Extracting more value from data silos: using the semantic web to link chemistry and biology for innovation

Tim P. Eyres

Syngenta AG, Switzerland

Received 5 July 2013; **Accepted** 16 September 2013; **Published** 14 September 2013

Competing interests: the authors have declared that no competing interests exist.

Abstract

In order to maximize the chances of finding novel crop protection molecules, that are safe for humans and the environment, it is necessary to bring together biological and chemical information from both inside and outside of an organisation. The integrated use of biological data can help eliminate false positive molecular candidates and improve the chances of finding the correct candidates for development.Information about the biological activity of compounds is captured in disparate systems within Syngenta and in the public domain. This research showed how highly curated bioactivity data from ChEMBL was linked to the Syngenta corporate chemical catalogue, along with other Syngenta research data and commercial patents indexes, using the Resource Description Framework (RDF).

Motivation and Objectives

To maximise the chances of finding novel and safe crop protection molecules it is necessary to join biological and chemical information located inside and outside an organisation. The integrated use of biological data provides a more holistic description of the activity of molecules, and thus helps eliminate false positive molecular candidates and improve the chances of finding the correct candidates for development.

Information about the biological activity of compounds is stored in disparate systems both within Syngenta and in the public domain. This research shows how semantic technologies were used to join public domain bioactivity data with Syngenta corporate data and commercial patents indexes.

The resulting linked data was used to support mode of action, spectrum and selectivity competency questions used in herbicide discovery.

The key outcomes of the research included:

1. increased speed and decreased effort required to retrieve the desired information related to chemical substances of interest (from weeks to days);
2. improved quality of the results when compared to existing approaches leading to large savings in terms of efforts and business benefits, particularly:
   a. discovery of new insights which would otherwise be missed, potentially reducing the number of late stage candidate failures and increasing the likelihood of identifying successful projects early;
   b. avoiding duplicating research efforts.

Methods

Technology

Federated search is an information retrieval pattern (Shokouhi and Si, 2011) allowing the simultaneous search of multiple disparate content sources with one query. A single query request is distributed to the multiple databases in real time, and upon collection the results are arranged in a useful form prior to being presented back to the user.

Prior to this research Syngenta integrated R&D data through the use of data warehouse systems. Although the Data Warehouse pattern offers very fast reporting capabilities on the available data, the approach has *cost and time to market* challenges due to complex processes of extraction, transformation and loading. The rapid increase in data volumes seen in Life Sciences further compounds the integration challenge when keeping internal and external data up to date.

Semantic web technologies provide a novel approach to the federated search pattern. Specific technologies used in this work included TopBraid from Top Quadrant, D2RQ[1], RDF[2], XML[3], SPARQL[4] and Web browsers.

Data

The data shown in Table 1 was integrated.

[1] http://d2rq.org/
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/2008/REC-xml-20081126/
[4] http://www.w3.org/TR/rdf-sparql-query

Table 1. This table reports essential information for all integrated data sources.

Data type	Source	Provider	Integration technique
Compound identifiers and activity	Syngenta chemistry repository (internal)	Syngenta	D2RQ federation
Document metadata	Syngenta document repository (internal)	Syngenta	XML import
Protein crystal	Syngenta protein crystal repository (internal)	Syngenta	D2RQ federation
Small molecule activity	Syngenta small molecule repository (internal)	Syngenta	D2RQ federation
Patent metadata	Derwent world patent index (http://thomsonreuters.com/derwent-world-patents-index/)	Thomson Reuters	XML import
Compound identifiers, activity, target and toxicity	ChEMBL (https://www.ebi.ac.uk/chembl/)	EBI	SPARQL

Architecture

The solution faced two key technical challenges of distributed queries and data connectivity. Distributed queries were implemented through the exploitation of SPARQL inferencing following a map/reduce pattern. Data connectivity required a variety of patterns and technologies due to the connectivity offered by the source systems. This included native SPARQL access, D2RQ connectivity to SQL based systems and file export/import. The native SPARQL access proved to be the simplest with the file transfer being the most time and resource consuming. An overview of the Syngenta Federated Search Architecture is shown in Figure 1.

Measuring Benefits

The research benefit was measured by answering competency questions defined in terms of:

1. increased efficiency in lead finding, reducing the time and effort for the scientists;
2. improved access to the internal data sources by linking them and providing a uniform way of accessing them;
3. improving the quality of generated leads and lead evaluation by giving the scientists insights into the external databases and research data.

A selection of example questions is given below:

1. Which chemical structures that are inhibitors of a given enzyme have a potency (threshold) above the potency threshold?
2. What tox data is available for a given enzyme inhibitor?
3. What references have been published for this structure or similar structures?
4. Which species for the given enzyme have IC50 data on plant weed varieties?
5. Find the given enzyme binding site of structures similar to a particular enzyme inhibitor.
6. Which species and compounds have reported measured bioactivity values for a given enzyme?
7. What X-ray crystallographic data is available for a given enzyme?

A method to assess and measure benefits was defined. Several iterations were repeated over four identical phases. Iteration 1 was performed without the tool to establish a baseline. Subsequence iterations were performed with the tool to measure the benefits in comparison to the baseline.

Phase 1: working with the information and available tools to answer competency questions.

Phase 2: completion of an evaluation questionnaire capturing experiences.

Phase 3: collection and formatting of the answers.

Phase 4: open discussion on the results to agree a score per question.

