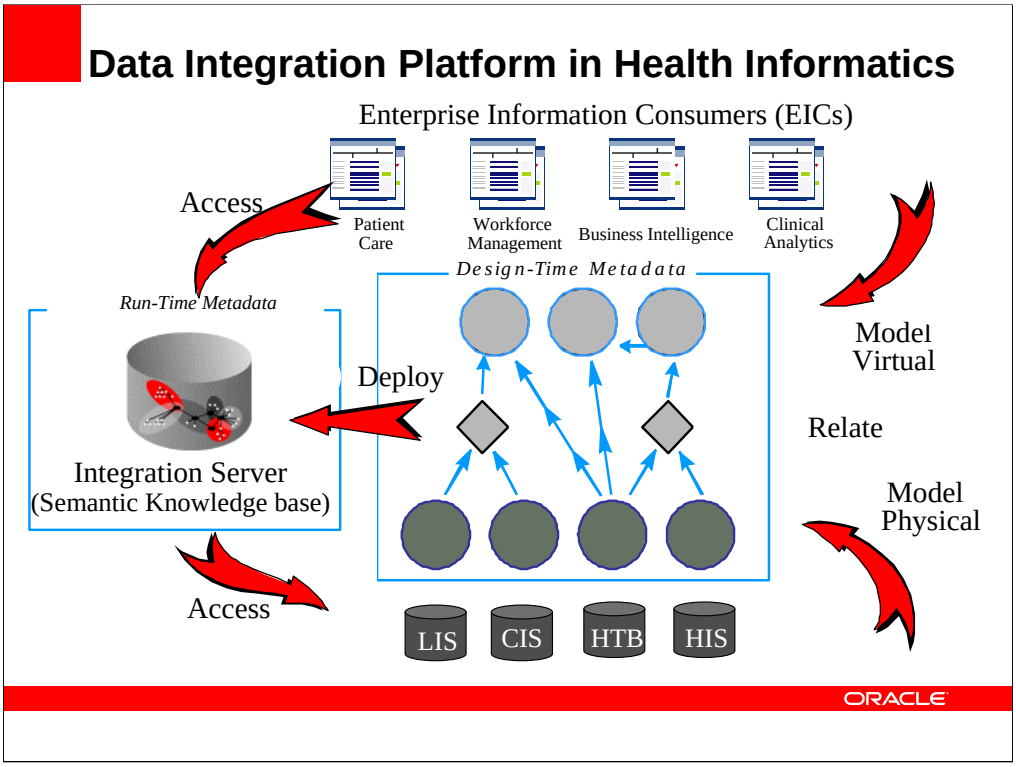
Biosurveillance

- Biosurveillance application: Track patterns in health data
- Data from 8 emergency rooms in Houston at 10 minute intervals
- Data converted into RDF/OWL and loaded into the database
- 8 months data is 600M+ triples
- Automated analysis of data to track patterns:
  - Spike in flu-like symptoms (RDF/OWL inferencing to identify a flu-like symptom)
  - Spike in children under age 5 coming in

ORACLE

A Biosurveillance application looks for events out of the ordinary of serious consequence such as epidemics. The University of Texas at Houston has an application of this type in production. The application takes dynamic event data from 8 hospitals in the form of RSS feeds (Web feed formats used to publish frequently updated content) and near realtime sensors. Event data is correlated with static reference data and historical data from databases to understand whether events are significant such as flu.

RDF/OWL are used to identify patterns that may not be obvious to a human analyst because of the many different indicators.

