Information Integration Intelligence with Semantic Technology

Ontolog Forum 2008-01-24

http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall_2008_01_24

Holger Knublauch

holger@topquadrant.com
http://www.topquadrant.com

About Myself

- Computer Scientist (PhD, 2002)
- 2003-05 Post-Doc at Stanford
  - Lead developer of Protégé-OWL
- 2006-now TopQuadrant, Inc.
  - VP, Product Development
  - Lead developer of TopBraid Suite

http://www.knublauch.com

TopQuadrant™

About TopQuadrant

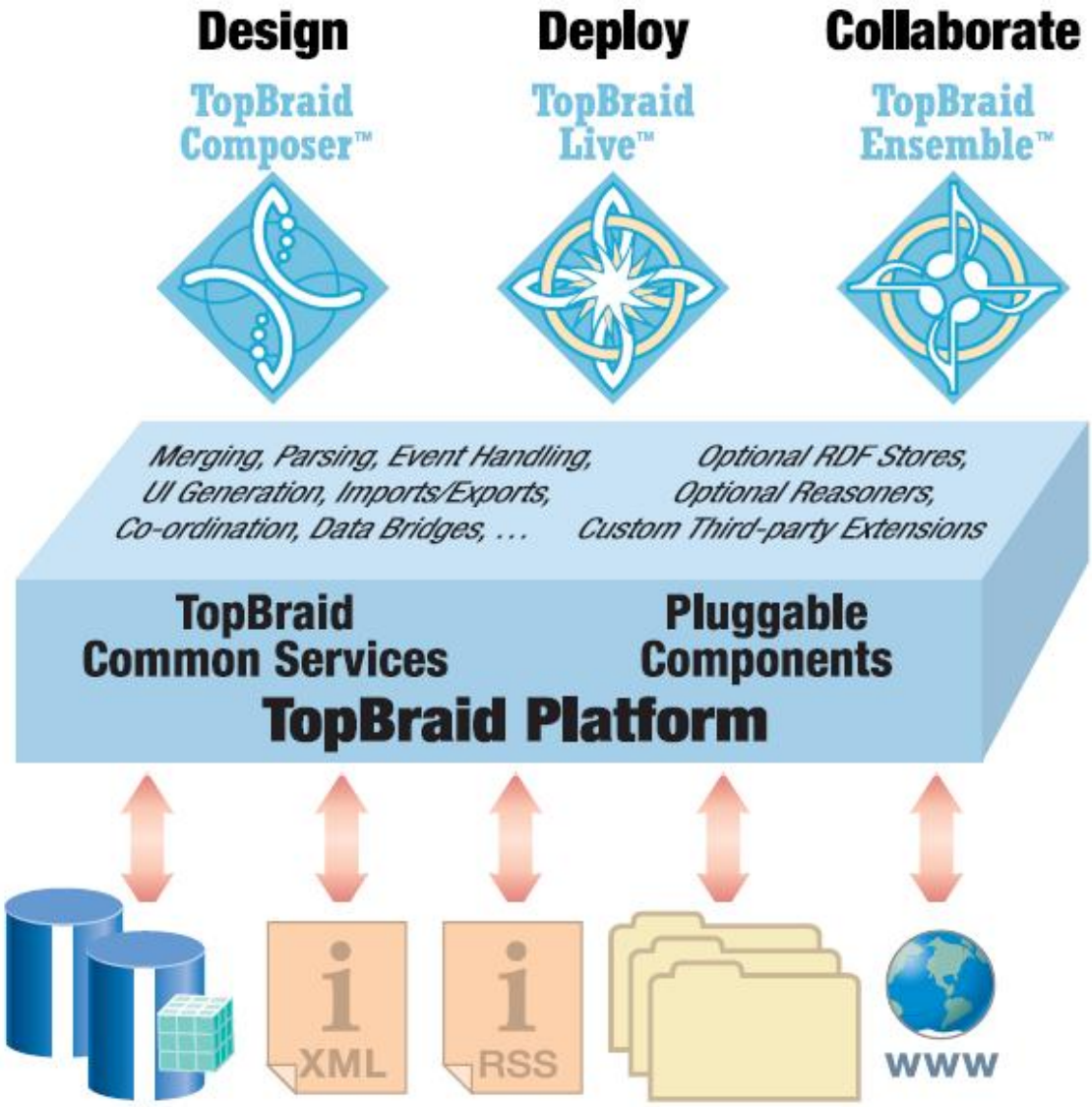
Headquarter: Alexandria, VA

Office: Mountain View, CA

Also: TopQuadrant Korea

TopQuadrant™

TopBraid Suite



TopQuadrant™

Information Integration Intelligence

- Heterogeneous data and schemas
  - Databases
  - Spreadsheet files
  - XML files
  - Newsfeeds
  - Online resources (HTML, GRDDL, RDFa)
  - Web services and data endpoints
- How to get integrated views to support business intelligence?

TopQuadrant™

Why Semantic Technology

- Class and property definitions (RDFS)
- Open architecture (URIs, triples, etc)
- Designed for linking (sameAs etc)
- Schema reuse (subClassOf etc)
- Explicit definitions of “semantics” (DL)
- Self-describing data (generic tools, discovery, schema evolution)
- Cross-schema querying (SPARQL)

TopQuadrant™

Semantic Technology Examples (1)

- Major retailer with an established name in Housewares, Lawn and Garden, Automotive and other products. Can we give our shoppers an integrated way to deal with warrantees, service records, proofs of purchase, etc. for all our product lines? An "Orbitz of Housewares"
- But they have hundreds of product lines, and new ones every day. How can they do this on this scale?
- TopBraid Semantic Technologies provides seamless integration of many and varied product lines
- Customers come to this retailer instead of competitor to get integrated support of new appliances with old

TopQuadrant™

Semantic Technology Examples (2)

- Consumer Electronics. Marketing and distributing information about products. Consumer electronics is notorious for new product categories with new features (game boxes? entertainment centers? HDTV? DVR?) and compatibility dependencies.

- How do we present our customer base with a seamless integrated picture of all possible products and how they combine, in the face of such a large set of product lines, with changing requirements?

- TopBraid / Semantic Technologies provides a flexible, extensible way to manage multiple products seamlessly

TopQuadrant™

Semantic Technology Examples (3)

- Health care solution built by CTG. Health care providers as well as patients require a seamless, integrated view of all health care information and services:
  - tests
  - available drugs
  - insurance information
  - clinic availability, etc
- Information is available for these things, but cannot be managed in a single seamless way.
- CTG is using TopBraid to create a seamless health care dashboard.

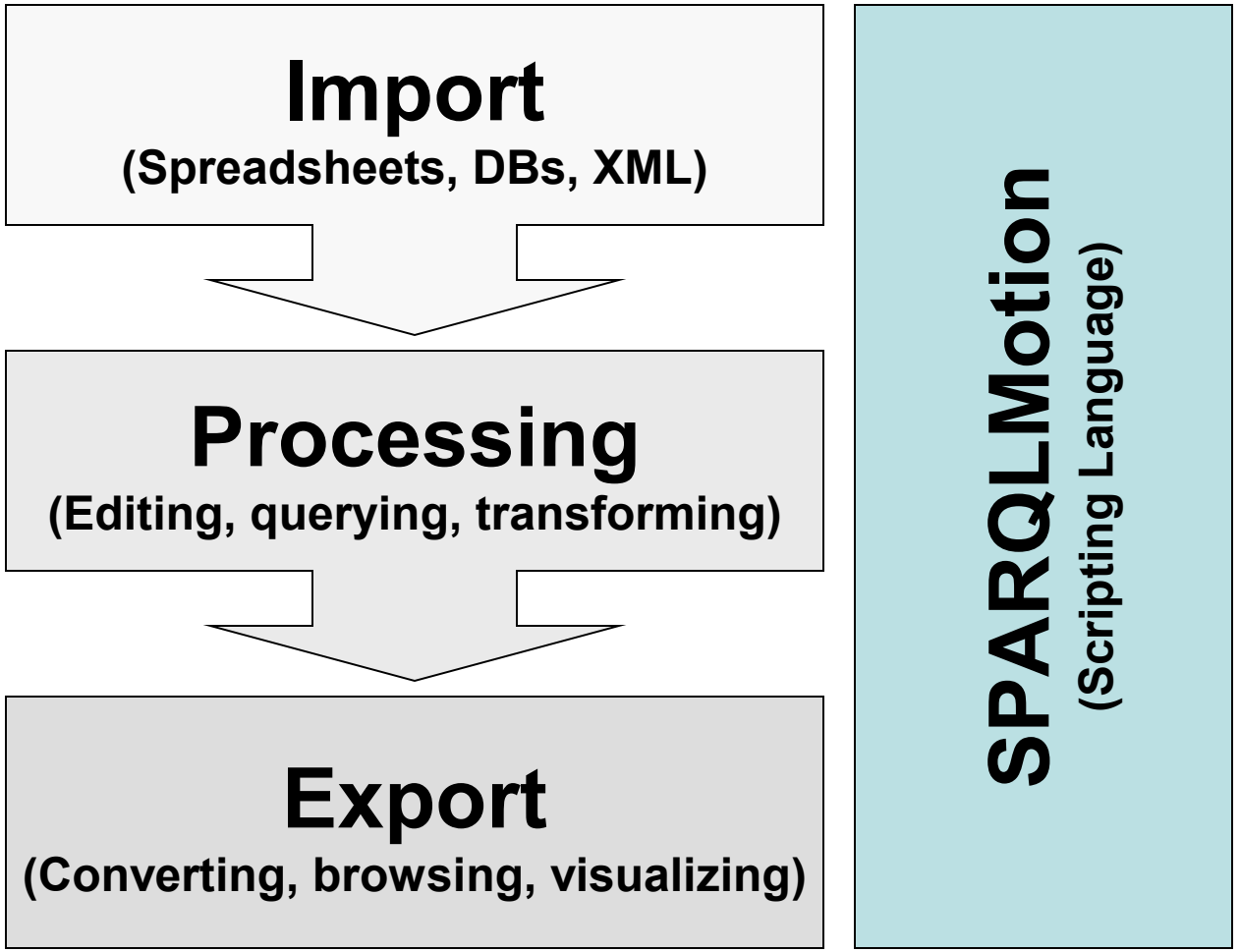
TopQuadrant™

Semantic Technology Examples (4)

- NASA Constellation project requires integration of information from an astonishingly wide variety of sources - different disciplines (hydraulics, electronics, mechanics, avionics, aerodynamics . . .).
- In the design stage any particular simulation (testing or evaluating design alternatives for a space system) will require a seamless view of a component from any number of perspectives. Even within a single discipline, different groups have information that contributes to a decision.
- Considering the operations and longevity requirements: Constellation project creates data that will be used 30 years into the future - think about the form of data 30 years ago (A lesson learned with the Space Shuttle, in which line drawings for designs had to be consulted 25 years later). The information architecture has to be flexible enough to withstand the passage of all those years.
- NASA is using TopBraid / Semantic Technologies to make flexible, future proof data systems to take a person to Mars.

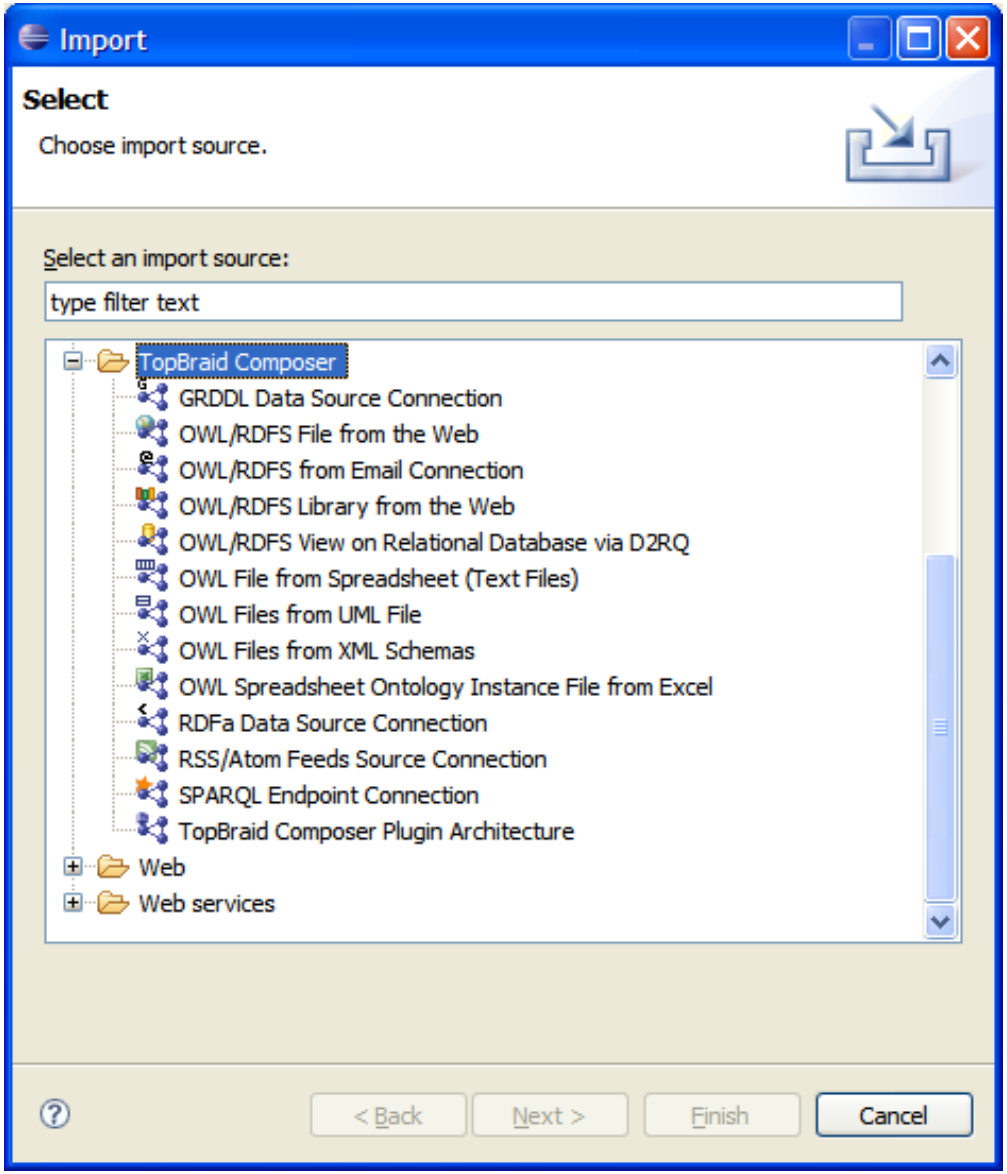
TopQuadrant™

Structure of this Talk



TopQuadrant™

TopBraid Import Features



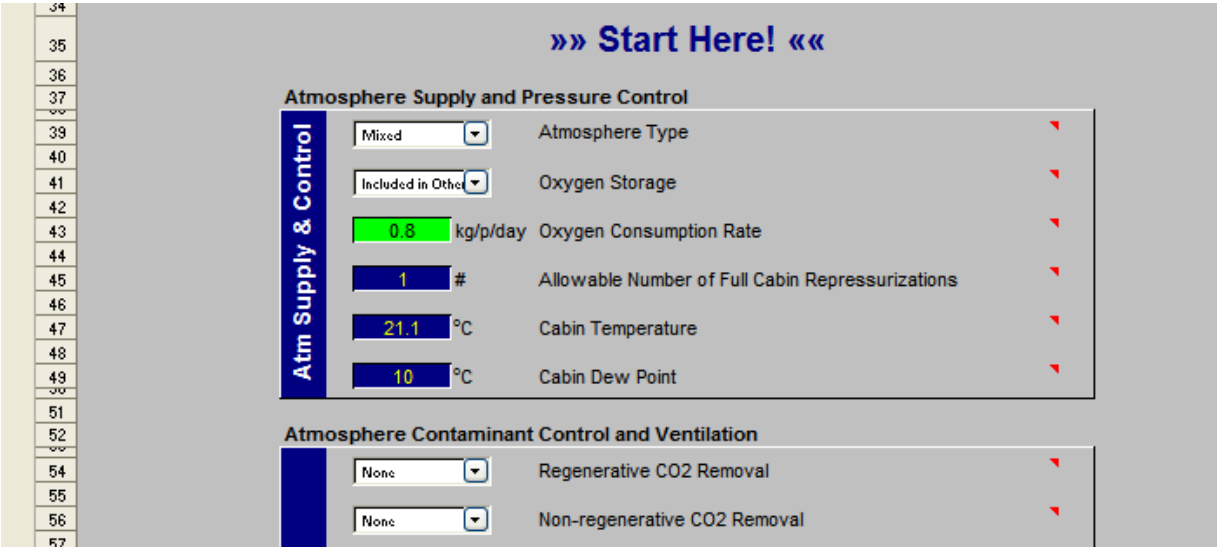
TopQuadrant™

Spreadsheet Import in TopBraid

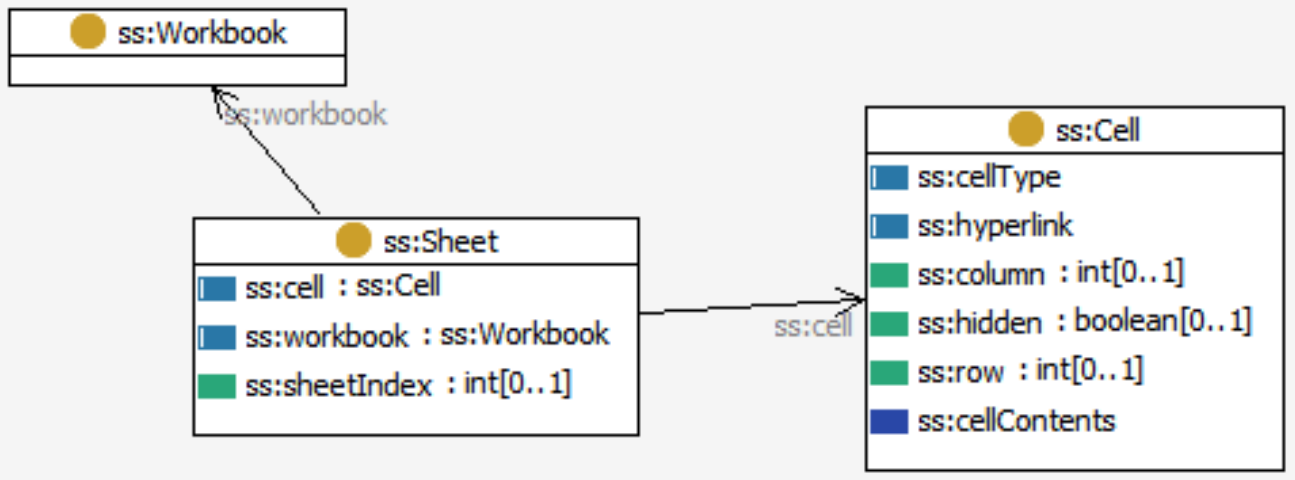
- In practice a lot (!) of useful data resides in spreadsheets
- Excel Spreadsheets can be quite sophisticated (programs on their own)
- TopBraid has two importing options
  - Excel files, each cell becomes an instance
  - Text files, each row becomes an instance

