Case Study:

# Extracting Value from Data Silos — **Syngenta Case Study**

## *Using a virtual semantic data warehouse to link chemistry and biology for innovation*

*"The resulting linked data from **this pilot was able to support the full range of action, spectrum and selectivity questions** in herbicide discovery that were presented as challenges. **Questions we were unable to answer before** without a significantly greater investment of time and resources. This was an excellent example of using a semantic web approach to link biological activity, patent and physical chemistry data to easily explore research questions."* **— Syngenta**

**PRODUCT**
TopBraid Insight — an "out of the box" virtual data warehouse that offers an agile, extensible approach to querying data from diverse data sources.

**BENEFITS**
- The ability to flexibly and quickly extend the data maps to cover as much or as little of data as needed.
- Rapid configuration of new data sources. Adding a new data-source takes less than a day. As a result, scientists can answer their questions much faster.
- Access to any data-source — structured or unstructured, internal or external leading to more reliable results of scientific investigations.
- Finding answers to unanticipated questions formulated by scientists as they explore and interact with data – unlike a traditional data warehouse that can only answer questions it has been specifically designed to answer.
- The ability to query diverse data without the need for data replication.
- On-demand creation of personal or team virtual data warehouses to support different investigations.
- Collaboration among colleagues around shared results of data exploration.
- A high performance, open architecture integration framework with enterprise scalability.
- An approach similar to that of Map-Reduce that avoids the distributed query problem and allows requests to be divided into tasks that can be resolved across data sources.
- Data emancipation – every data-source participating in the solution becomes a member of a linked data cloud and can be integrated in different ways, combinations and configurations.
- A solution based on Semantic Web Technology standards using the power of RDF, RDFS, SKOS and SPARQL to unify data for querying and aggregation of results.

### Background on Syngenta

Syngenta is one of the world's leading companies with more than 28,000 employees in over 90 countries dedicated to our purpose: Bringing plant potential to life. We are using our deep knowledge of agriculture to develop fully integrated offers on a global crop basis, combining our innovation in genetic and chemical solutions. The company contributes to meeting the growing global demand for food, feed and fuel and is committed to protecting the environment, promoting health and improving the quality of life.

### Challenges

In order to maximize the chances of success in developing and registering novel crop protection products, it is necessary to bring together biological and chemical information from both inside and outside of an organization. The integrated use of biological data can help eliminate false positive molecular candidates and improve the chances of finding the correct candidates for development.

***Working with TopQuadrant, Syngenta initiated a pilot project with the following key objectives:***

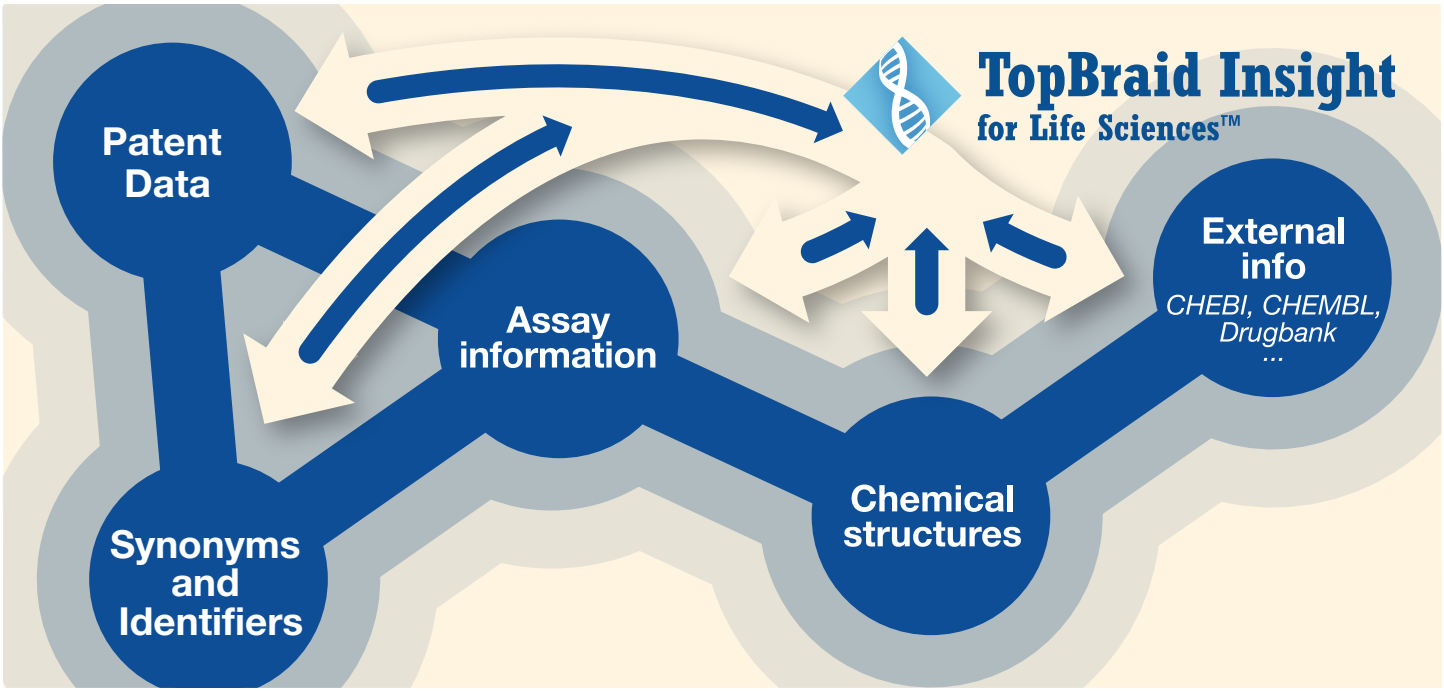
- Increase efficiency in finding a candidate substance, reducing the time and effort for the scientists.

