TopBraid Life Sciences Insight™

In the Life Sciences industries, making critical business decisions depends on having relevant information. However, queries often have to span multiple sources of information. In accessing, searching and using information, you may have needs or goals such as the following:

- increasing the efficiency of the target and lead drug discovery processes
- providing knowledgeable researchers, physicians and payers with credible information
- correlating data from clinical trials across multiple studies
- ensuring the right treatment to the right person at the right price

The quality of decisions that you and your customers will be making directly relates to the accuracy, completeness, timeliness and correctness of information. Sourcing, aligning and generating insight from aggregated information, at a given time and in a given context, are the key challenges.

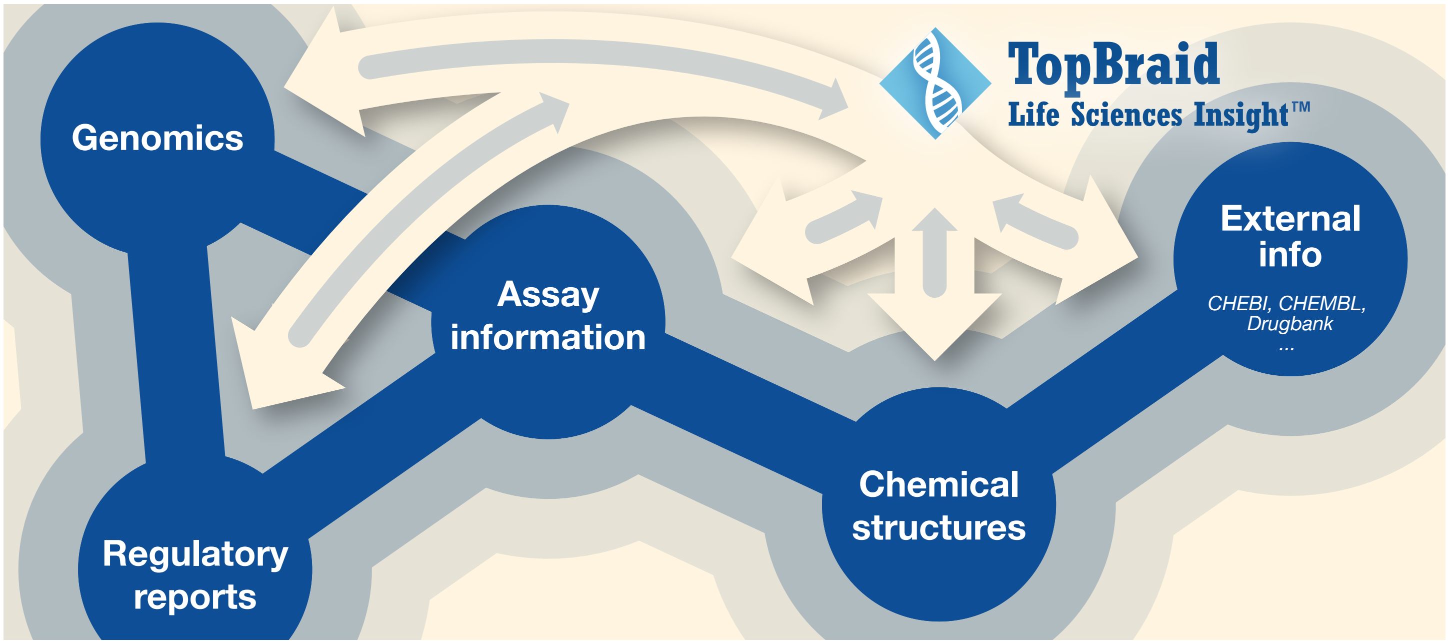


Figure 1: TopBraid Life Sciences Insight presents data from multiple datasets as if it were within a single data warehouse.

TopBraid Life Sciences Insight (TopBraid LSI) provides an "out of the box" Logical Data Warehouse that allows an agile, extensible approach to querying data from diverse data sources using a high performance open architecture that has proven enterprise scalability. TopBraid LSI decreases the time and cost for clinical trials and drug discovery by enabling diverse databases to appear as a single logical data space.

TopBraid LSI includes an upper ontology specific to Life Sciences and configuration of popular publicly available data sources such as CHEBI, CHEMBL SIDER, LinkedCT, DRUGBank and more. Customers' internal data sources and public data sources not within the pre-configured set are easily configured into the solution.

TopBraid LSI uses strategic technology standards to future-proof customers' investments while complementing existing infrastructure through an open architecture.

