"Semantic Data Virtualization: Extracting More Value from Data Silos"

Featuring Syngenta's report on its successful pilot
Webcast Agenda

• Overview of Problem and Solution:
  – Cost and complexity of distributed data integration
  – Semantic virtual data integration

• Syngenta Report on its Successful Pilot

• Approach:
  – Why Semantic Technologies are uniquely suited to solving the Problem
  – TopBraid Insight provides a platform for this approach

• Discussion and Questions
The Context: Data Distribution

An Information Tsunami

- Increasing volumes of data of multiple types from multiple sources

combined with

An Informatics Tower of Babel

- Distributed systems using heterogeneous technologies
The Problem (Domain Experts’ perspective)

- Domain experts know that data exists
  - Within an organization
  - In external data sources

...and know both *actual* and *potential* questions that could be asked of the data *and this is a big IF*...

...the data could be integrated.
Find all the 3-way valves across all vehicles that correspond to the valves in this vehicle that are showing intermittent malfunctions at this point in checkout since we changed to this new supplier and the associated change orders and work authorizations.

Example (the problem is not domain-specific)

Semantic technologies allow the meaning of data to be expressed so that data can be related across databases with different schemas.
*Re-purposing* of data requires separation of context from data/meta-data, i.e. explicit representation of context-specific meta-data.
The Solution: (IT Technologists’ perspective)

Physical Data Integration - Technology-centric

• Proven tools and technologies (data warehouse, data mart, commercial solutions, etc.)
• Time consuming, expensive

Virtual Data Integration

• Lower cost?
• Technology dependencies?
• Time?
Virtual Data Integration

The Semantic Technologies Solution:

- Based on an integrated set of W3C standards

...that expose the meaning (semantics) of distributed data in a technology-independent representation (RDF)

...and use SPARQL (semantic query standard) to run federated queries over the distributed meta-data/data.
Why Semantic Virtual Data Integration?

- Lower cost
- More user /domain-expert-centric – Less technology centric
- Enables Deeper insight into data
- Less development time
- Increased flexibility and **evolutionary capabilities**

Domain experts can ask “unexpected questions” after the integration has been done.
Syngenta’s Successful Pilot

• Working with TopQuadrant, Syngenta initiated an extensive pilot project that:
  – Speeded the retrieval of the desired information from weeks to days
  – Decreased the effort necessary to integrate data
  – Enabled answering questions that they were unable to answer before

• Syngenta’s requirements helped to motivate development of TopBraid Insight.

• Problem is important, pervasive and there is high value in addressing it

• Encouraging other organizations to initiate pilots
Extracting more value from data silos

Using the semantic web to link chemistry and biology for innovation

Tim Eyres
Syngenta is one of the world's leading companies with more than 27,000 employees in some 90 countries dedicated to our purpose: Bringing plant potential to life. Through world-class science, global reach and commitment to our customers we help to increase crop productivity, protect the environment and improve health and quality of life.
Crop Protection Research & Development Progression

- Synthesize
- Test
- Analyse
- Design
- Efficacy
- Safety

Classification: PUBLIC

$/cycle $$$$$/cycle
The Challenge – Expand along dimensions beyond potency and AI spectrum

1 Importance of this axis may differ by indication
2 Compatibility with new delivery technology, formulation breadth, seed care, germplasm, traits, etc.

Competency Questions of Potency & Spectrum

What **bioactivity** is known for small molecule ‘x’, or it’s similar compounds, targeting protein ‘p’?

Competency Questions of Registrability and Innovation

References to all **Published Documents** where Chemical Substances that are similar to Chemical Substance 's' with structure property 'sp' are **Subject Of Paper**.
The Linked Data Enterprise
(It’s strategy not just a single solution)

“… **Linked Data** describes a method of publishing structured data so that it can be interlinked and become more useful.”


- An organization in which the act of information creation is intimately coupled with the act of information sharing
- Information is produced and consumed in ways for specific business needs, but produced in a way that can be connected to other aspects of the enterprise
- Key roles for:
  - Metadata: Controlled vocabularies, Taxonomy, tagging, ontologies
  - Metadata governance
  - Project/System publication of data & concept models
  - Architectural [reuse] of existing information & meta-data
The Approach - Combine internal and external data using “Linked Data” technology

What **bioactivity** is known for small molecule ‘x’, or it’s similar compounds, targeting protein ‘p’?

Syngenta Scientists ask **Competency Questions** of the Data, phrased using “real world” terms from the Concept Model

**Concept models** describe R&D knowledge and the semantic vocabulary for Syngenta science

**Data Models** link the scientists’ “real world” terms to the raw data and information

**Data** sets, both **internal** and **external** can be easily added to the federation.
1. **Dream the ideal** – assume a world of effortlessly linked data at your finger tips.
2. **Brain storm** – what questions would you like to ask of the knowledge base? Try to be specific. What do you really mean?
3. **Assess** – remove duplicates and cluster similar questions.
4. **Categorise** – importance, needs to be answered today or in the future, routine or non-routine.
5. **Identify concepts** and the **properties** of those concepts.
Competency Question Example

Competency Question:
All the **ChemicalSubstances** (and their structure properties) that have activity (of type inhibition) against target 't' with an assay result where assayOrganism is a plant.

Concepts:
- ChemicalSubstance
- Target
- Assay
- Bioactivity
- AssayOrganism

Properties:
- hasSubstanceStructure
- activityType
- assayResult
Now the technical problem ...