Excel Import in TopBraid

- Sometimes, spreadsheets are not just single tables
- Each cell may have a distinct meaning
- Information about cell position must be preserved

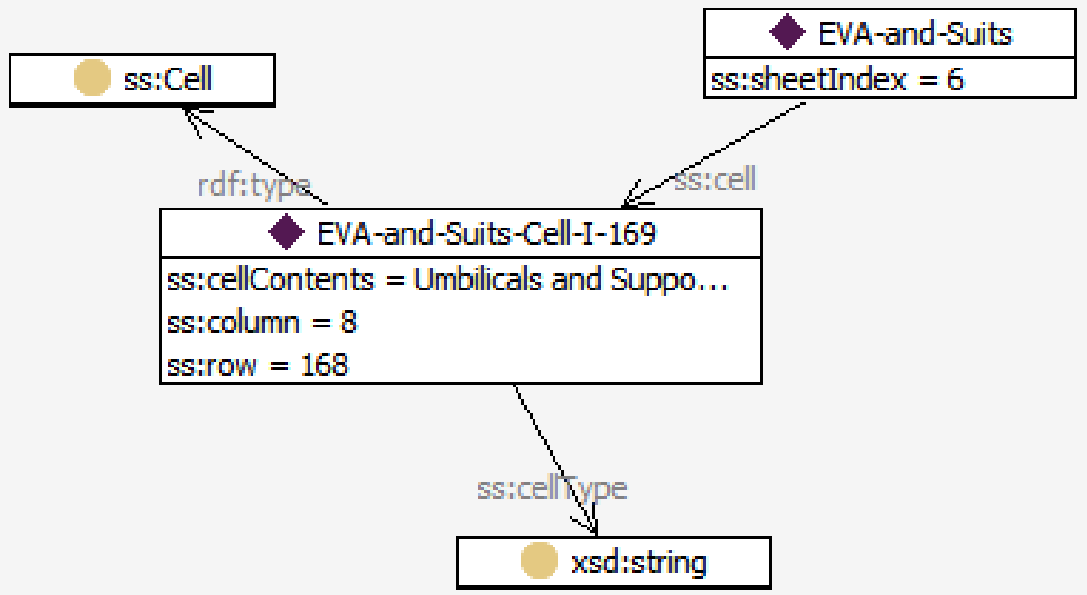


TopBraid Spreadsheet Ontology



Classes

Example Instance

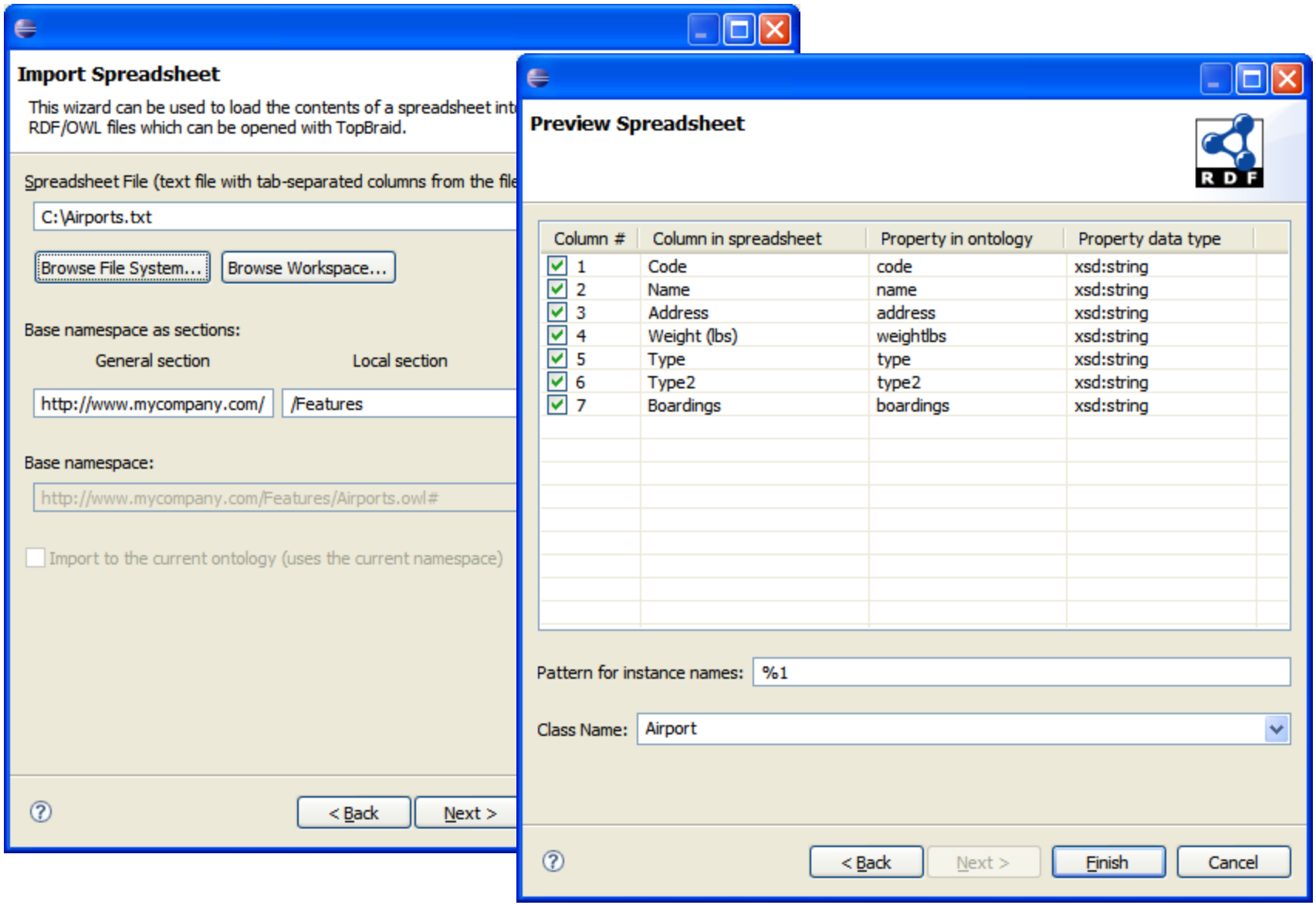


TopQuadrant™

Spreadsheet Import

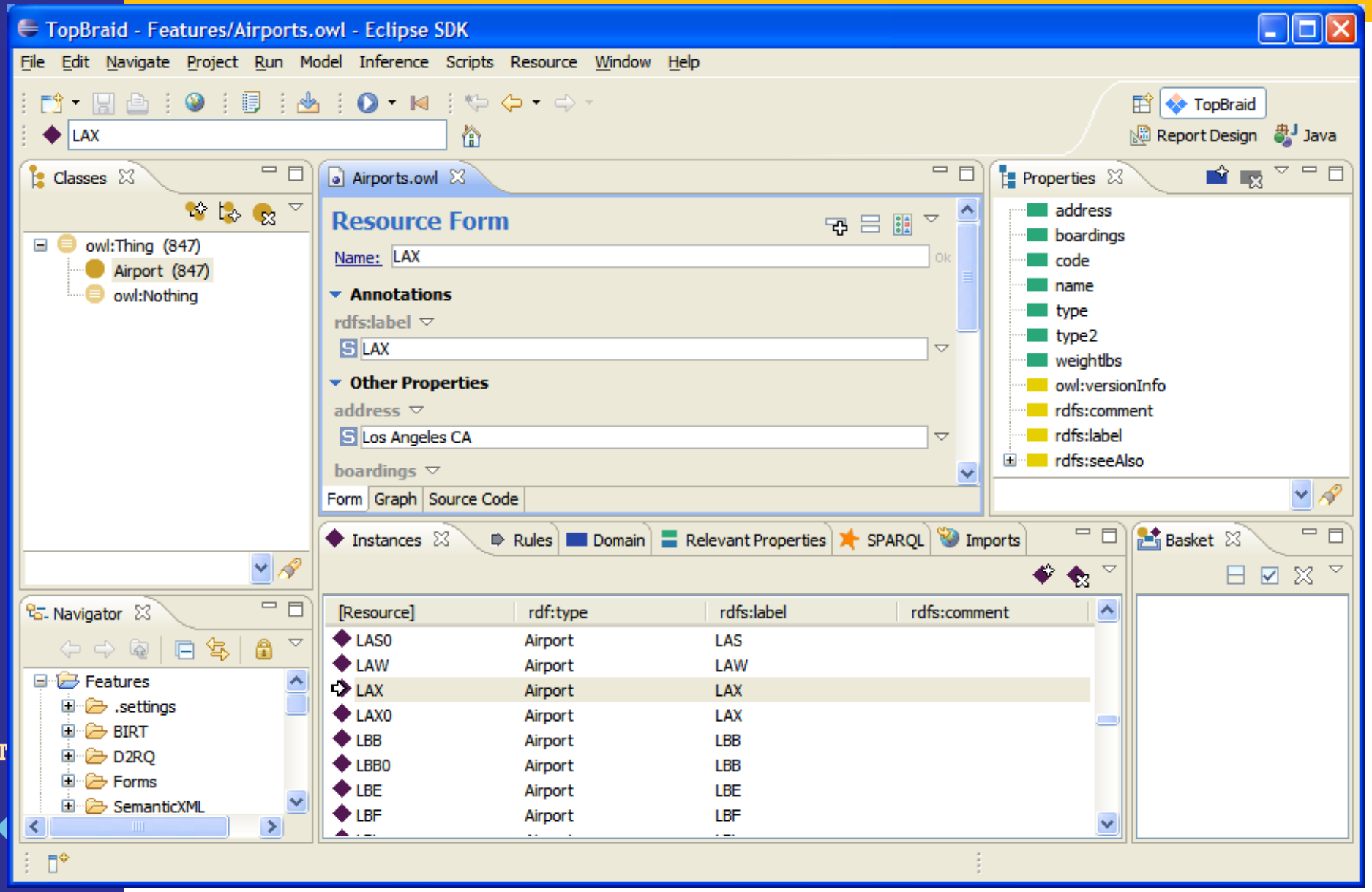
- Input: Tab-separated text files
- Table is interpreted as class
- Columns can be mapped into properties
- Rows become instances
- Import wizard can be used to fine tune

TopBraid Spreadsheet Import (1)



TopQuadrant™

TopBraid Spreadsheet Import (2)

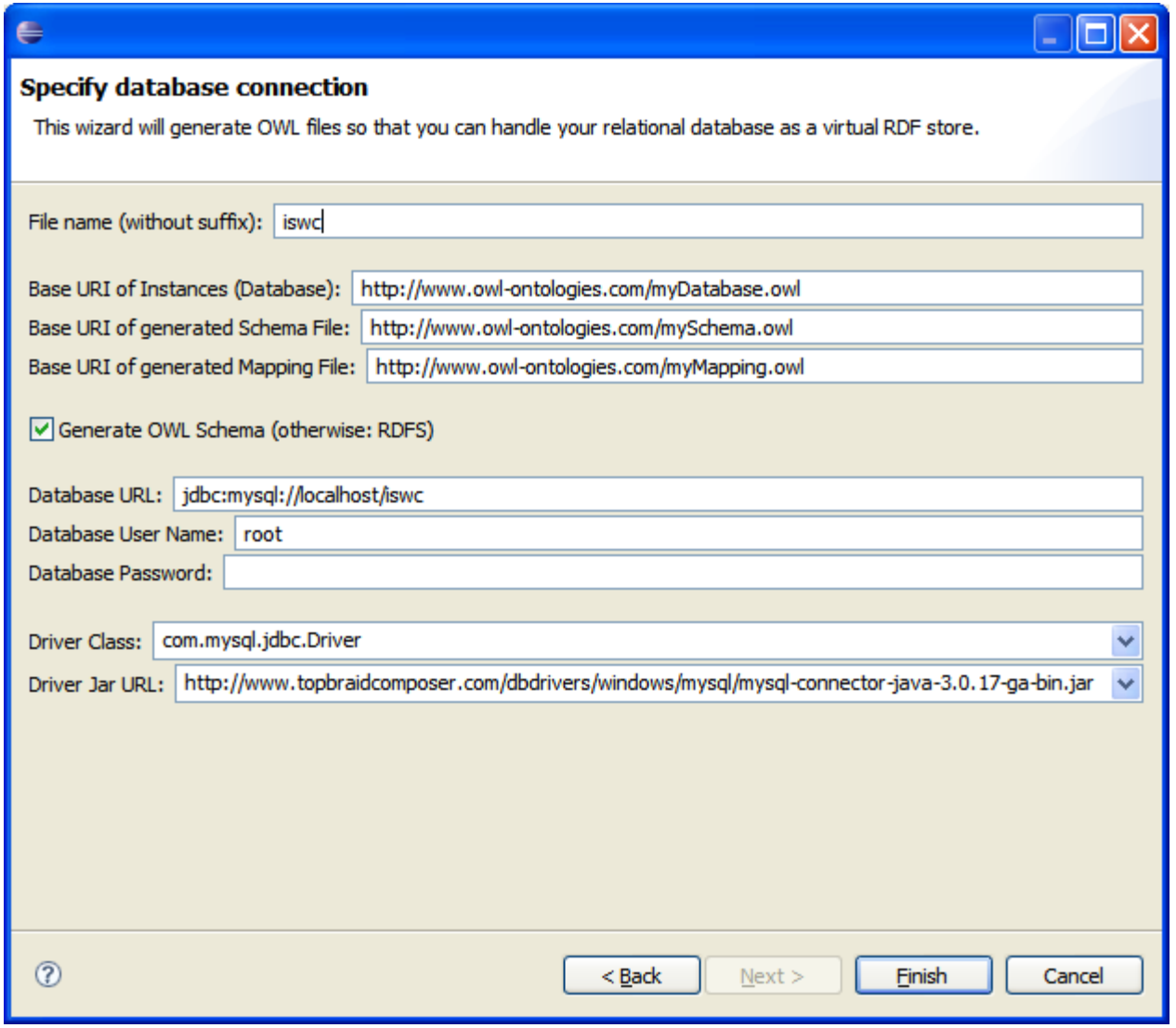


Relational Database Import

- Much enterprise data resides (and needs to stay) in relational databases
- Relational database importer (D2RQ) built into TopBraid
- Static import of schema
  - Tables become classes
  - Columns become properties
  - Link tables become object properties
- Dynamic import of actual data
  - Rows become instances
  - On the fly, i.e. data can stay where it is

TopQuadrant™

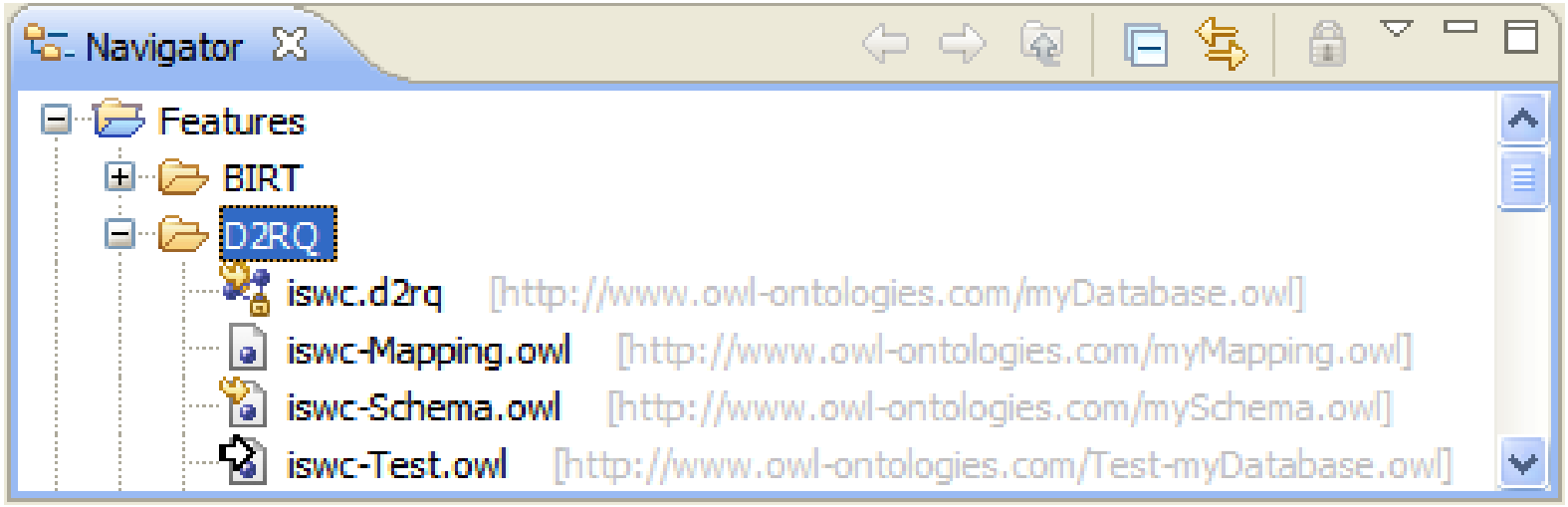
Database Import in TopBraid (1)