The Data Integration Challenge

Diverse data needs federated queries

The need to integrate data from multiple silos of information has been a problem plaguing businesses ever since databases first appeared. Commonly, a query that needs to access multiple sources must be mapped to a series of requests to distributed data sources. Brokering these queries, optimizing their execution, and aggregating their results are time-consuming and expensive tasks.

Many different approaches to integrating access to diverse data sources have been attempted. All are complex. Often, integration is provided by applications that can talk to one of several data sources, depending on the user's request.

Hard-wiring of data sources and data alignment results in poor maintainability

In these systems, the data sources are typically "hardwired"; replacing one data source with another means rewriting a portion of the application. In addition, data from different sources cannot be compared in response to a single request unless the comparison is likewise wired into the application.

Moving all relevant data to a warehouse allows greater flexibility in retrieving and comparing data, but at the cost of re-implementing or losing the specialized functions of the original source, as well as the cost of maintenance. Furthermore, each data warehouse is designed to answer predefined queries. Modifying or extending it to answer new questions is expensive and time-consuming.

Connecting data from multiple sources is a hard problem

Yet another approach to accessing data from multiple sources is to create a homogeneous object layer to encapsulate diverse sources. This encapsulation makes applications easier to write, and more extensible, but does not solve the problem of meaningfully connecting data from multiple sources.

Pros and Cons of Current Approaches

1: Build a Mega-application Database

Avoid the data-silo problem entirely by adopting a single vendor for all application systems. One vendor is responsible for merging all data.

Pros: In theory the integration problems go away.

Cons: It is unlikely that one vendor can provide a single application with sufficient scope to incorporate all business information.

Data silos still appear.

2: Use data warehouses to extract data from silos

Extract data from data silos, with transformation to a composite schema, then load into a data mart.

Pros: Query performance of the data within each data mart is good. OLAP tools are good at querying the data marts and performing data analysis.

Cons: Inadequate composite schema. New concepts require their own data mart. New data marts require a specialist to create ETL for the source datasets.

Loaded data is often aggregated, and access to original data is lost. Provenance of the data is also usually lost.

This encourages and often drives the creation of multiple copies of data. The advent of big data, with its large and growing data volumes, argues strongly against duplication of data. Since data is replicated it is never as up to date as the actual data.

3: Perform SQL distributed querying

RDBMS vendors, and others, offer the concept of distributed querying. A common query engine resolves the query by visiting all connected databases.

Pros: The federated query engine uses SQL, which is well supported. Since data is federated the overhead of replication is avoided.

Cons: Federated query performance is often very poor as large quantities of data are fetched.

The distributed query has to resolve the mismatch of schemas, resulting in complex federated queries.

As more datasets are added, the distributed query performance is likely to degrade geometrically.

A universal enterprise schema must be created—often a lengthy process—prior to the use of distributed queries. Once this is created, it is difficult to adapt to changing requirements or to add additional concepts introduced from new data sources.

Advantages of a Logical Data Warehouse Approach

TopBraid LSI is an Integration Framework in the form of a Logical or Virtual Data Warehouse, an approach that is emerging in response to problems with warehousing and distributed querying. Instead of creating an actual data warehouse, TopBraid LSI presents data from multiple datasets as if it were within a single data warehouse. Different virtual data warehouse models can be tailored to different users without duplication of data.

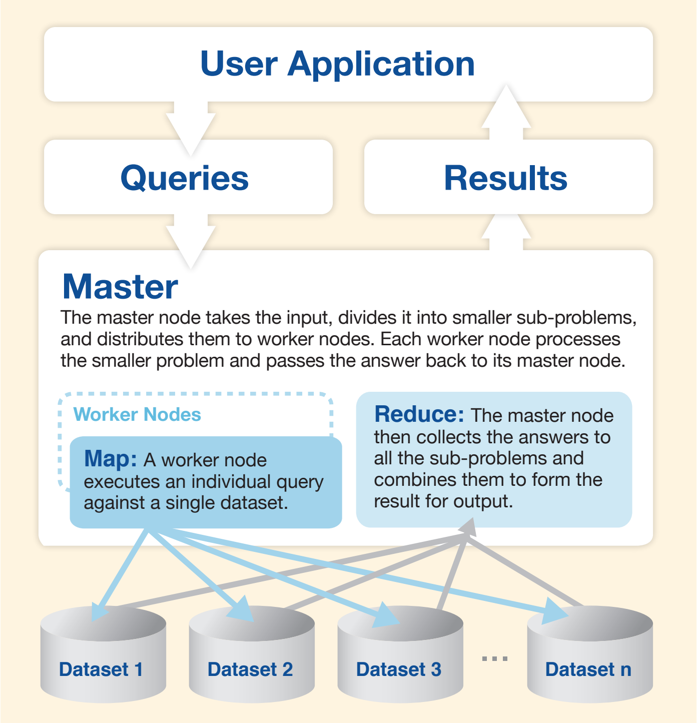

User Application
Queries
Results
Master
The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. Each worker node processes the smaller problem and passes the answer back to its master node.
Worker Nodes
Map: A worker node executes an individual query against a single dataset.
Reduce: The master node then collects the answers to all the sub-problems and combines them to form the result for output.
Dataset 1
Dataset 2
Dataset 3
...
Dataset n