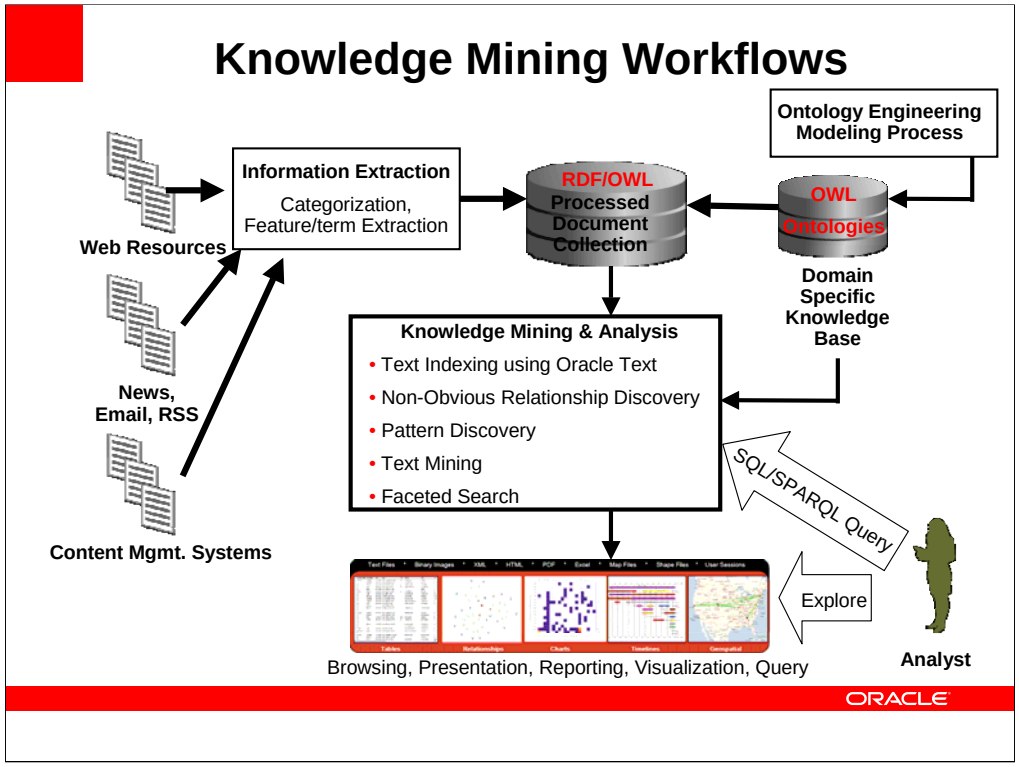


In a hospital, patient information consumers across the top want to get results that are in different data sources from different departments across the bottom

The integration server provides a metadata layer in the form of an integrated ontology derived from the schemas in the source databases, and manages it as a semantic knowledge base

Oracle Database 11g plays the role of metadata server storing the RDF model and persisting the inferences that can get quite large and evolve over time.

The metadata server insulates the applications across the top from the underlying physical structure of data sources so they don't need to write PL/SQL to access all these data sources or revise source code when schema is extended or a new database is added.



The objective of the knowledge mining workflow is to provide meaning and better intelligence on the large reams of data coming into the system from unstructured and possibly structured data sources. Metadata stored in RDF format and OWL ontologies are used to provide a meaningful way to categorize the underlying RDF data.

TopQuadrant TopBraid is a market leading tool for to defining ontologies and add new artifacts to evolve the ontologies. There are a range of tools for ETL, reporting and visualization of graph data and integrating with the geospatial and temporal data.

Only Oracle Database…

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 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 persisted RDF data model and analysis platform 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

Oracle provides an open, persistent, analytic semantic data management platform.

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

Enterprise Information Integration

- Hutchinson 3G Application
  - Comprehensive Service Availability monitoring solution
  - For control of individual access to services and individual services options
- Why RDF
  - Suited for explicit representation of relations amongst data from various domains
  - Suited for providing different views for different users of data, Ex: Service availability perspective, Statistics perspective, etc.
- Why Oracle Semantics Technology
  - Embedded in database technology
  - Versioning and schema support
  - Programming language interfaces like PL/SQL
  - Could use in-house expertise of DBAs and database developers