TopQuadrant™

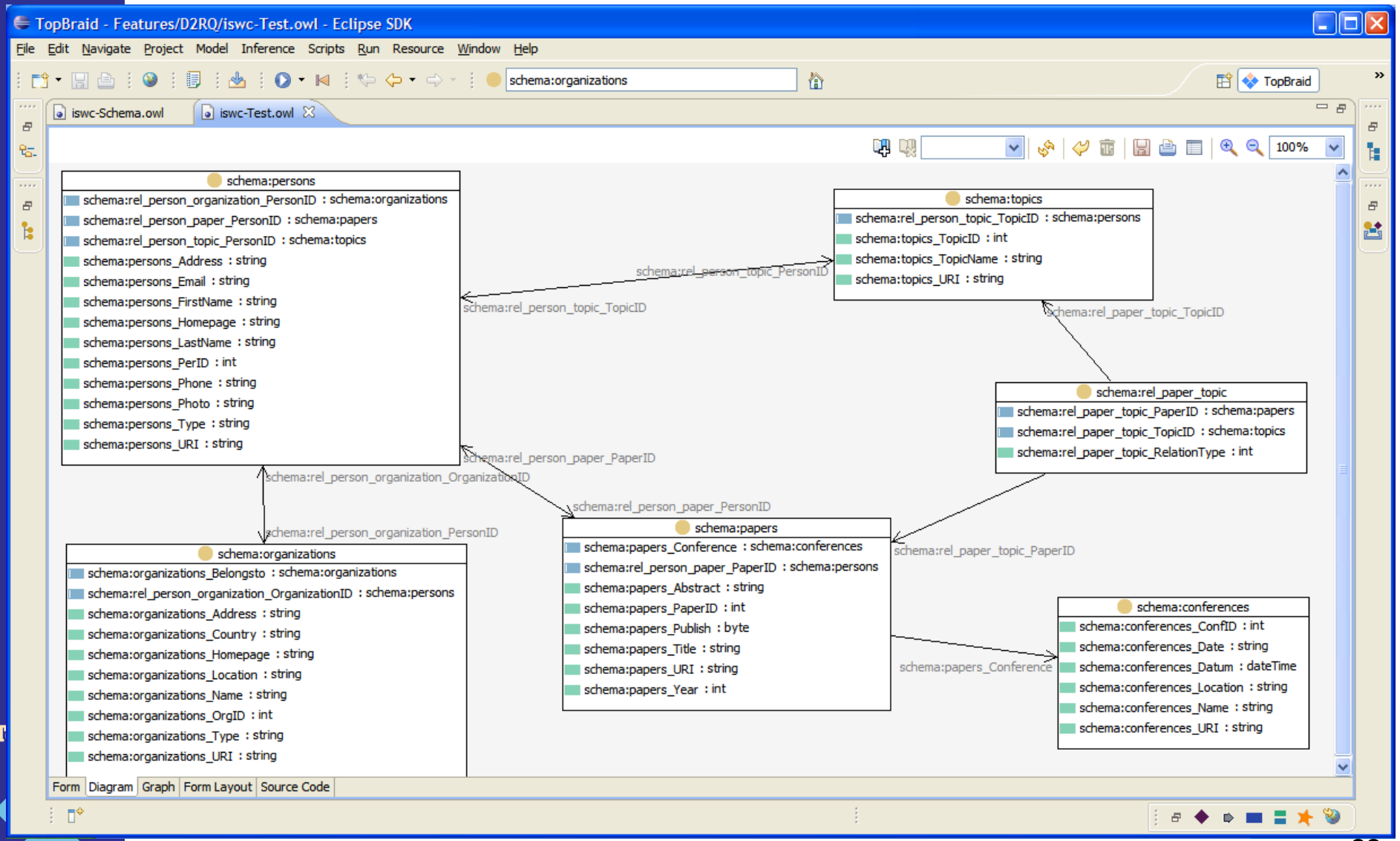
Database Import in TopBraid (2)

- TopBraid automatically generates
  1. Schema
  2. Instances placeholder file (.d2rq)
  3. Mapping file (table-to-class mapping)
  4. Test file that imports 1. and 2. based on 3.

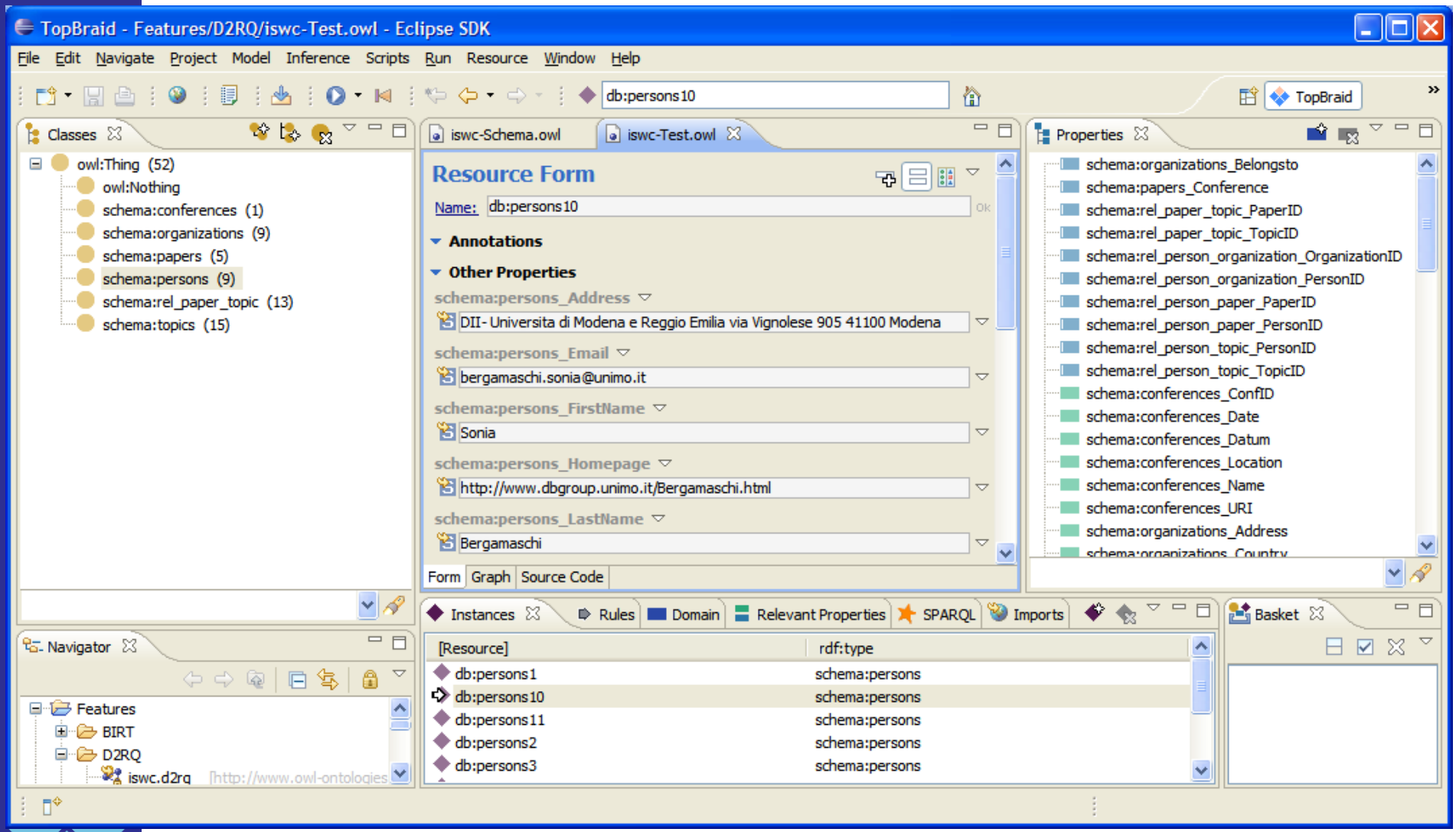


TopQuadrant™

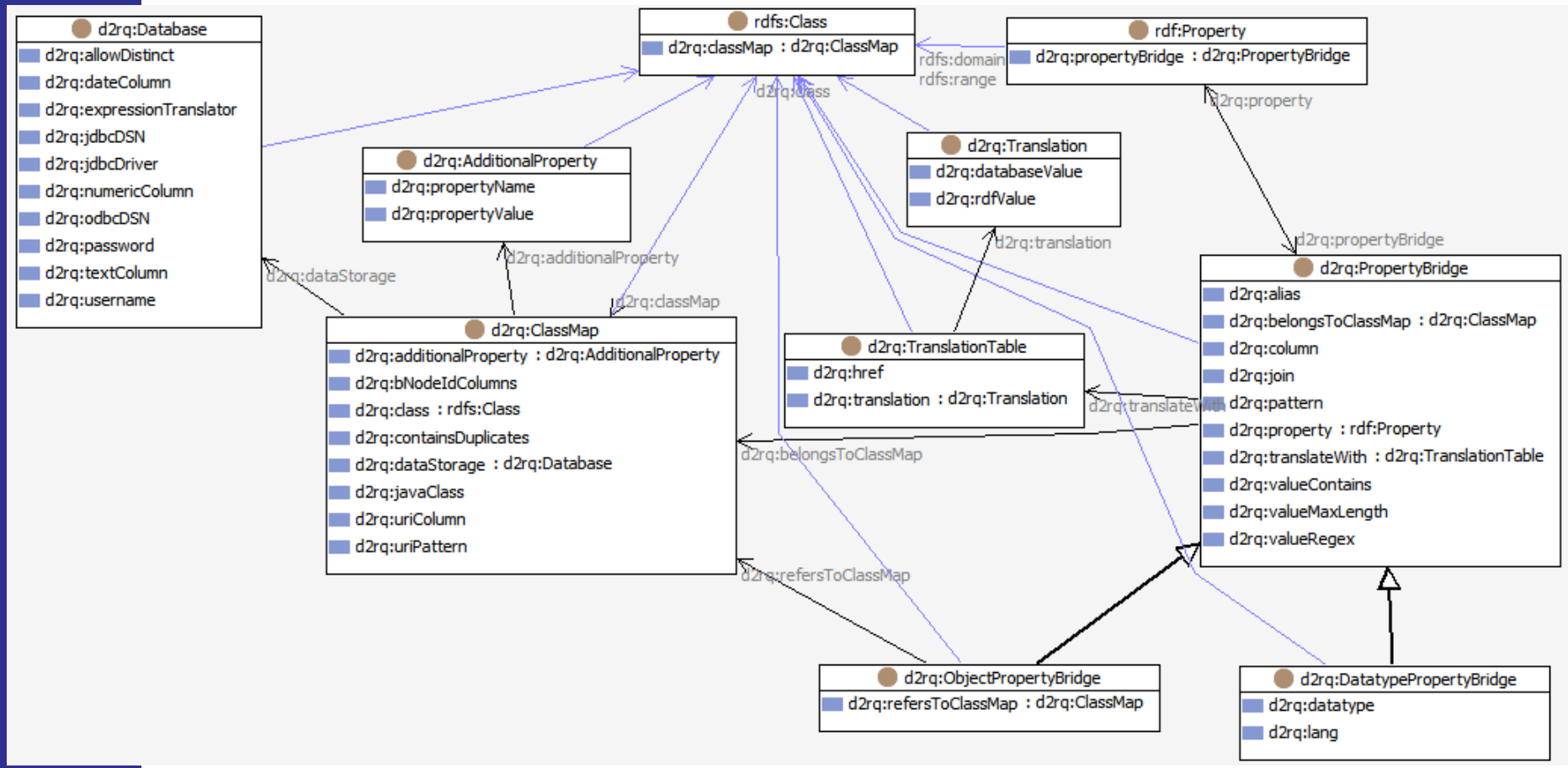
Database Import in TopBraid (3)



Database Import in TopBraid (4)



Database Import in TopBraid (5)



D2RQ Mapping Ontology

TopQuadrant™

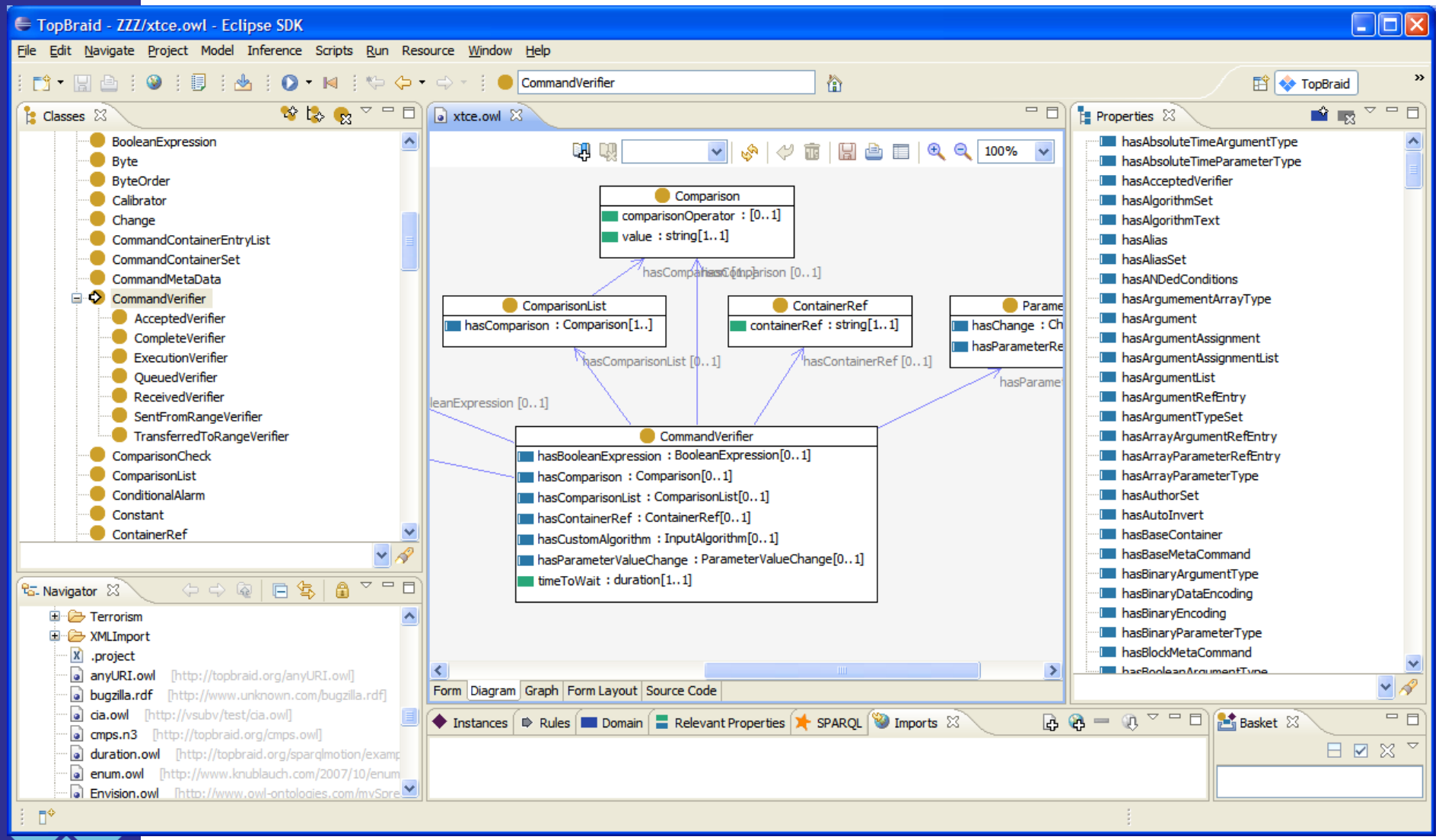
Database Import in TopBraid (6)

- Relational databases imported by D2RQ become triple sources like any other – but original data can stay where it is
- Resulting mapping can be fine-tuned
- Full range of generic RDF/OWL tools can be executed
  - Inferencing
  - Merging
  - Mapping
  - Querying
- Not all of these perform equally well

XML Import/Export

- XML is the favourite syntax in many areas, e.g. data exchange between tools, web services
- TopBraid supports two approaches
  - XML Schema import to ontology
  - Semantic XML

XML Schema Import/Export



Semantic XML

- Converts arbitary XML to OWL
- Keeps reverse-engineering info in the resulting ontology, using annotation properties
- Can create XML files from OWL
- Lossless round-tripping of XML
- Mapping ontologies can be edited

TopQuadrant™

Semantic XML Example

sxml.owl | plugin.xml

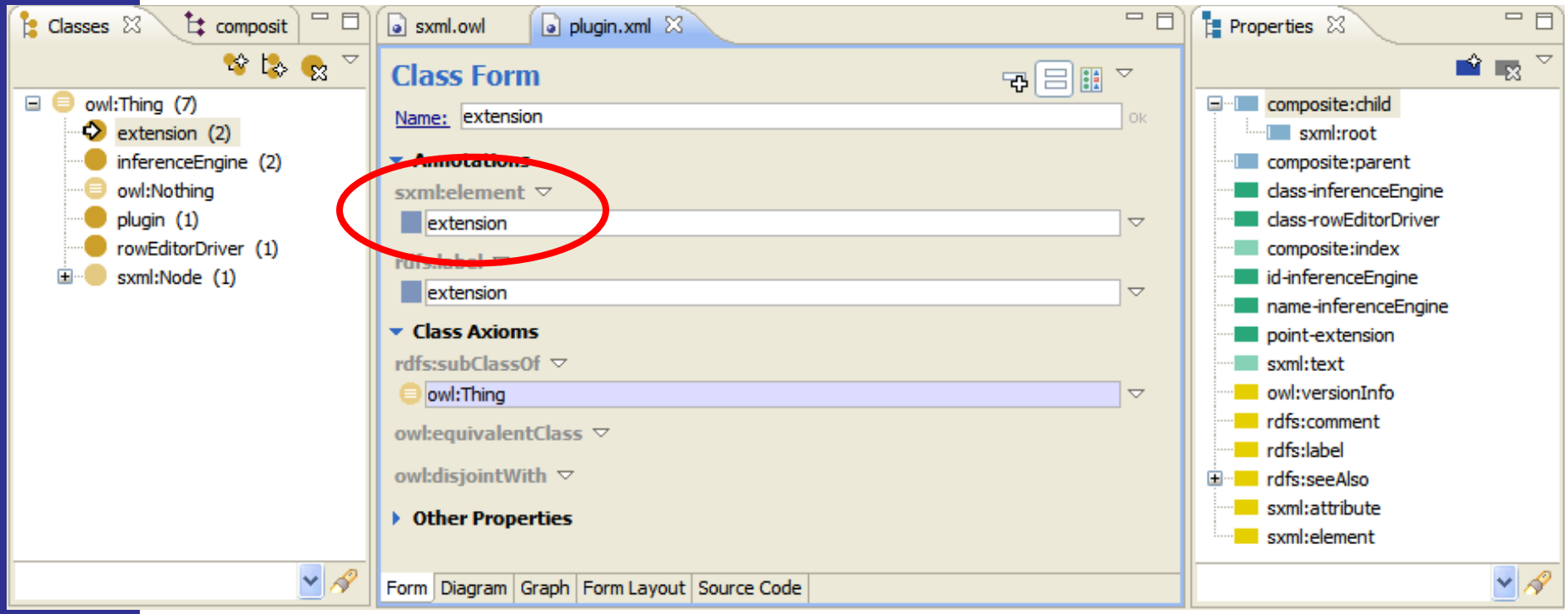
```
<?xml version="1.0" encoding="UTF-8"?>
<?eclipse version="3.2"?>
<plugin>
   <extension
         point="org.topbraid.inference.inferenceEngines">
      <inferenceEngine
            class="org.topbraid.sparql.inference.SPARQLInferenceEngine"
            id="org.topbraid.sparql.inference"
            name="SPARQL CONSTRUCTs"/>
      <inferenceEngine
            class="org.topbraid.sparql.inference.SPARQLResourceConstructorInferenceEngine"
            id="org.topbraid.sparql.constructors"
            name="SPARQL Resource Constructors">
      </inferenceEngine>
   </extension>
   <extension
         point="org.topbraid.strings.rowEditorDrivers">
      <rowEditorDriver class="org.topbraid.sparql.editor.SPARQLRowEditorDriver"/>
   </extension>

</plugin>
```