The questionnaire measured the business benefits in the following categories:

1. General tool usability
2. Efficiency gains
3. Quality of results compared to the current ways of working.

Results and Discussion

The Federated Search implementation was tested by five scientists, biochemists and chemists. They had no prior experience of the tool and developed experience during the benefit measurement activity. Where applicable results were on the scale of 1-4 where 4 is the best. The over-

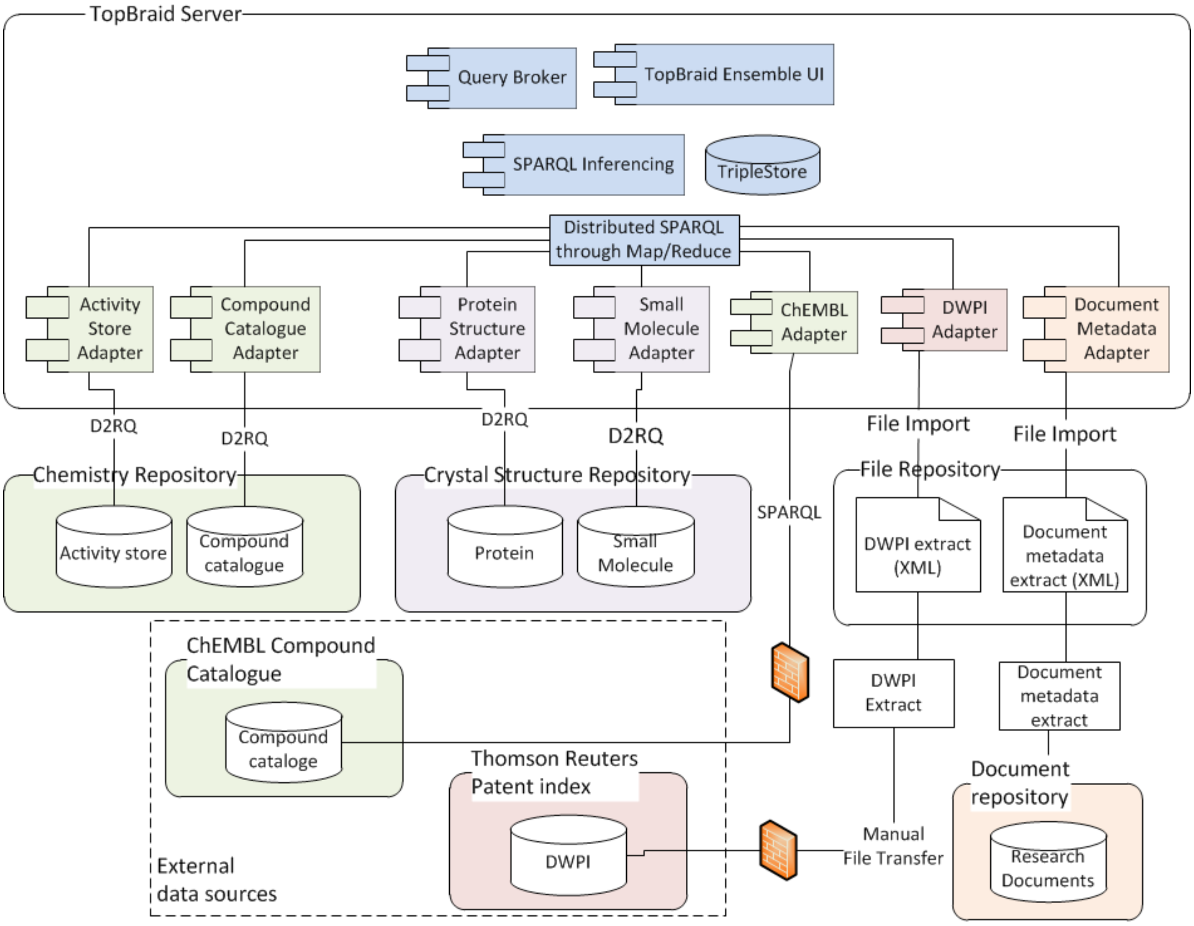

Figure 1. Syngenta Federated Search Architecture overview.

all scores were the final results after all iterations were complete.

How simple was the tool compared to current alternatives?
Overall score: 3. with the following caveats:

1. the tests have been done on a prototype of a Federated Search tool;
2. explanations were missing to form the query correctly;
3. scripted queries worked perfectly, but looking for other things was tricky.

Was there an increase in speed to complete the process from query to result compared to current alternatives?
The efficiency gains were in range of 1:6 for the tool. However, comparison against the baseline was difficult due to certain parts of the existing process being too time consuming without the tool.

The Federated Search solution was recognised as much simpler than the current alternatives.
The proposed Federated Search solution tool effectively replaces a longer manual process.

How does the quality of output compared to current alternatives?
Overall score: 4.

While the output was comprehensive, the presentation of the results made them difficult to analyse.

How was the level of detailed compared to current alternatives?
Overall score: 4.

Complicated search formulation for the x-ray related competency questions made the solution difficult to use with ad hoc queries.

Overall, the approach has been 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 R&D in Syngenta.

Acknowledgements

The project team at Syngenta. Our collaborators at ChEMBL. Our collaborators at Thomson Reuters.

References

Shokouhi M, Si L. (2011) Federated Search. *Foundations and Trends® in Information Retrieval* **5**(1), 1-102. doi:10.1561/1500000010