Figure 2: Illustration of query execution using Map-Reduce type strategy over federated data

The approach adopted by the TopBraid LSI integration framework is similar to that of Map-Reduce. To avoid the distributed query problem, each request for data is divided into tasks (queries) that can be resolved by a single data source. The worker nodes may do this division of tasks again based on data that is returned leading to a multi-level tree structure. The data is then re-assembled into the result set that was required by the original query. Since the data is federated, the overhead of replication is avoided. It is simple to incrementally add new datasets. Unlike a data mart, the consolidated schema can be easily changed or updated to reflect new concepts without the need to change any underlying datasets.

According to Gartner[1] "…the Logical Data Warehouse (LDW) is a new data management architecture for analytics which combines the strengths of traditional repository warehouses with alternative data management and access strategy."

Query performance is excellent as the Map-Reduce approach only fetches the data from sources that are required. Expensive cross-database joins are

[1] *Gartner Hype Cycle for Information Infrastructure, 2012* - http://www.gartner.com/id=2101415

***How Gartner re-defines the Enterprise Data Warehouse for the Era of Big Data*[2]**

The new type of warehouse—the Logical Data Warehouse (LDW)—is an information management and access engine that takes an architectural approach which de-emphasizes repositories in favor of new guidelines:

- **The LDW follows a semantic directive to orchestrate the consolidation and sharing of information assets, as opposed to one that focuses exclusively on storing integrated datasets.**
- **The semantics are described by governance rules from data creation and business use case processes in a data management layer, instead of via a negotiated, static transformation process located within individual tools or platforms.**
- **Integration leverages both steady-state data assets in repositories and services in a flexible, audited model via the best available optimization and comprehension solution available.**

avoided. Provenance of individual statements is retained. Since data need not be replicated there are no concurrency problems nor large data storage problems. The TopBraid LSI solution, based on Semantic Web technologies, uses the power of RDF and SPARQL to unify data for querying and aggregation of results.

Traditional data warehouses have limited dimensions and measures. In comparison, TopBraid LSI has been proven to support more than 100 dimensions and measures. Using TopBraid LSI allows users to drill down through dimensions and measures into other, more complex, data structures. Moreover, unlike traditional data warehousing, the approach used for Topbraid LSI makes it possible to dynamically refactor the model with no further Extract, Transform and Load (ETL) operations.