- Each element name becomes a class
- Each attribute becomes datatype property
- Nesting is mapped into a dedicated object property (composite:child)

TopQuadrant™

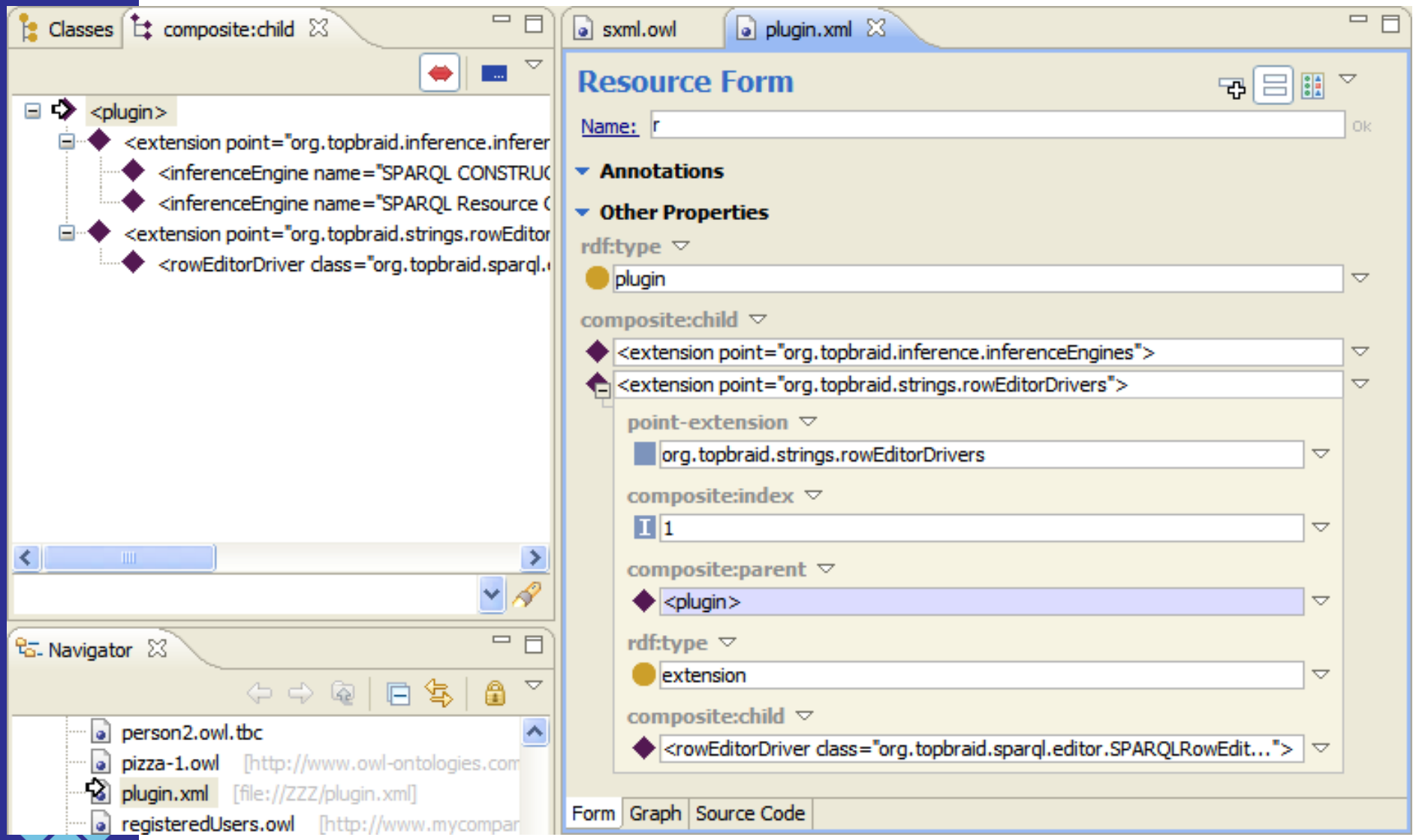
Semantic XML Classes



- Each generated class contains an annotation that points back to the XML element where it came from
- Similarly for the properties

TopQuadrant™

Semantic XML Instances

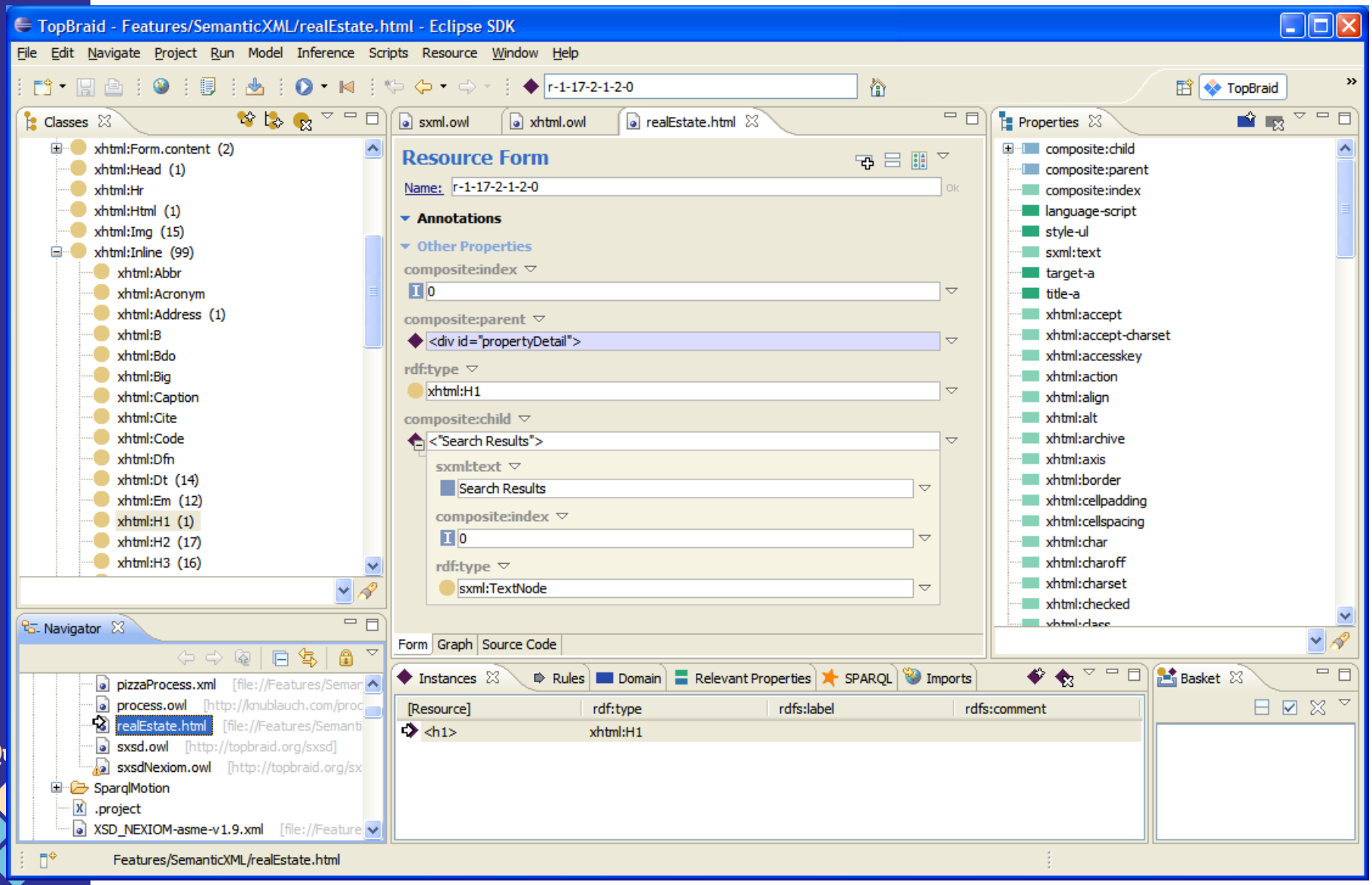


Semantic XML Profiles

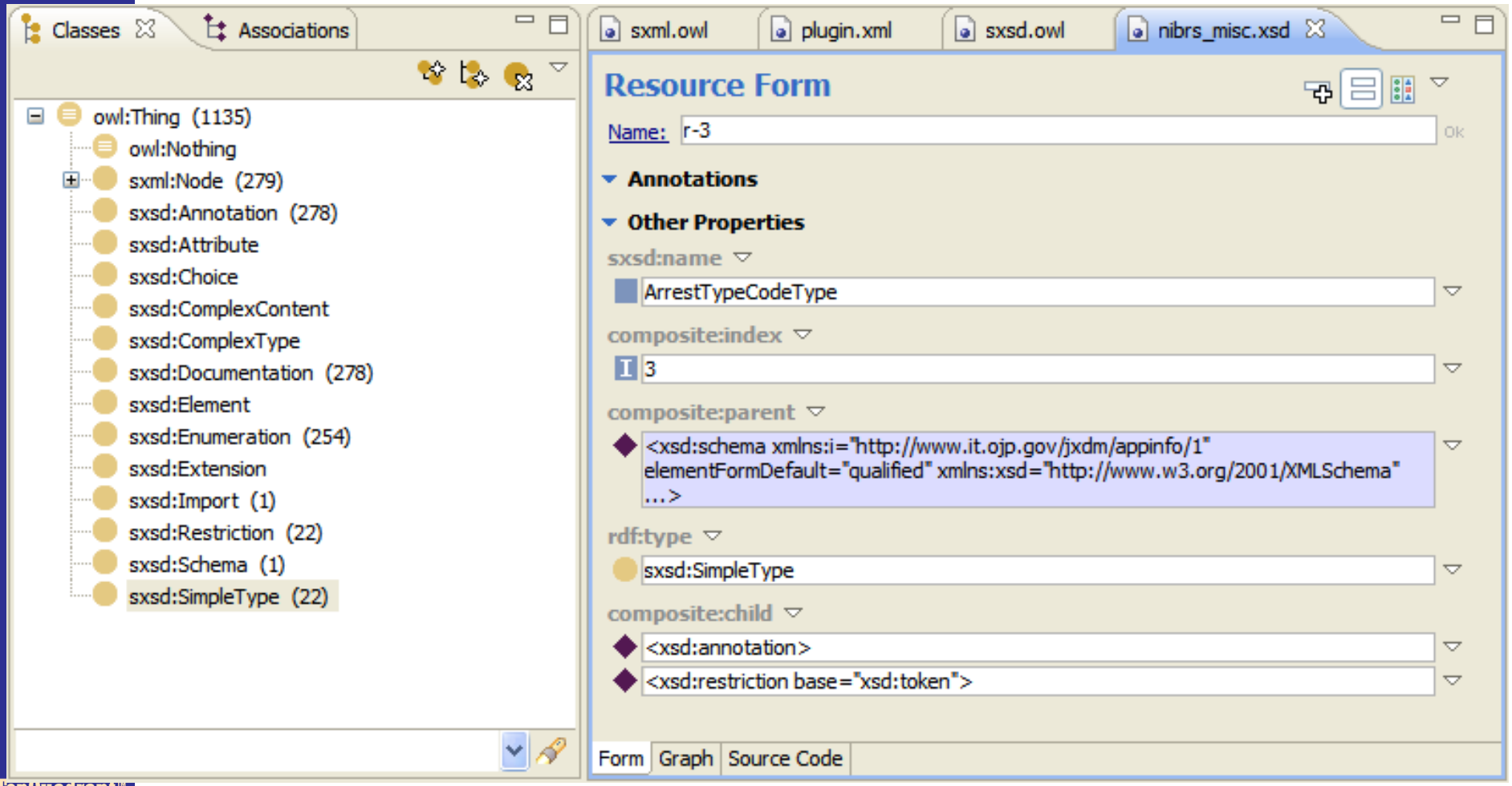
- The Semantic XML class models can be edited and fine-tuned
- TopBraid provides a couple of standard profiles
  - XHTML to open .html files (including tidy)
  - XSD to open XML Schemas
- More profiles are planned/prepared
  - X3D
  - SVG

TopQuadrant™

Semantic XML Profile for HTML



Semantic XML Profile for XSD



Semantic XML Summary

- Load, query and generate arbitrary XML files (even without XSD)
- Generated schema can then be fine tuned and reused for other XML files of the same kind
- SPARQL, rules and inferencers can be used to extract or convert the XML

Other Importers

- UML Class Diagrams
- Direct Triple Sources
  - Files (RDF/XML, N3/Turtle, N-Triples)
  - RSS/Atom Feeds
  - GRDDL
  - RDFa
  - SPARQL Endpoints
  - RDF databases (Oracle 11g, Jena, AllegroGraph, Sesame)

Data Processing

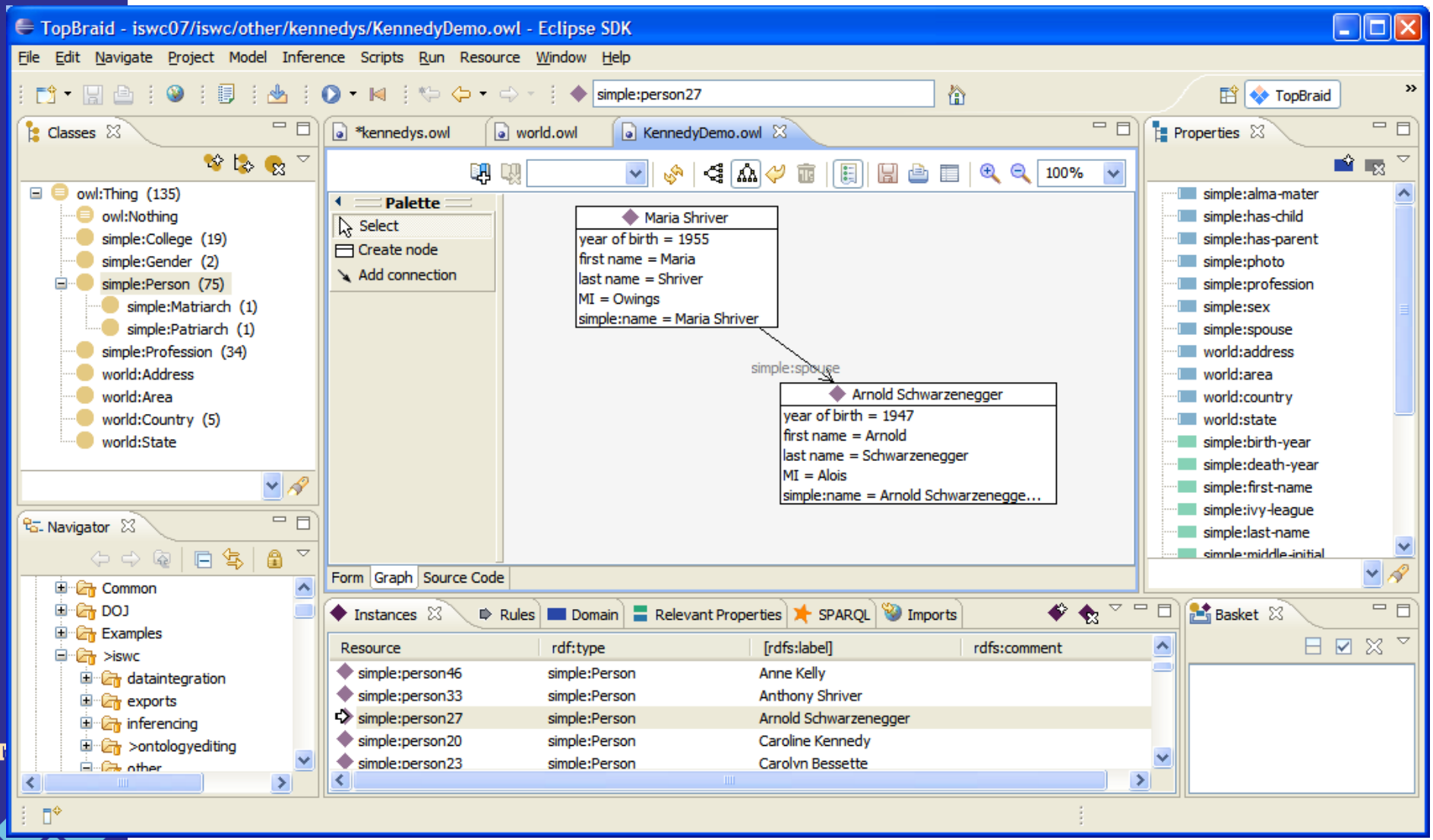
- So far: data physically converted to a uniform language (RDF/OWL)
- Semantic integration
  - Ontology editing
  - Mapping by built-in inferences
  - Mapping by constructing new triples

TopQuadrant™

Ontology Editing

- TopBraid Composer is the most sophisticated professional editor for OWL and RDF on the market
- Modular ontologies
- Refactoring
- Form-based & visual editing
- Customizable and extensible
- Driven by requirements from real-world projects (NASA etc)
- Several hundred users (and counting)