TopQuadrant™

TopQuadrant's standards-based solutions enable a semantic ecosystem among people, applications and data—lowering the cost of ownership and enabling intelligent, data-driven action. TopBraid Insight connects data by accessing, linking and combining internal and external data sources faster, better, more broadly and in a more future-proof and flexible way than conventional data integration products.

For more information visit www.topquadrant.com, or contact us at tbi-info@topquadrant.com or by phone at +1 919 300 7945.

Ask us about scheduling a demo to explore how TopBraid Insight meets your specific requirements.




- Improve the quality of generated candidates and candidate evaluation by providing the scientists with the ability to access, explore and analyze available data earlier in the research cycle.
- Improve access to the internal and external data sources by linking them and providing a uniform way of accessing them.
- Improve capability of understanding the associated risks and opportunities for a research project and thereby increasing Syngenta's chances of developing the right crop protection product for the market.
- Achieve the correct balance between human and environmental safety and efficacy in order to achieve high quality, broad registrations.

To meet these objectives, a successful system was developed and piloted in 2013. Syngenta's requirements have motivated development of TopQuadrant's newest product — TopBraid Insight.

## Solution

A virtual semantic data warehouse was created to merge many of these internal and external data silos so that researchers could ask questions that would draw answers from all of these sources. TopQuadrant's first step was to add adapters to all data-sources so that the contents of these databases would appear as RDF triples. RDF is an industry standard designed to facilitate data merging even if the underlying data-source schemas differ, and to support the evolution of data schemas over time without requiring all the data consumers to be changed. With the RDF data model all information appears as facts or subject-predicate-object triples. Facts connect to other facts forming a knowledge graph or RDF (linked data) cloud.

To allow the resulting RDF cloud to answer questions, TopQuadrant turned to a Map-Reduce like approach, a procedure in which you can take users' questions and use a map of data-sources to concepts and properties to decompose any question into sub-questions appropriate to each data source. These sub-questions are then asked of each data-source and the results are merged back (reduced) into the virtual data warehouse. Once a logical data warehouse is established, the results can be returned to the user in a conventional format displayed as a form, grid of results, exported to Excel, and so on.

Using this solution, a user can explore all cross-referenced information, expanding the contents of the virtual data warehouse much like a snowflake navigation through a physical data warehouse. This can continue as needed to provide the user with enough information to answer their questions, analyze information, form new hypothesis and explore them to reach conclusions. The user can also link to external sites. External data is treated just like internal data and is merged into the query results.

## Results

The approach has shown to be very valuable, both for integrating data sets and answering key scientific questions. The ability to apply semantic web technologies to a federated search approach has opened up new avenues in exploiting the large volumes of R&D data within Syngenta. This novel approach is being extended to other parts of Syngenta R&D.

Additionally, the TopQuadrant and Syngenta pilot team received a CEO Certificate of Recognition for the successful outcome of this project.