ORACLE

Large Public Dataset for Data Integration

- Uniprot dataset
  - Protein dataset, links several datasets together
  - Very large – 1 billion triples
- Why RDF
  - Life Sciences datasets are cross-referenced with each other and with third-party data sets
  - Data model is not simple and is evolving
  - Without RDF data was sub-optimally used
- Why Oracle
  - Enabled quick, low budget deployment
  - Handled complex queries that went beyond simple lookups based on protein names, ex: mapping gene names to diseases (required traversal of multiple links)

ORACLE

Data Integration

- University of Cincinnati Children's Hospital
  - Systematic approach to discover novel relationships between drugs and diseases
  - Knowledge sets using pharmacological substance, drug target, pathway, disease and clinical features had to be linked
- Why RDF
  - Ability to link data across datasets
  - Ability to identify embedded associations in the genome-phenome-pharmacome network
- Why Oracle
  - Commercial-grade software

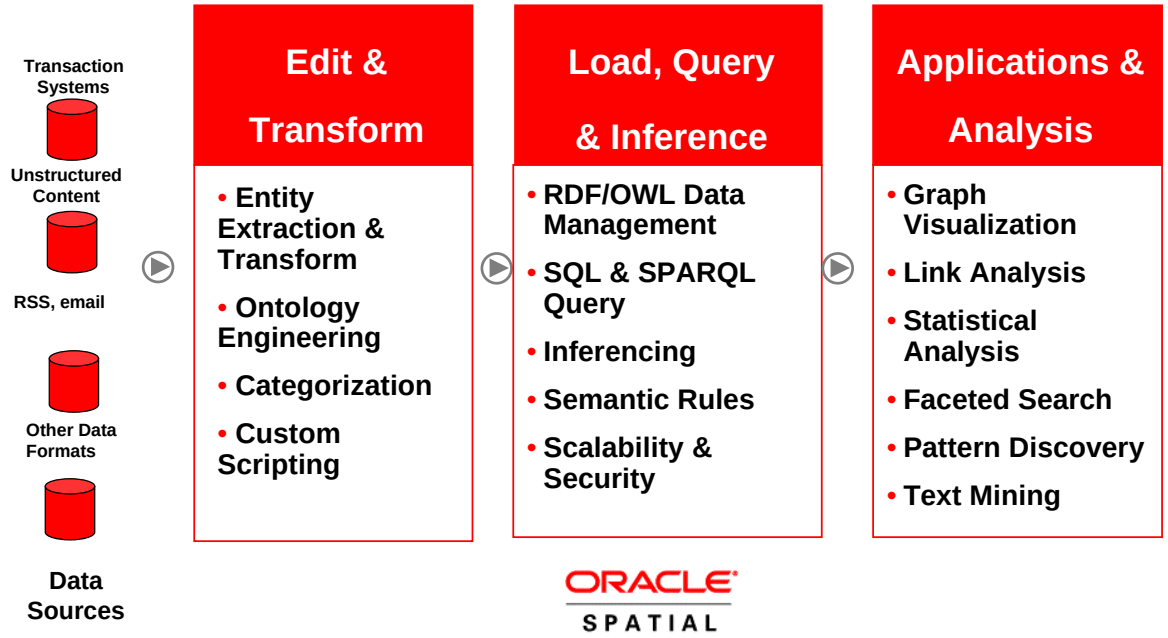
ORACLE

Re-use of Legacy Data

- Internal Compound Re-purposing at Pfizer
  - Save time and cost by re-using internal compounds and associated research
  - Challenge: how can compounds be identified across different databases and tools?
- Why RDF
  - Could store and represent any type of data
  - Ontology used to model the data could be easily modified as new data came in and underlying Science changed
  - Enabled rapid response to changing customer needs
- Why Oracle
  - Ability to combine relational queries with semantic queries: key facts were exported as RDF triples but primary sources of data were retained as relational and XML
  - Could use in-house expertise of DBAs and database developers

ORACLE

Semantic Data Management Workflow



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.