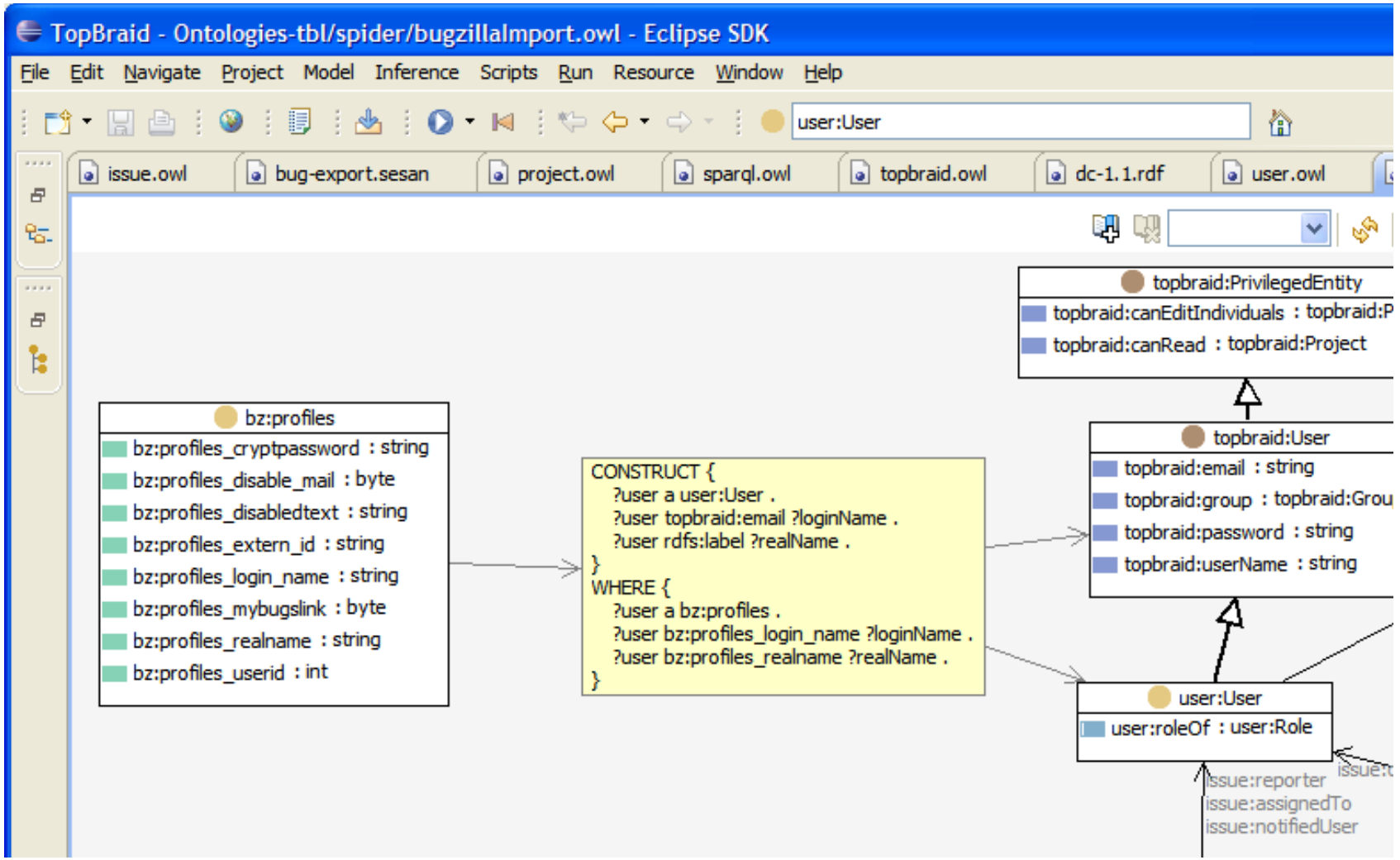
TopBraid Composer



Ontology Mapping via RDFS/OWL

- rdfs:subClassOf/owl:equivalentClass
- rdfs:subPropertyOf
- Then run inferencing
- Only suitable for trivial cases
- Limited expressivity

Ontology Mapping with SPARQL



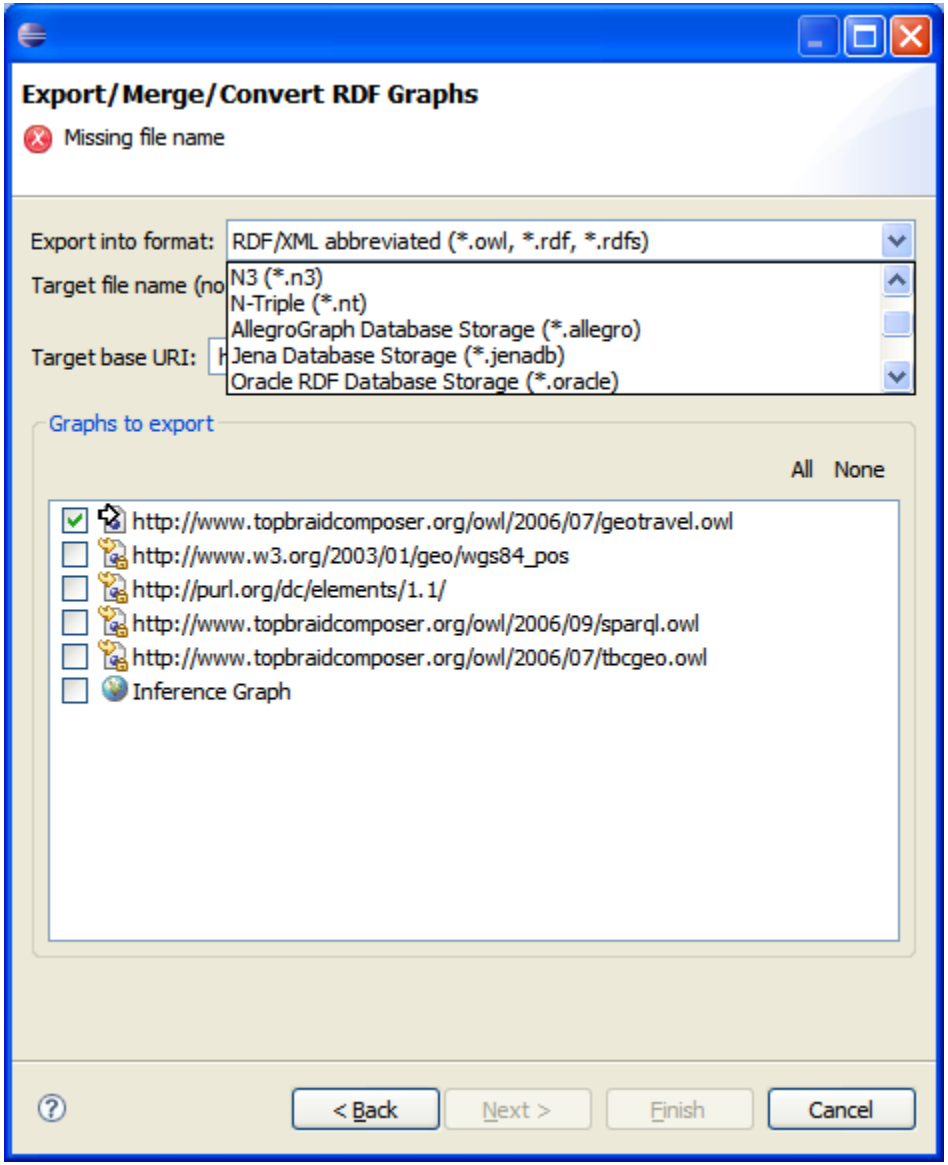
TopQuadrant™

TopBraid Export Features

- Triples (databases, files)
- HTML documentation
- Semantic Java Server Pages
- Google Maps, Calendars
- Spreadsheets, Matrix
- Business Intelligence Reports
- Browsing and querying (TopBraid Live)

TopQuadrant™

Export/Merge/Convert Triples



TopQuadrant™

HTML Export

Ontology Documentation

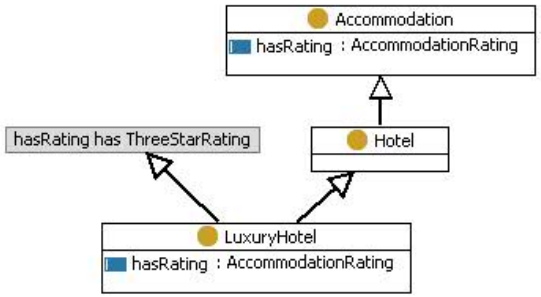
Navigation

Ontologies
Classes
Datatype Properties
Object Properties
Individuals
All Resources

- Destination
- FamilyDestination
- Farmland
- Hiking
- Hotel
- LuxuryHotel
- Motel
- Museums
- NationalPark
- QuietDestination
- Relaxation
- RetireeDestination
- RuralArea
- Safari
- Sightseeing
- Sports
- Sunbathing
- Surfing
- Town

http://www.owl-ontologies.com/travel.owl#LuxuryHotel

Class LuxuryHotel



Class Axioms

rdfs:subClassOf

- hasRating has ThreeStarRating
- Hotel

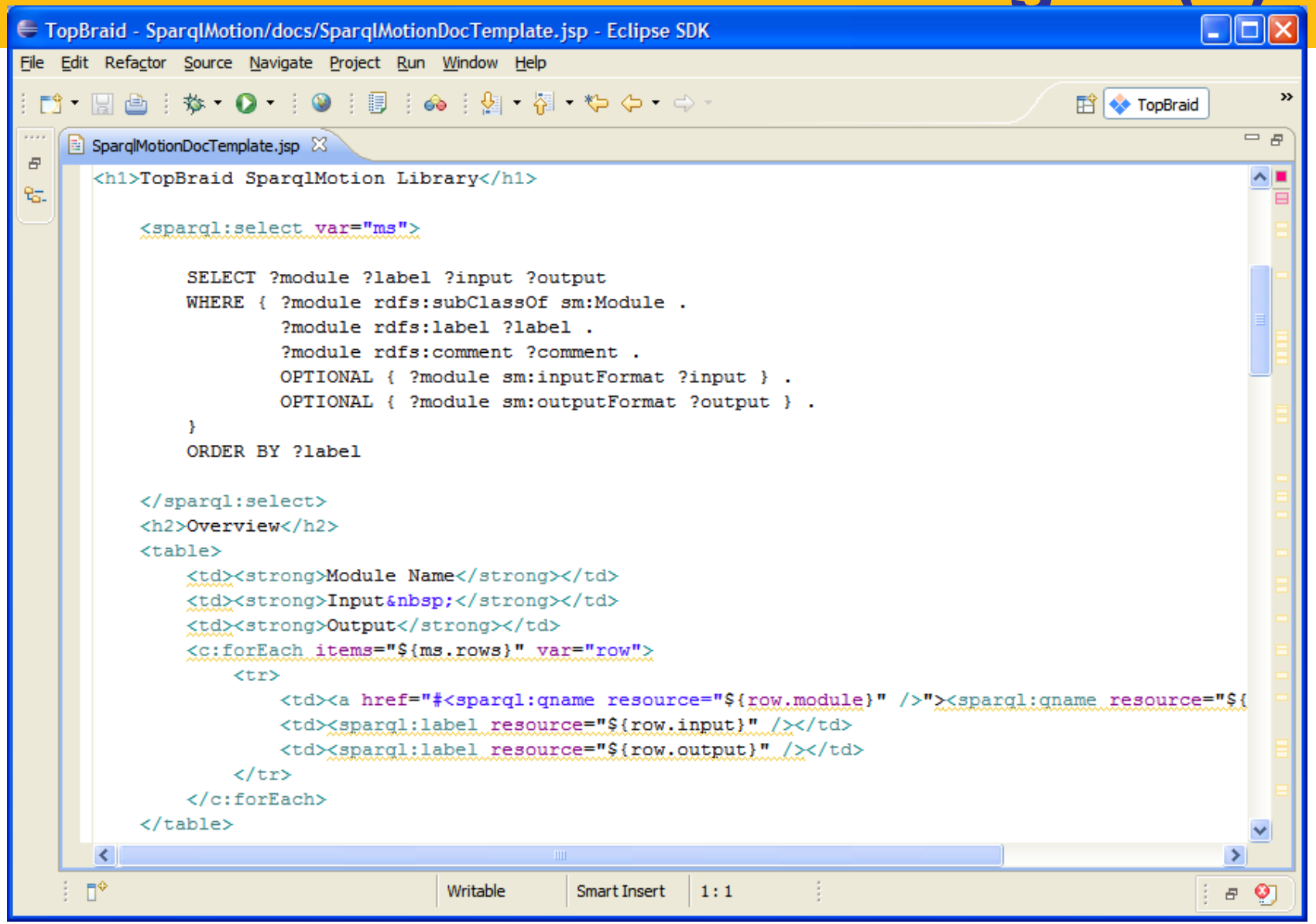
Other Properties

rdf:type

- owl:Class

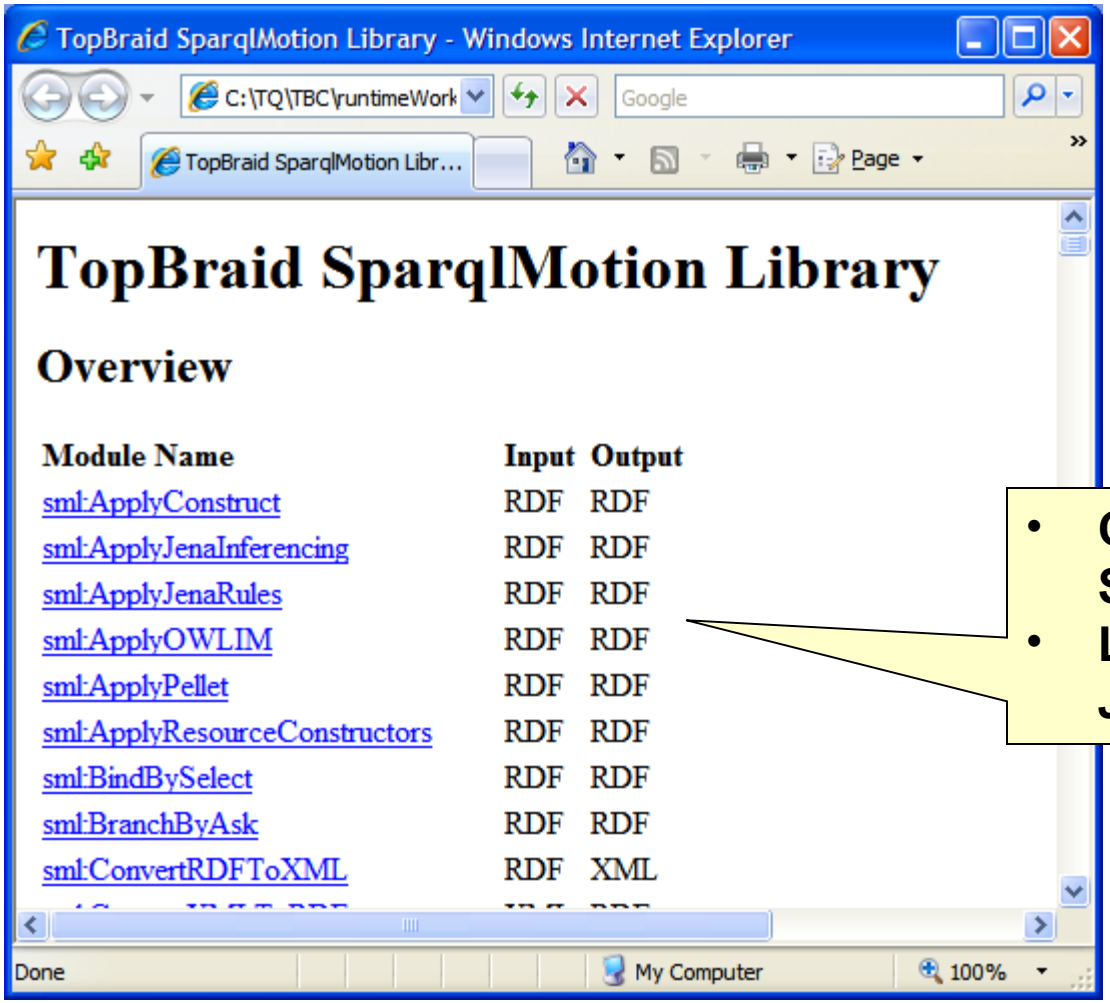
Instances

Semantic Java Server Pages (1)



TopQuadrant™

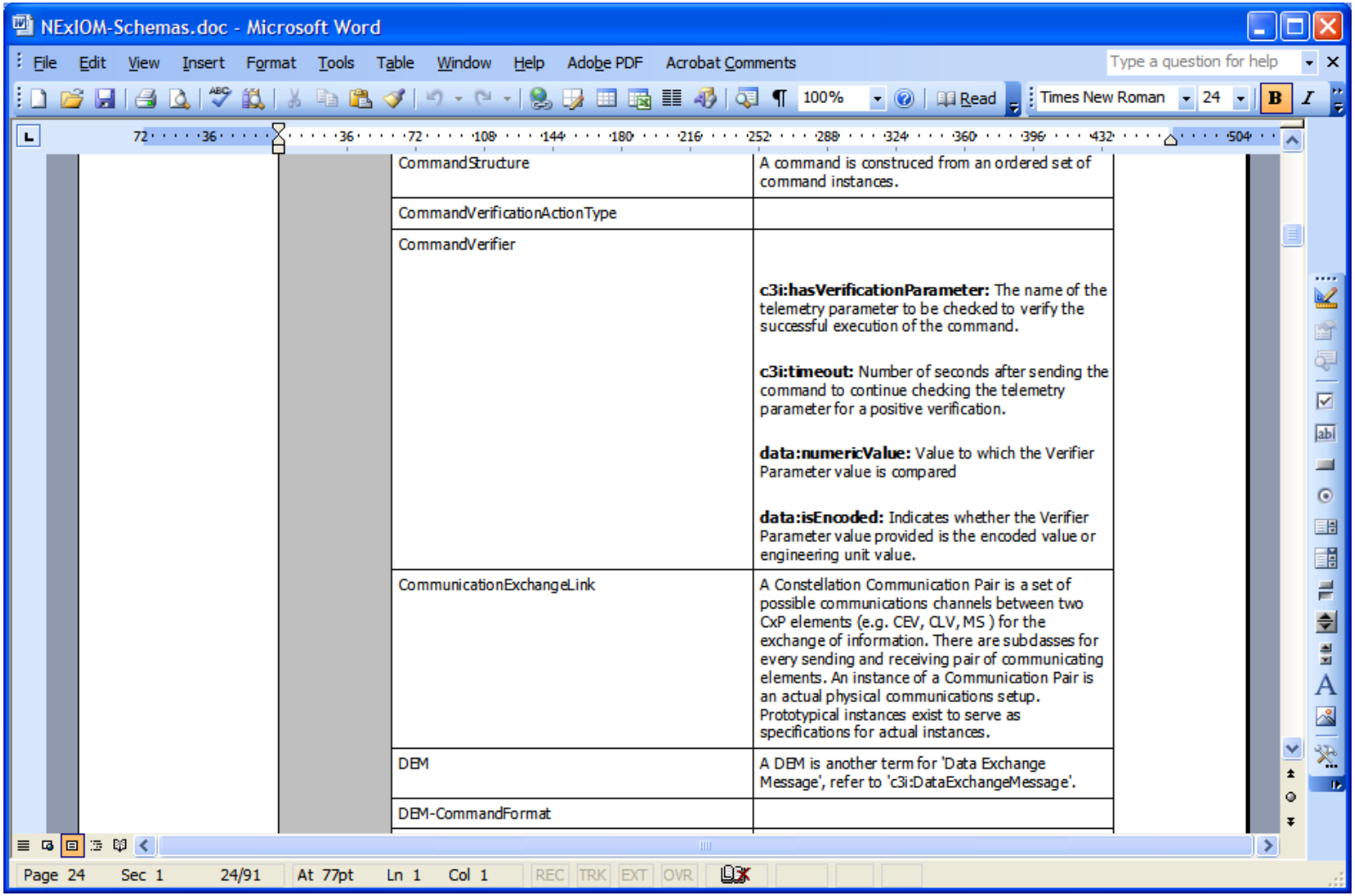
Semantic Java Server Pages (2)