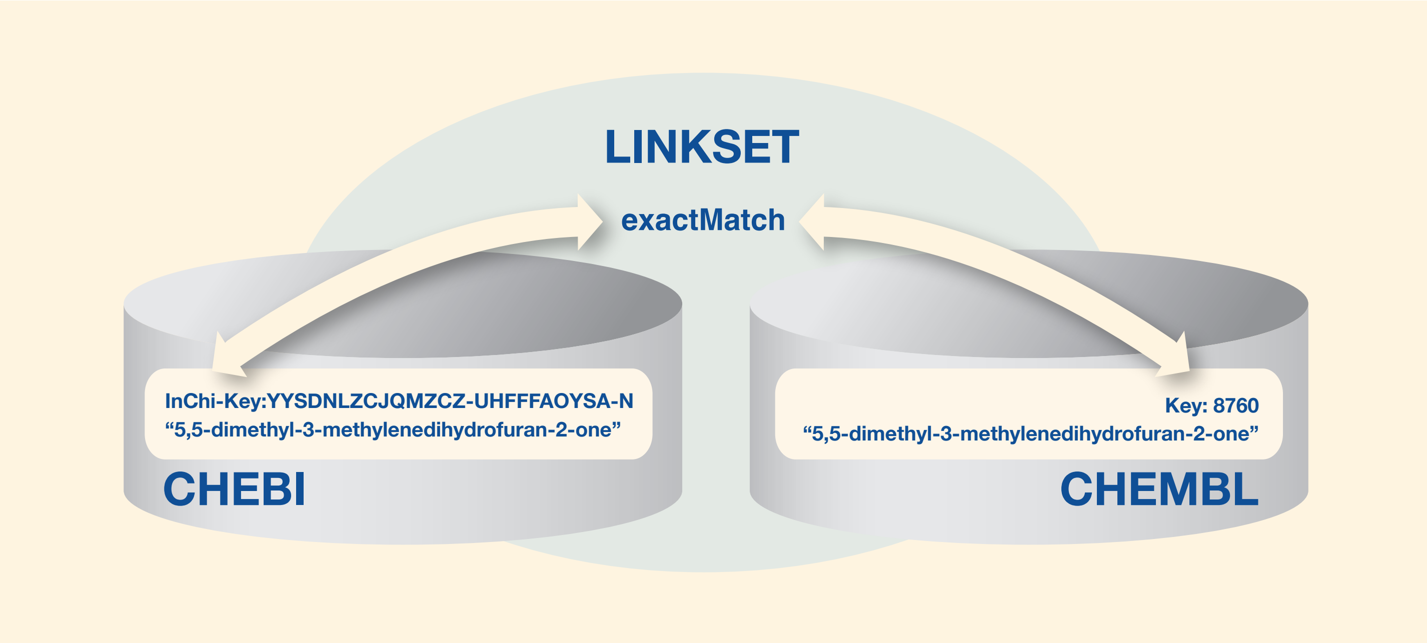

LINKSET
exactMatch
InChi-Key:YYSDNLZCJQMZCZ-UHFFFAOYSA-N
“5,5-dimethyl-3-methylenedihydrofuran-2-one”
CHEBI
Key: 8760
“5,5-dimethyl-3-methylenedihydrofuran-2-one”
CHEMBL

Figure 3: Illustration of a linkset that maps compounds from CHEBI to compounds from CHEMBL, showing just one example mapping within the linkset

[2] Does the 21st-Century “Big Data” Warehouse Mean the End of the Enterprise Data Warehouse? -http://www.gartner.com/id=1775719

TopBraid LSI is Easy to Configure and Use

Configuring data sources for use in the logical data warehouse is simple and consists of the following steps:

1: Configure a new data source

A new data source is identified to the system by registering it. Federation is not limited to RDBMS datasets, but includes other formats such as XML, spreadsheets, Linked Open Data and web services. TopBraid LSI examines the data source and auto-populates its schema.

2: Configure concepts within a data source and define how the data within it links to other sources

Relevant concepts and their properties from the data source are mapped to existing concepts and properties in the upper ontology (so-called universal concepts). If they are new concepts or properties, you can easily add them to the upper ontology using a simple user interface. Additionally, properties to be used for keyword searching are identified. Since this mapping is virtual, it is easy to change it as requirements evolve, or even change the upper ontology universal concepts as the scope changes. These modifications do not require re-extracting data and re-populating warehouses.

This step also defines how identifier keys in the new data source are related to identifier keys in other sources. Such relationship information is captured using VoID[3] linksets (see Figure 3).

A linkset provides a mapping from data entities in one dataset to another. Since multiple datasets might share the same data items, mappings can grow geometrically. To avoid the geometric growth, TopBraid LSI dynamically creates a chain of linksets. As a result, if data source A is linked to data source B and data source B is linked to data source C, data sources A and C become automatically linked without needing any additional configuration steps. This way, the number of linksets only grows linearly as new datasets and concepts are added. See Figure 4, illustrating that the brown dataset (lower left) is linked to the dark green dataset (upper right) even though there is no direct arrow (linkset) between them.