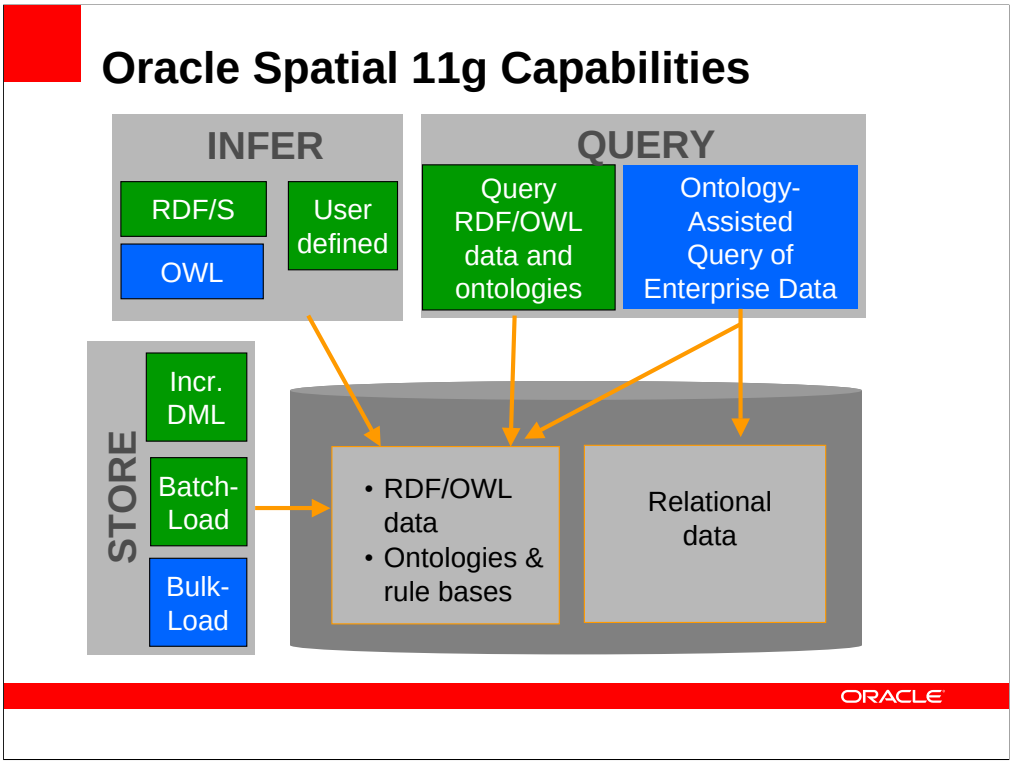
•ETL entities from unstructured & structured data sources OS, CMS, Web, databases

•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 provides an open, persistent, analytic semantic data management platform with scaleable storage, persistent inference and robust Semantic and SQL query capability:

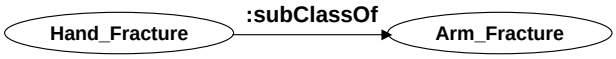
•Storage model, loading, and management for data represented in RDF/OWL

•SQL-based query of RDF/OWL data

•Ontology-assisted query of Relational data

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

Store Semantic Data



- Scalable native graph data store in Oracle Database
  - Oracle Database 11g stores up to 8 exabytes
- Semantic data stored optimally in relational tables
- Load Options: Bulk, Batch, and SQL INSERT
- Single management environment for all your data

ORACLE

The graph is composed of triples (walk through the example triple) Collections of triples comprise a model.

Oracle Spatial provides a native semantic data store in Oracle Database based on the W3C RDF standard to store semantic models. Testing to date scales beyond a billion triples. 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.

Inference RDF Data

- Native inferencing in the database for
  - RDF, RDFS, OWL subset
  - User-defined rules
- New relationships/triples are inferred and stored ahead of query time
  - Forward Chaining
  - Minimizes on-the-fly computation and results in fast query times
- 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 has the unique ability to do mixed queries: relational & graph queries in the same SQL statement. It uses a forward-chaining mechanism that infers new relationships from the existing model and store them persistently in the database. This is valuable for large triple stores of 100's of millions and larger. Persistent storage means relationships can be precomputed and inferenced at a convenient time for later querying. Walking through the simple example of inference, querying on upper extremity fracture returns hand fractures even though there is no direct relationship.