- **Content driven by SPARQL queries.**
- **Layout defined by JSP template.**

TopQuadrant™

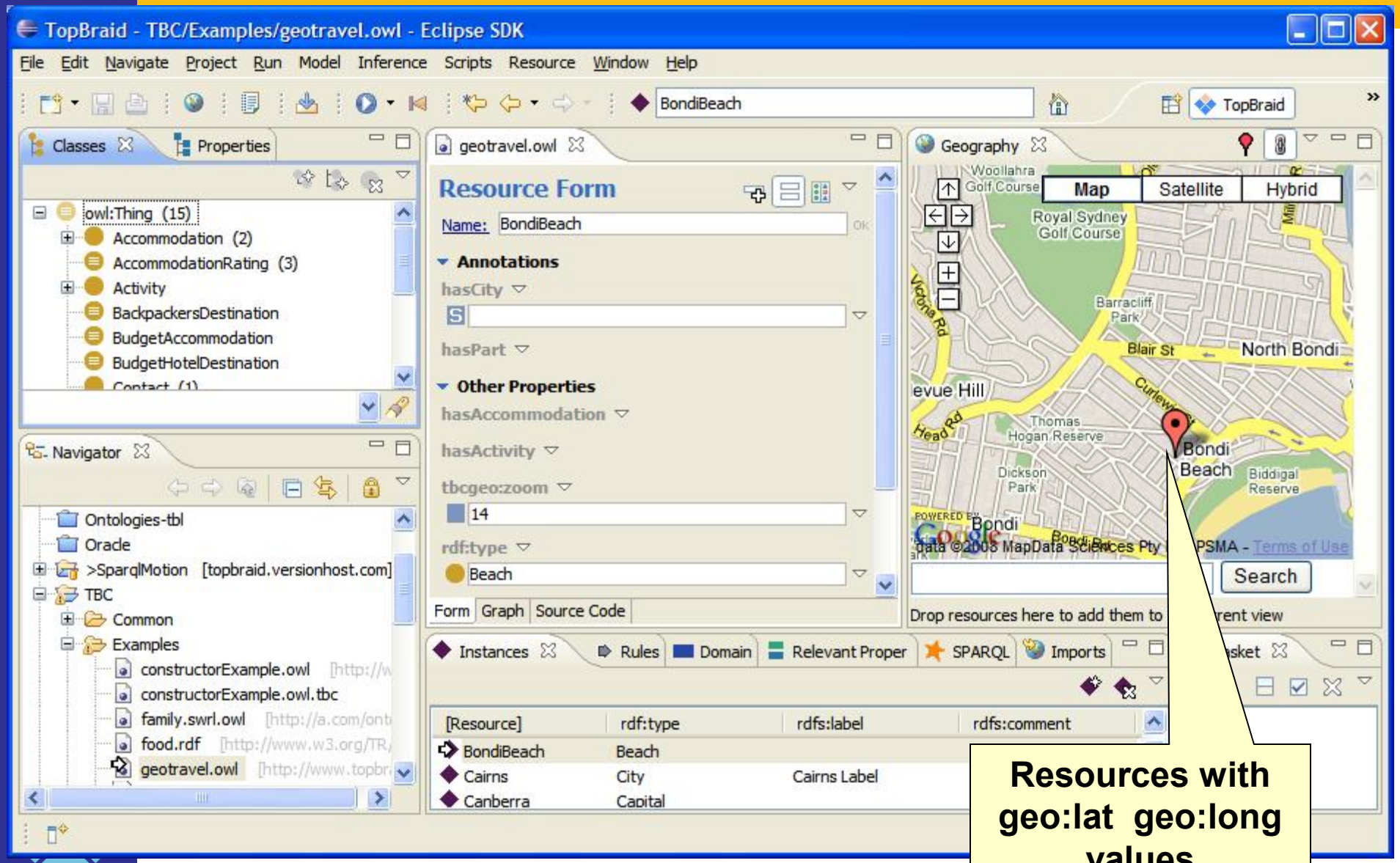
Semantic Java Server Pages (3)



Used extensively to generate all kinds of documents, deliverables

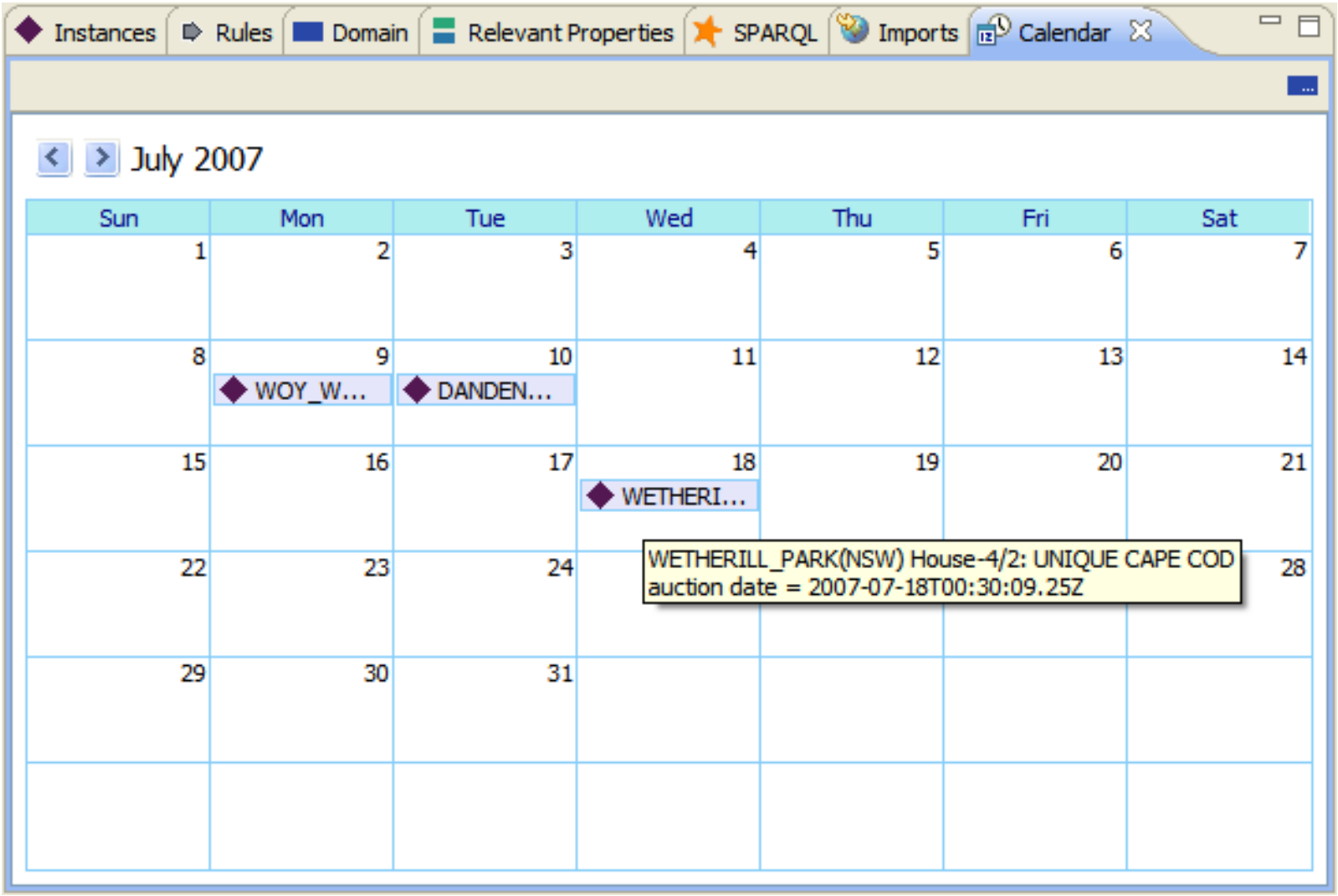
TopQuadrant™

Google Maps



Resources with geo:lat geo:long values

Calendars



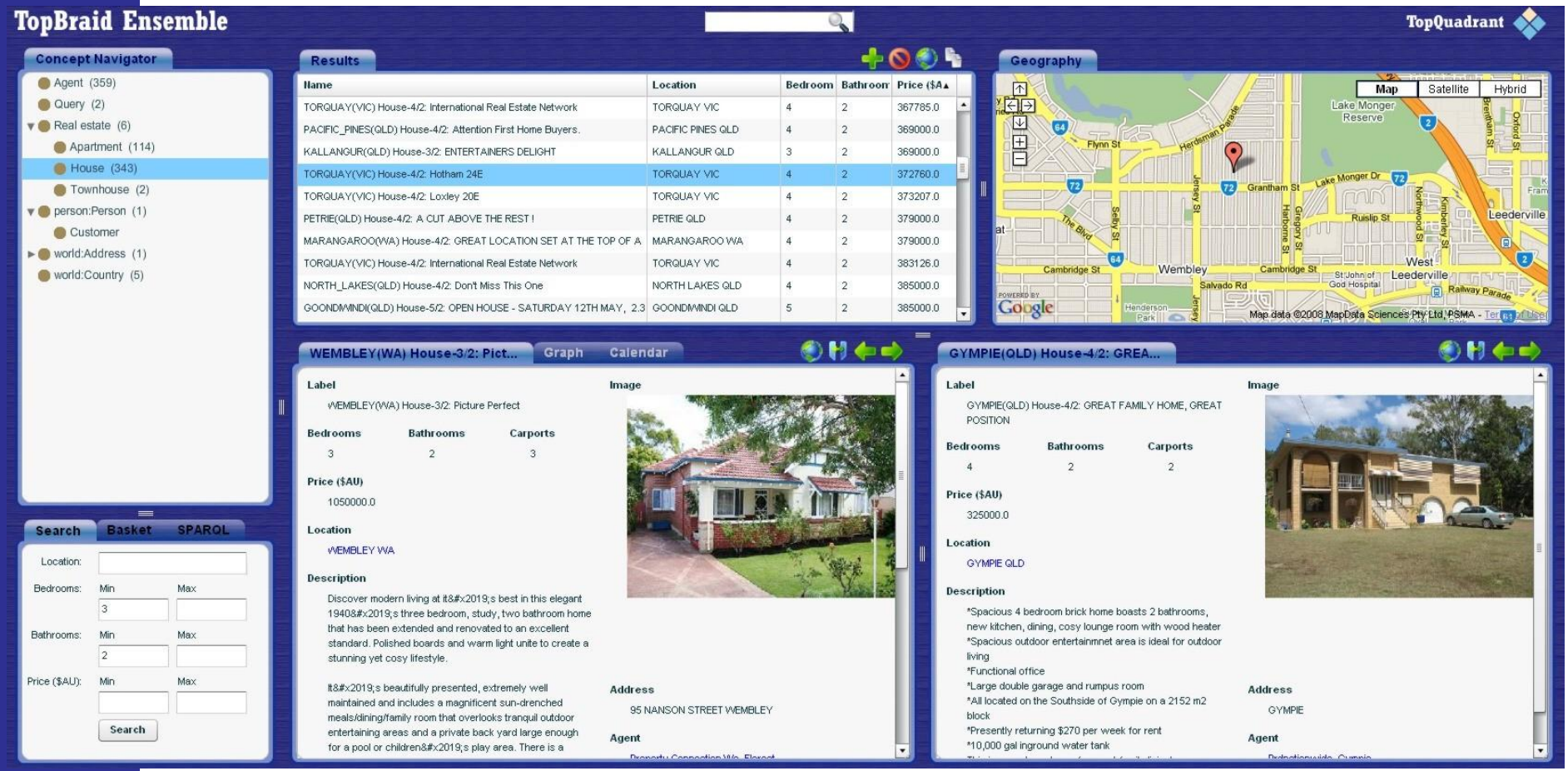
TopQuadrant™

BIRT Reports

Input: Tabular data from SPARQL queries

TopQuadrant™

TopBraid Ensemble (1)

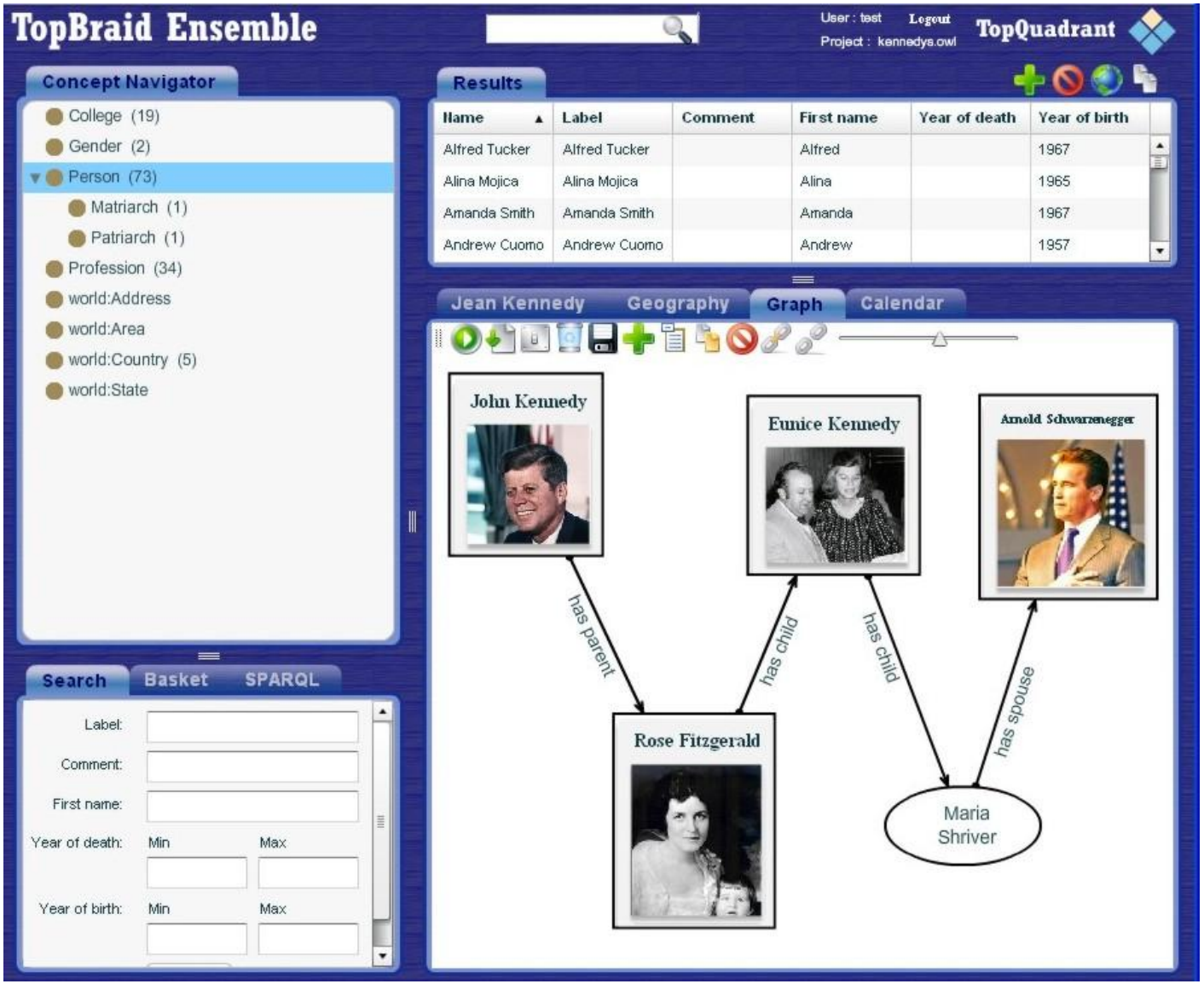


Rich Internet Application for browsing and editing RDF/OWL

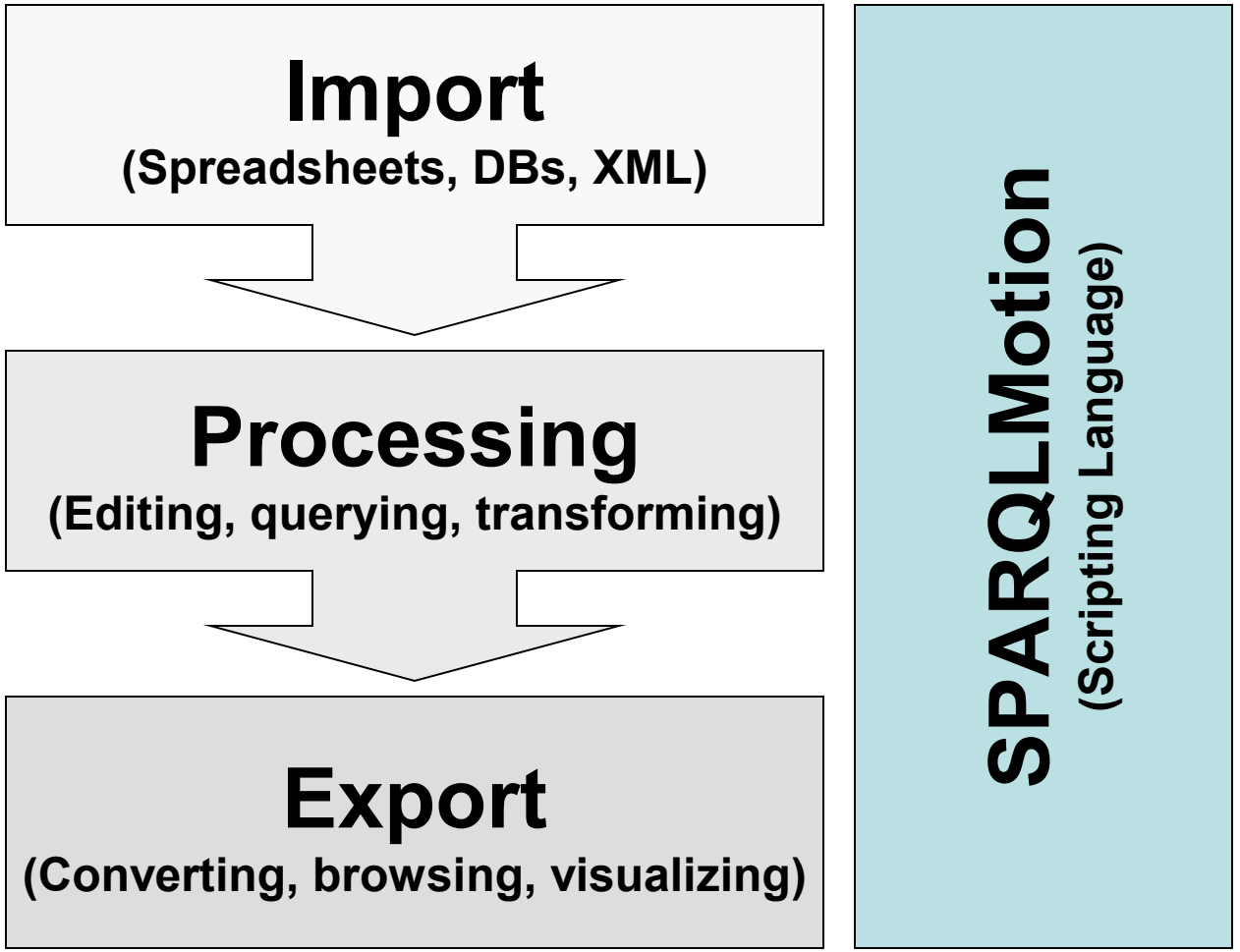
TopQuadrant™

TopBraid Ensemble (2)