What Chemical-Substances (and their structure properties) are we working on that have activity (of type inhibition) against target 't' with an assay result where assayOrganism is a plant and this substance is similar to any patented substance?
The ‘lift and shift’ solution

Internal Data Sources

- Bio activity of small molecules
- Compound catalogue
- Protein crystal structures
- Small molecule crystal structures
- Research documents

External/Public Data Sources

- ChemSpider
- Derwent Patent Index
The physical data warehouse solution

Internal Data Sources
- Bio activity of small molecules
- Compound catalogue
- Protein crystal structures
- Small molecule crystal structures
- Research documents

External/Public Data Sources
- ChEMBL
- Derwent Patent Index

ETL
Vision – Virtual Data Warehouse

Internal Data Sources

- Bio activity of small molecules
- Compound catalogue
- Protein crystal structures
- Small molecule crystal structures
- Research documents

External/Public Data Sources

- ChEMBL
- Derwent Patent Index

Map-Reduce

SQL

SPARQL
Challenges – Virtual Data Warehouse

Internal Data Sources

- Bio activity of small molecules
- Compound catalogue
- Protein crystal structures
- Small molecule crystal structures
- Research documents

External/Public Data Sources

- ChEMBL
- Derwent Patent Index
Actual - Virtual Data Warehouse

Internal Data Sources

- Bio activity of small molecules
- Compound catalogue
- Protein crystal structures
- Small molecule crystal structures
- Research documents

Internal Links

Map-Reduce

External/Public Data Sources

- ChEMBL
- Derwent Patent Index

MAP-REDUCE

File import

D2RQ

SPARQL

SQL

Syngenta
Outcome

- Federated Search succeeded as measured by its scientific criteria laid down in the project definition
  - R&D was able to ask **adhoc questions** of **Potency, Novelty and IP** across a broad set of internal and external data sets
  - Scientists can **pose questions using scientific terms**, without the need to understand the underlying data model

- Federated Search delivered
  - a **fully functional** “linked data” federation
  - **concept models** which are **reusable** by future BI or integration projects
Technical Proofs

- Linked Data can be used to **federate live** operation data sets

- New **datasets are simply added** by linking their data model to the scientific concept model

- Using **Competency Questions** ensures that:
  - The **right things are linked** and we do not end up boiling the ocean
  - **Business benefit can be evaluated** at the end of any data set integration

- **Federation** was able to make disparate data sources appear to the scientists as one **unified** data set.
What else in 2013

- Linking new data sets: weather, soil, topology, field trial, mechanistic crop modelling
- Broadening acceptance through reporting and visualization solutions

What next in 2014

- Extending in the genetic, chemical and environment domains
- Learning how to scale semantic solutions for enterprise wide usage
Acknowledgements

- The project team at Syngenta
- Our collaborators at TopQuadrant
- Our collaborators at ChEMBL
- Our collaborators at Thomson Reuters.
- The “community” on whose shoulders we stand
Data Integration is hard because of... 

- Multiple data sources and storage technologies
- Multiple data-collection contexts
- Multiple data-usage scenarios
Perspectives from a (converted) non-believer

- “Connecting the dots” means “machine understanding the semantics” of the data/meta-data being “connected”

- Data integration (by machines) means *sharing semantics*

  \[ \text{Two (or more) systems given same data/meta-data will produce the same results when performing the same function} \]

- There are no silver bullets or secret sauces...
Semantic technologies – What’s different?

• Semantics as a “first-class citizen”
  – *Use* of data (by domain experts) rather than *storage* or *transmission* of data (by technologists)

• Bypass the “non-semantic” barriers to data integration
  – Differences in relational models/tables or XML document hierarchies
  – Vendor-specific technologies/implementations

• Use of semantic technologies is *evolutionary*:
  – *Not* “rip and replace”
  – *Instead* “integrate and evolve”
Semantic Data Virtualization

- The “secret sauce” isn’t really secret
- It is based on the use of W3C semantic standards

  **However,** aligning semantics isn’t simple...

- TopBraid Insight has a UI and associated tools to make as easy as it can be...

- ...and when it’s done, it opens the door to insights that were previously difficult, time consuming, and expensive to achieve
TopBraid Insight (TBI)

Connect the dots $\rightarrow$ $\rightarrow$ New Insights from virtually integrated data

Interactive Exploration
- Federated queries
- Over distributed data
- Surface answers using W3C semantic standards
- Running on commodity internet technologies
  HTTP, URI, ReST, etc.
Key Problem-Solving Features

Benefits of this approach to semantic data virtualization:

- **Virtualized Data Storage** — Federates data from different sources. Resolves data variety. Eliminates data replication and ETL. High performance, flexible, open, scalable, architecture.

- **Simple change management** — Add new datasets as needed. Change or update consolidated schemas for new concepts and questions without changing underlying datasets.

- **Easily configured** — Consolidated schemas and datasources for different users without duplication of data or new ETL scripts.

- **Interactive Dynamic Exploration** — Explore and discover intuitively using Search, Query, Filter, Browse, Navigate, Visualize, Share.
TopBraid Insight™ - Today
The whole is greater than the sum of its parts

Aristotle

...Questions?
“A wonderful harmony is created when we join together the seemingly unconnected.”

- Heraclitus