Query Semantic Data

- Choice of SQL or SPARQL
- SPARQL-like graph queries can be embedded in SQL
  - Key advantage – semantic queries can be combined with relational data
  - Ex: find me all fractures related to upper_extremity_fracture that occurred in patients between ages 5 and 10
- New Jena plug-in for Oracle can be used, includes a full SPARQL API
- Oracle plans to natively support SPARQL

ORACLE

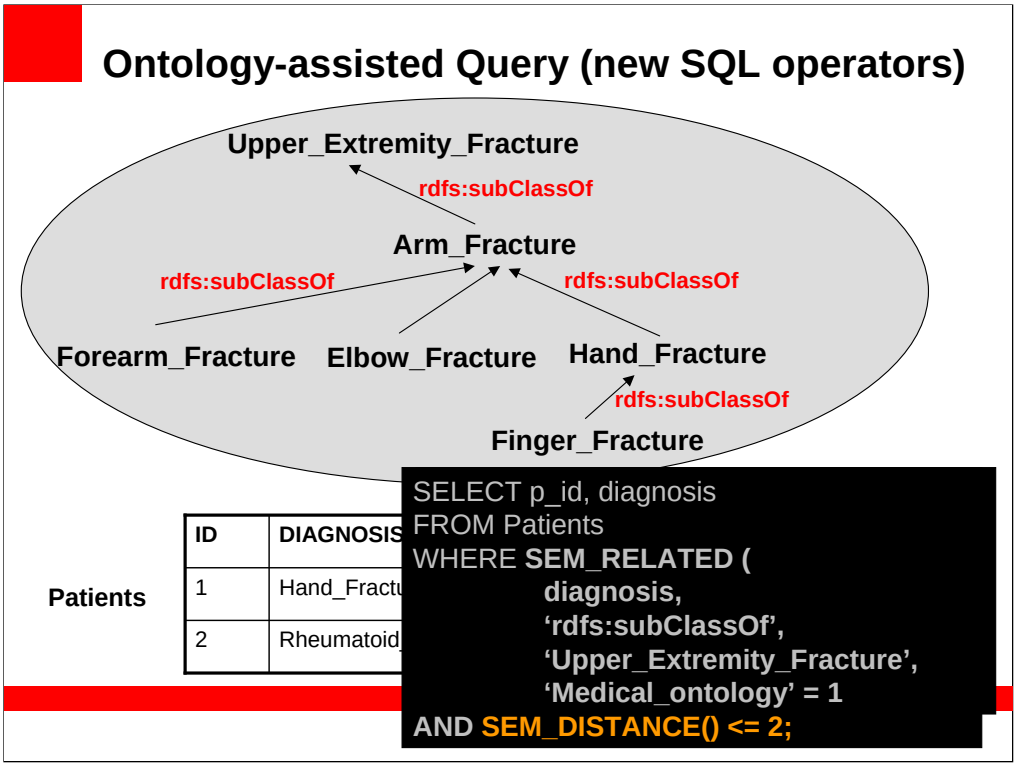
Oracle provides The ability to:

•access semantic data through SQL and SPARQL

•do mixed queries – relational and graph queries in the same SQL statement

•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 the next major release.



Build slide: key point: mixed semantic / relational queries return more semantically complete results

- No diagnosis in table = Upper_Extremity_Fracture
- Using SEM_RELATED returns rows with Arm_Fractures only
- Using SEM_Distance returns row with Hand_Fracture

Example: Semantics Enhanced Search

"Find me all DICOM images that contain the term 'Jaw'"

- Images might be annotated using multiple sets of terms
  - Might be annotated with 'mandible' but query with 'Jaw' should find the image
- Semantic relationships enhance keyword only search

ORACLE

You can also integrate RDF/OWL data with multimedia content stored and indexed in the database using Oracle Multimedia.