TopBraid Ensemble (3)



Structure of this Talk



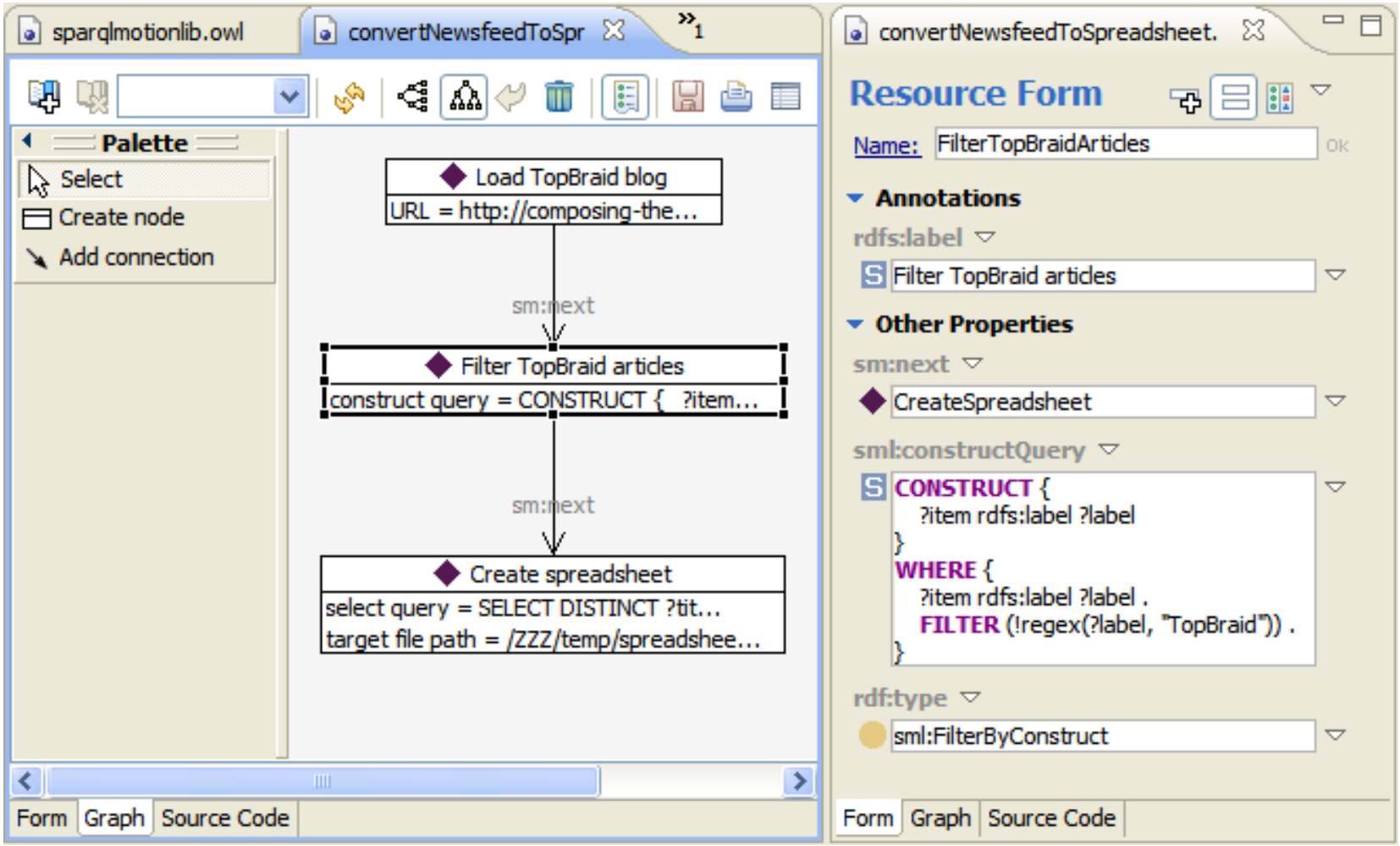
TopQuadrant™

SPARQLMotion

- A visual scripting language for Semantic Web Technology
- Import – Process – Export
- Use case: repeatable data processing and information integration tasks
- SPARQLMotion itself is defined as an OWL ontology
- Instance scripts can be edited with any OWL editing tool
- Has an extensible architecture

TopQuadrant™

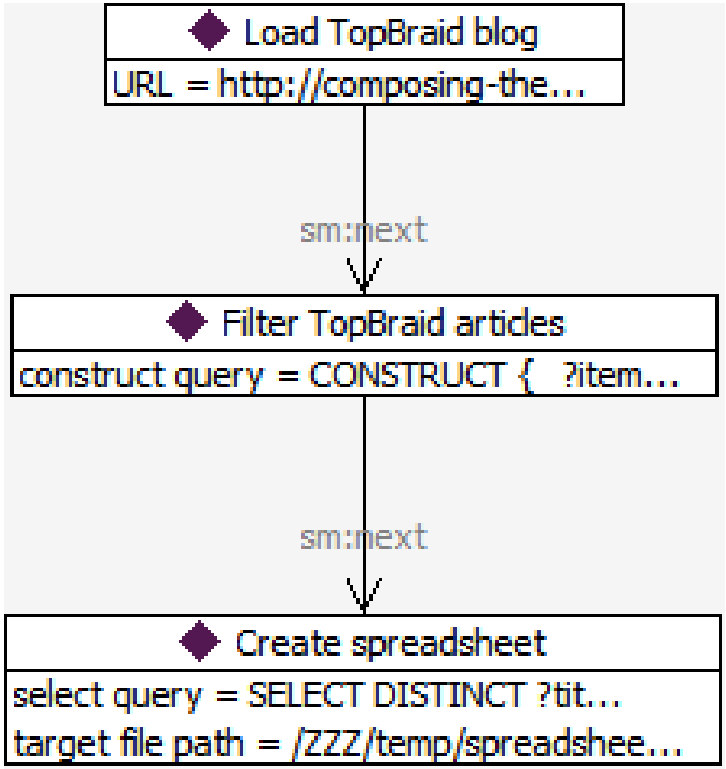
SPARQLMotion Example



TopQuadrant™

SPARQLMotion Language

- Scripts consist of modules
- Modules have a type (e.g. ApplyPellet)
- The output of one module is the input to its successors (RDF, XML and/or variable bindings)
- Branching (if-else), Iterations (while) and merging supported

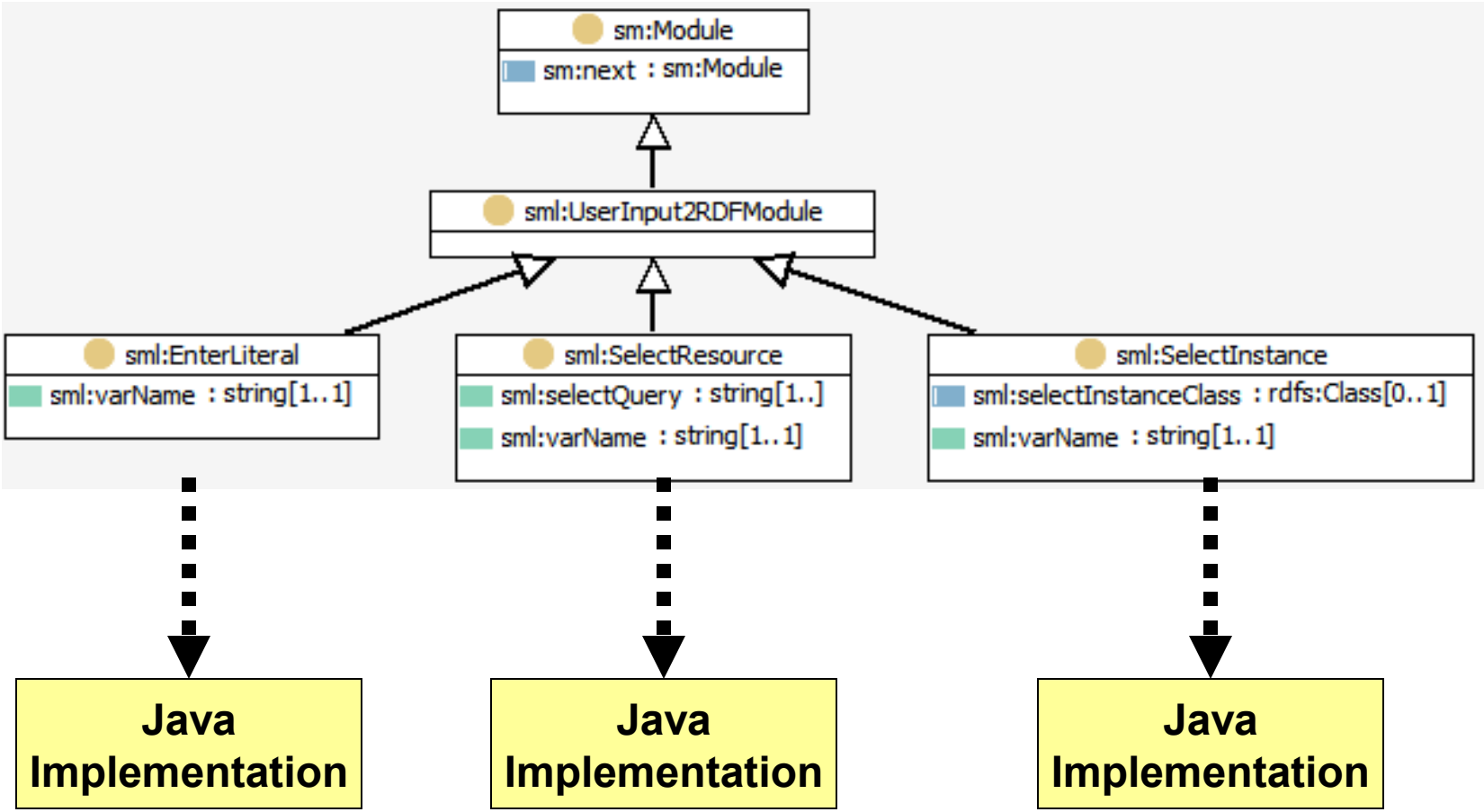


TopQuadrant™

SPARQLMotion Module Types

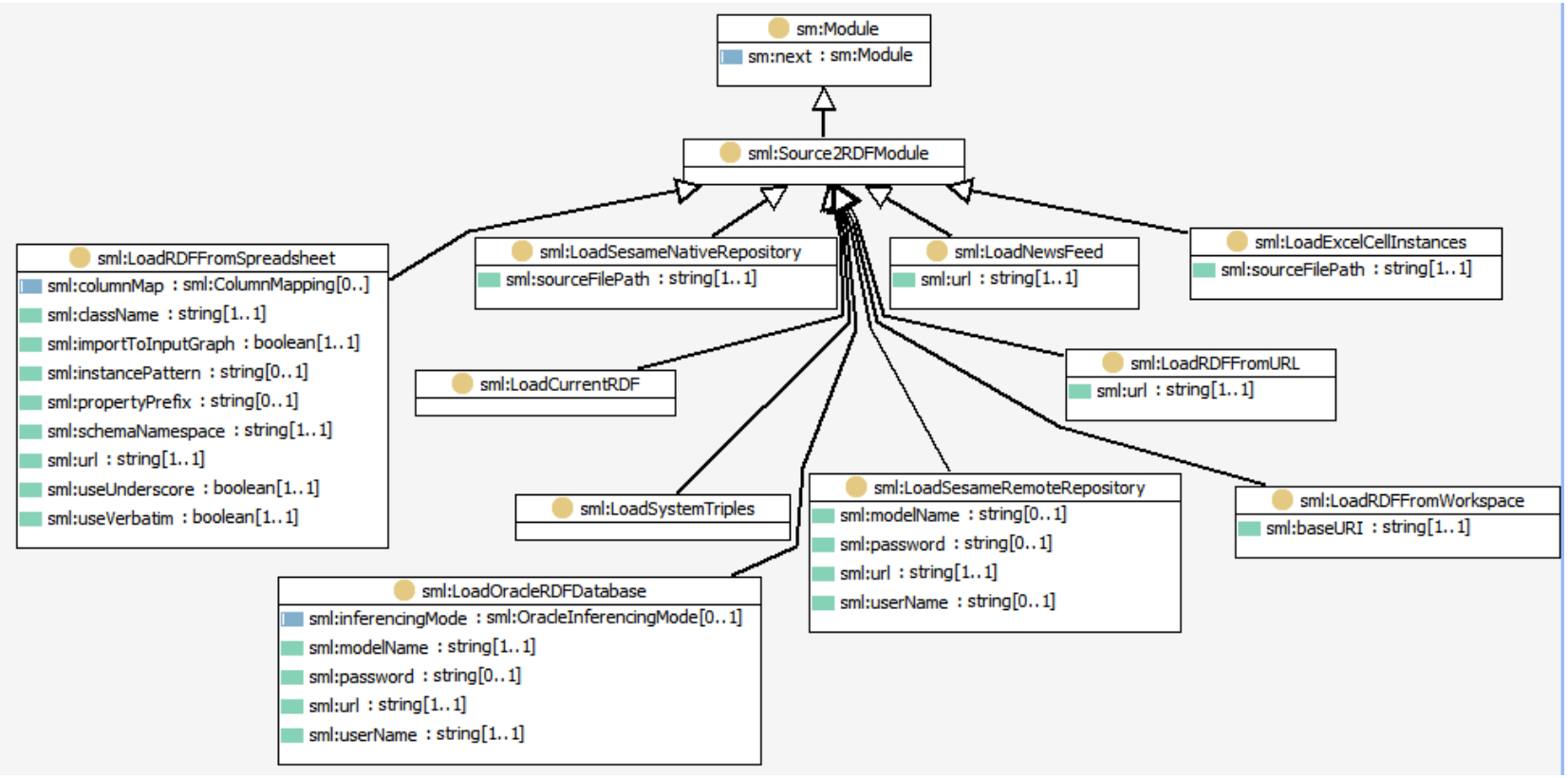
- Input
  - Something-to-RDF
  - Something-to-XML
  - User Input
- Processing
  - RDF-to-RDF
  - XML-to-RDF
  - RDF-to-XML
- Output
  - RDF-to-Output
  - XML-to-Output

SPARQLMotion Module Library (1)



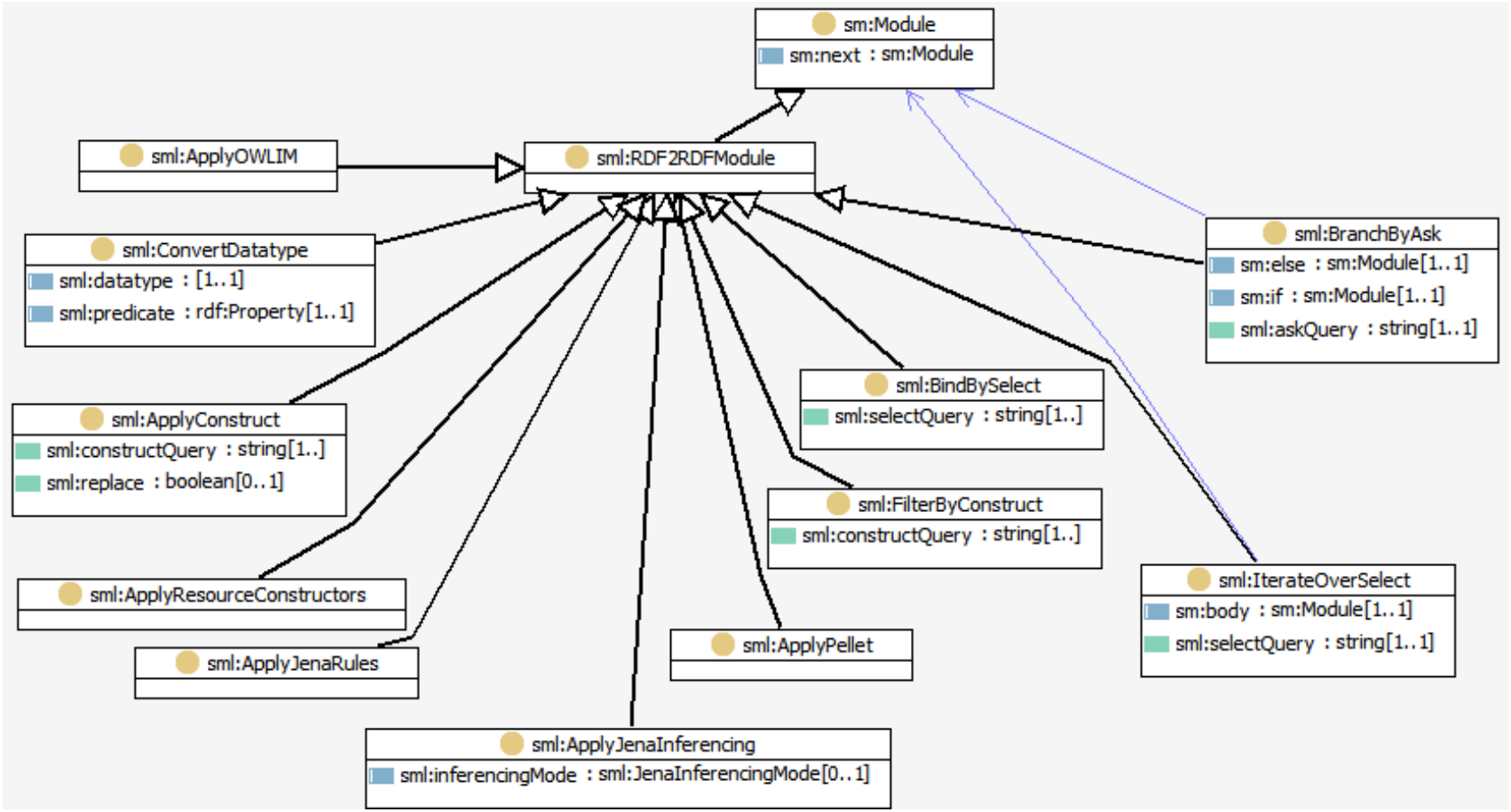
OWL Representation is backed by Java classes in execution engine
Module implementations are plug-ins to the engine