In this case Oracle Multimedia stores DICOM images. The name of an image is a node in the triple. The name of the image can then link to the image in the database. The advantage of storing the image title as an RDF node, is that that you can begin to perform queries such as "find me all DICOM images that contain the term 'jaw'". This query retrieves all images that relate to jaw or a part of a jaw.

You can also create a user-defined rule to state that a 'mandible' is SameAs 'jaw'. This would allow you to retrieve images that had been annotated with either phrase. You can also use the user-defined rules to state what the various components of a 'jaw' are, so that you can retrieve all images that have been annotated as a component of a 'jaw'. It is hugely important to be able to query images in this way, as it helps ensure that you are retrieving all relevant images.

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

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

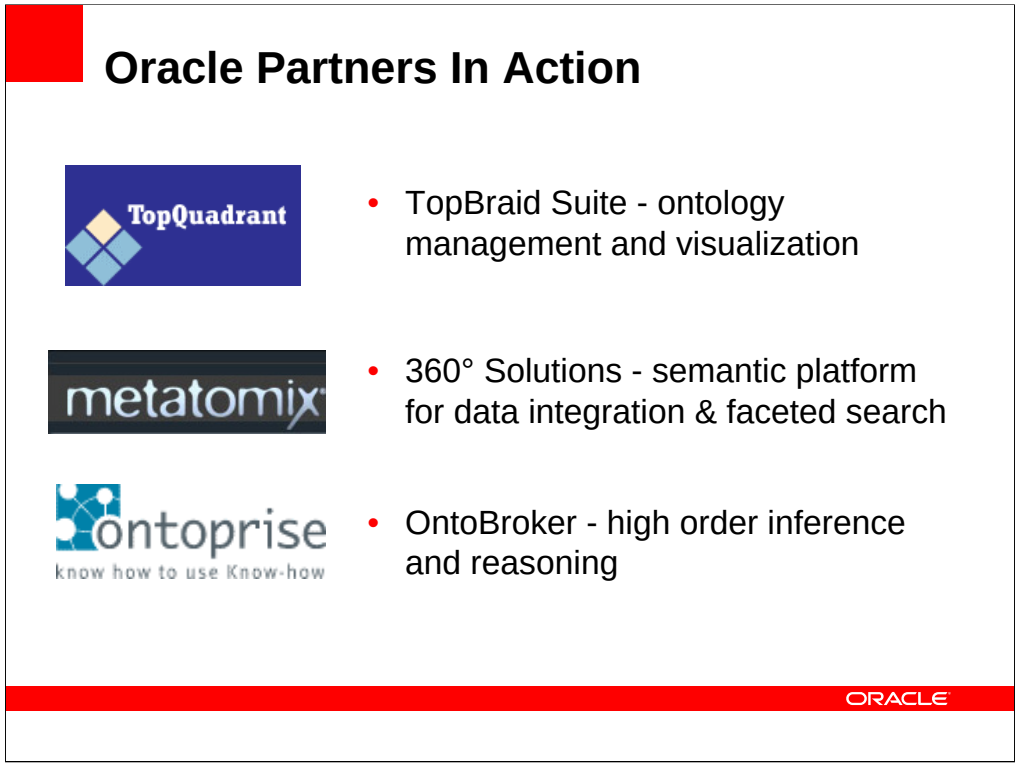
Commitment to W3C Semantic Standards

- 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 Semantic Web Education & Outreach (SWEO)
  - W3C Health Care & Life Sciences Interest Group (HCLS)
  - W3C Multimedia Semantics Incubator group
  - W3C OWL 1.1 Working group
  - W3C Semantic Web Rules Language (SWRL)

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

Oracle Database provides query access through SQL and SPARQL (SPARQL Protocol and RDF Query Language). SPARQL queries to Oracle Database today are supported through Jena and the Jena Adaptor Oracle provides on OTN.

Oracle is planning to provide full SPARQL support in Oracle Spatial in the next major release.



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

or

oracle.com

ORACLE