SPARQLMotion Module Library (2)



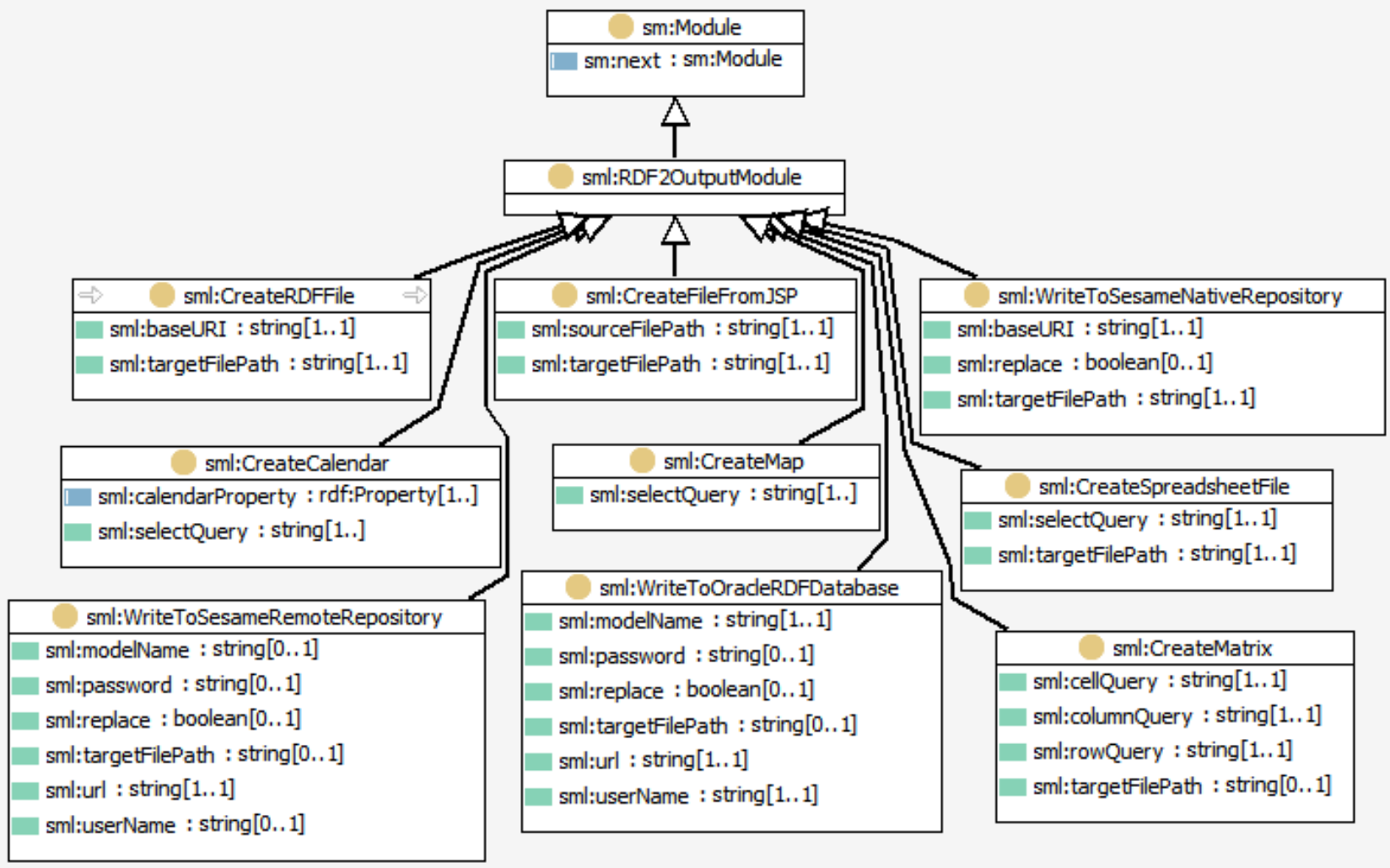
TopQuadrant™

SPARQLMotion Module Library (3)



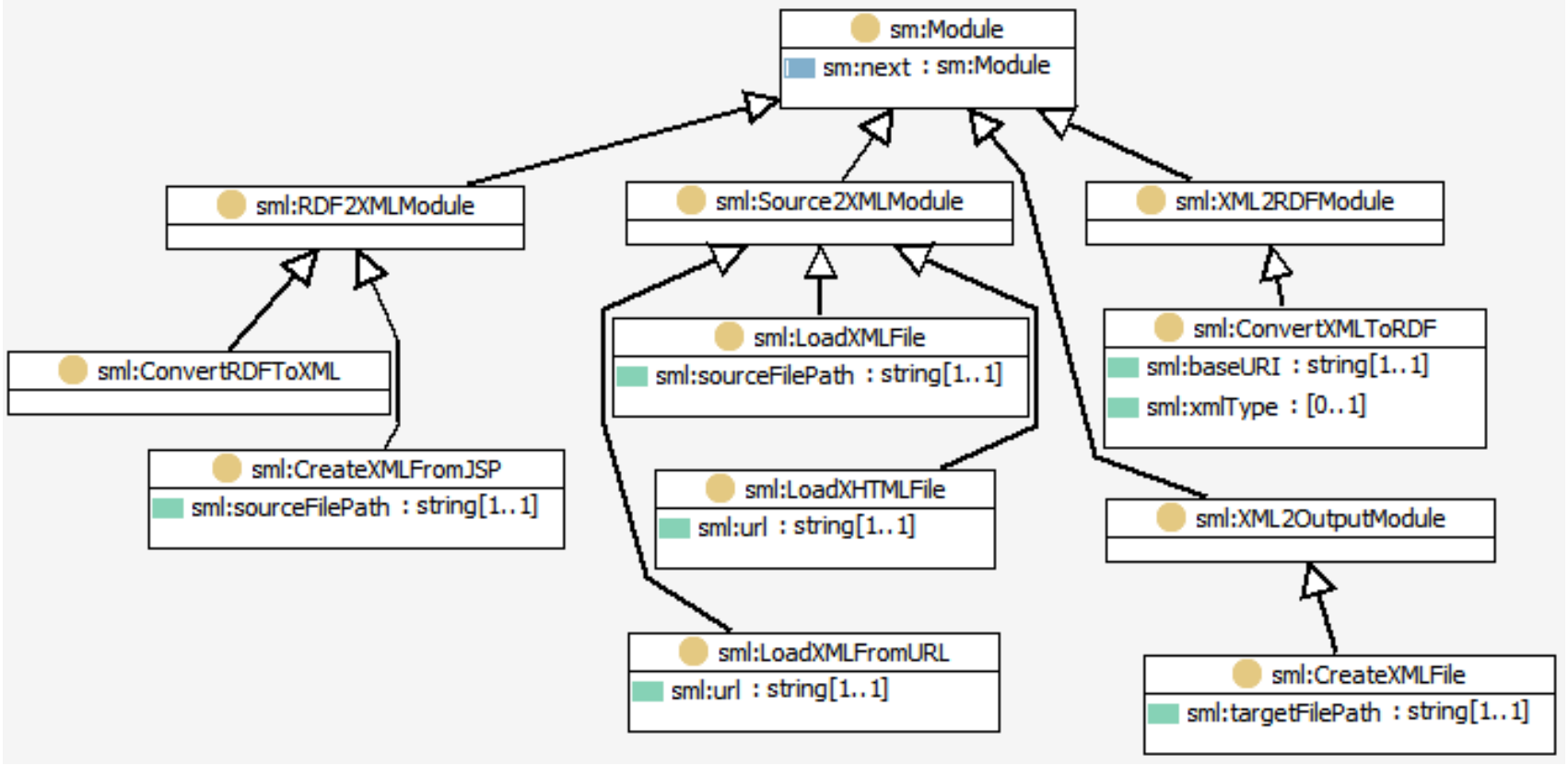
TopQuadrant™

SPARQLMotion Module Library (4)



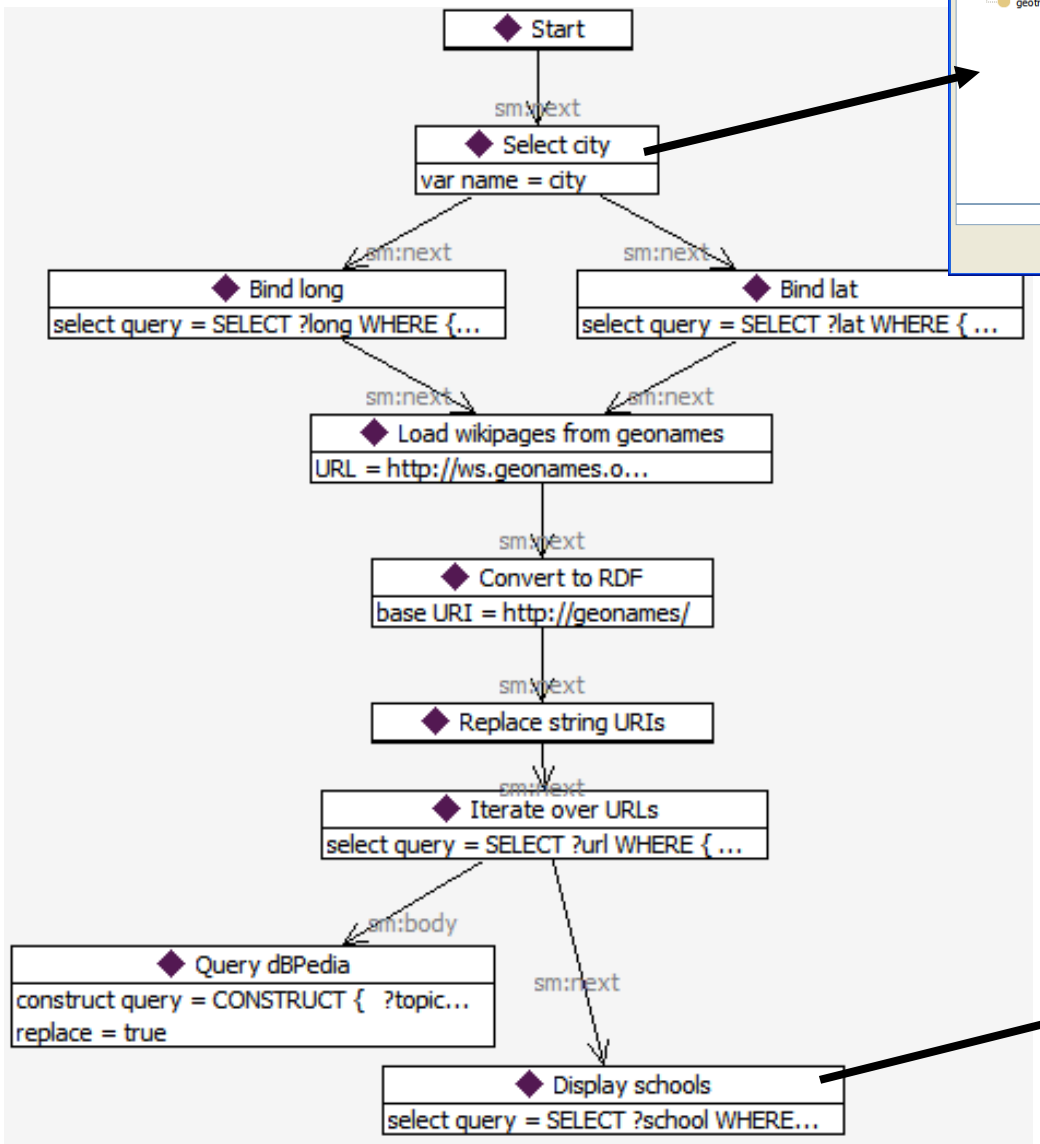
TopQuadrant™

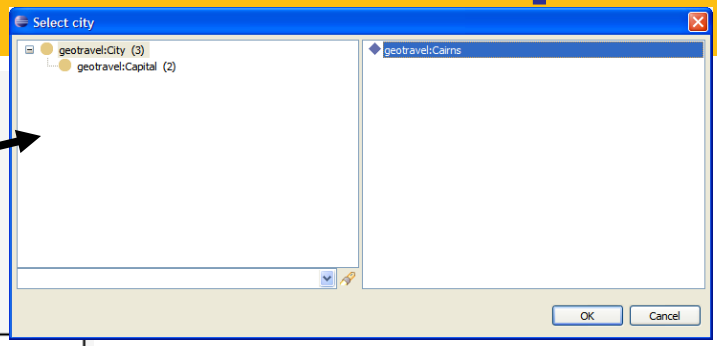
SPARQLMotion Module Library (5)

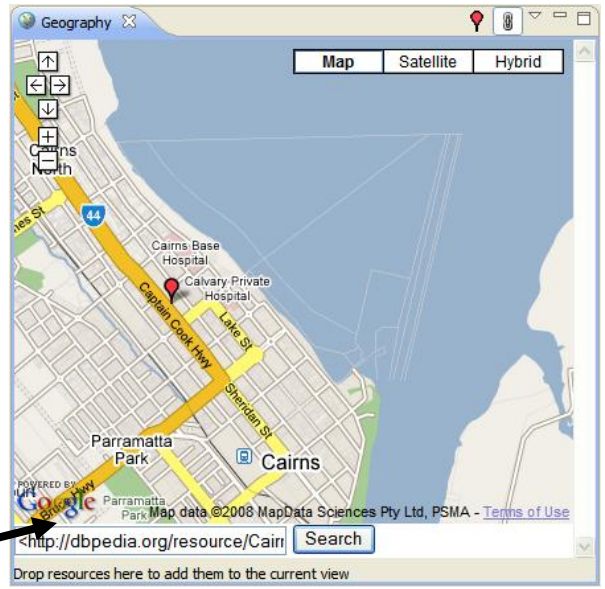


TopQuadrant™

Complex SPARQLMotion Example







TopQuadrant™

SPARQLMotion Use Cases

- Convert files to databases
- Combine multiple RSS feeds
- Create spreadsheets and charts
- Run periodic background checks
- Create XML input for other tools
- Control web pages
- Create maps and calendars
- Run inferences periodically
- …

TopQuadrant™

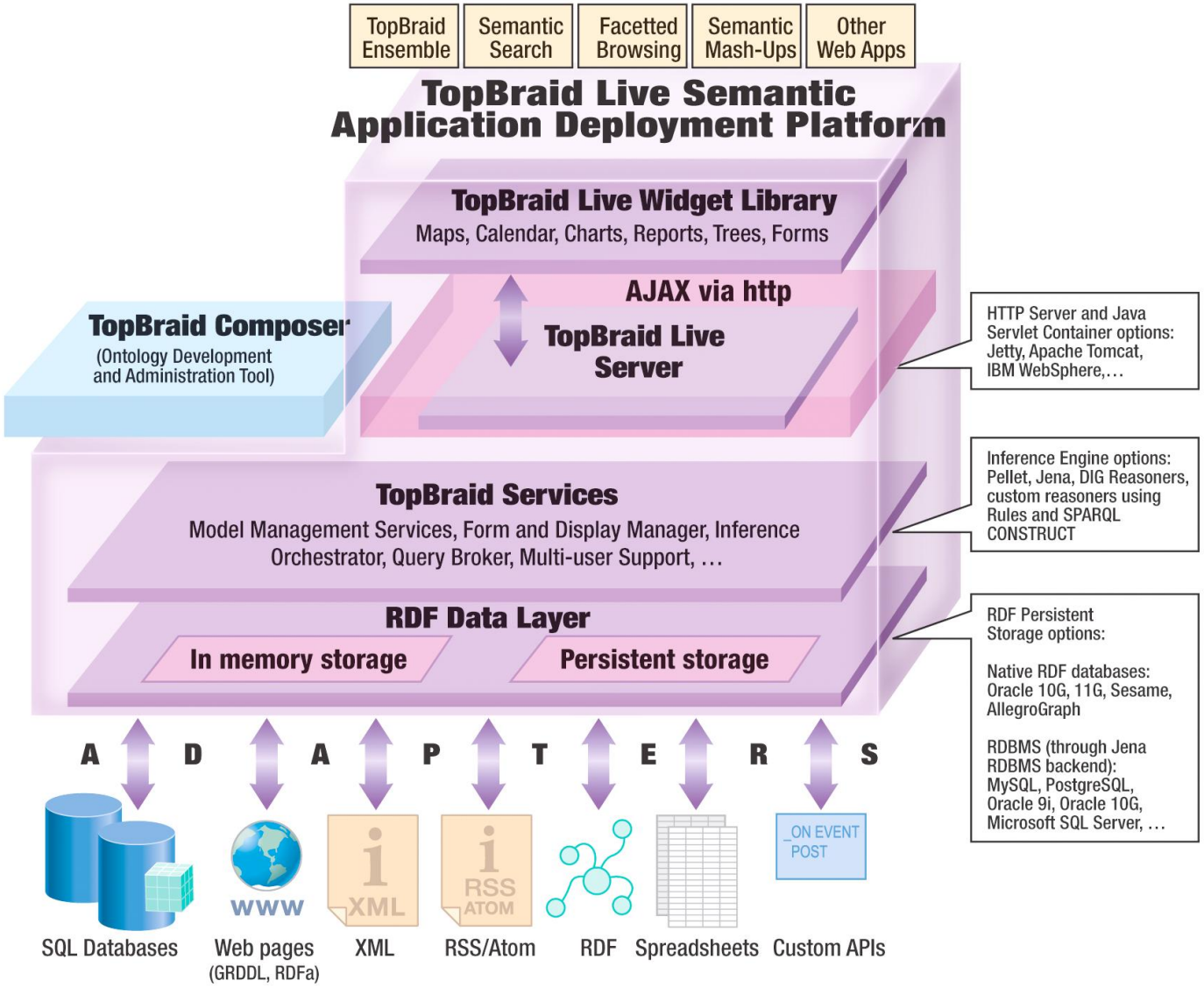
Summary

- Semantic Web languages are an attractive foundation for data integration tasks
- Generic methods and tools can be used, exploiting ontological metadata
- The TopBraid Suite product family is a comprehensive solution covering import, processing and export.

TopQuadrant™

Extra Slides

TopBraid Suite



TopQuadrant™