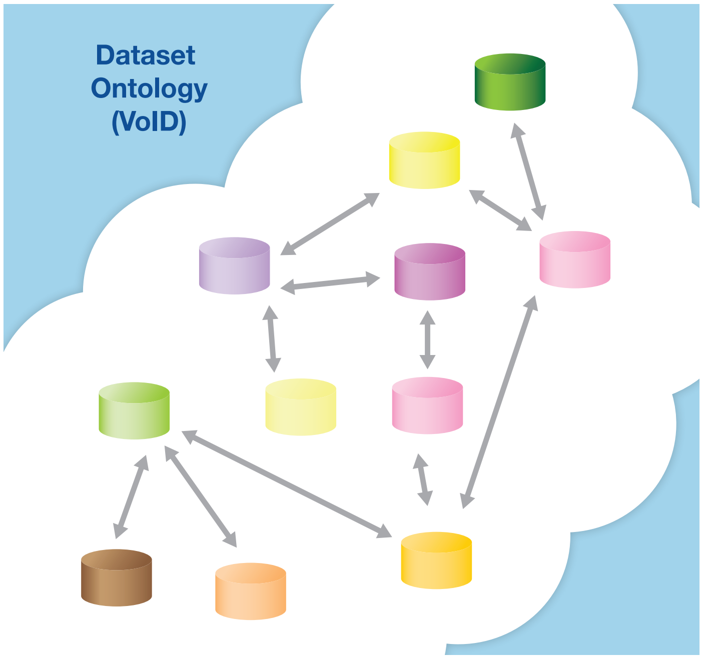


Figure 4: Linksets for multiple data sources

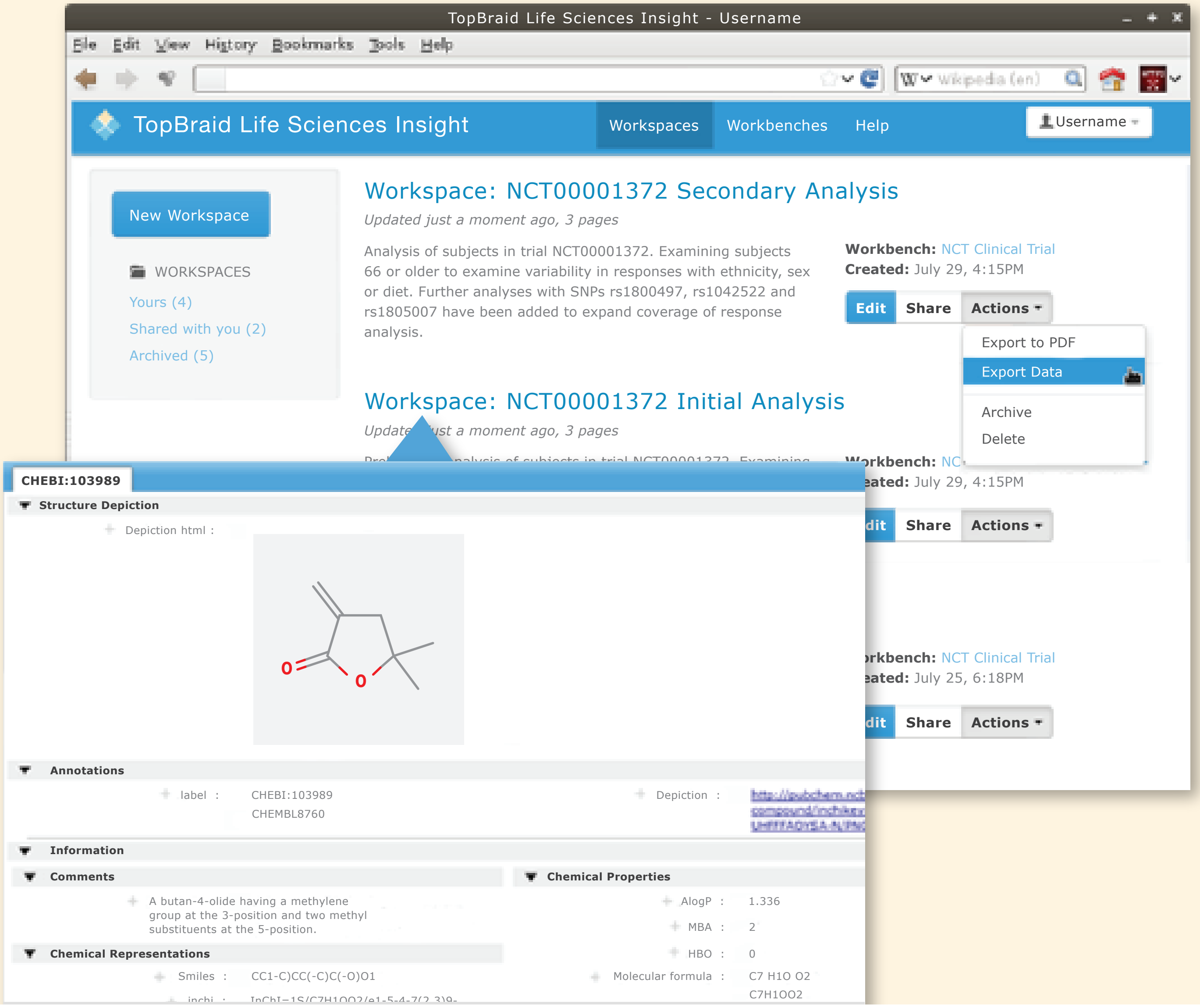
3: Create new or update existing workbenches

A workbench consists of a package of datasets and selected concepts and properties with their mappings. This way, the same artifacts can be used for various search and analysis purposes in different workbenches and workspaces.

Workbenches can be thought of as templates used to create workspaces shared by one or more users to perform research and data analysis. A workspace is an instance of a workbench populated by the retrieved and discovered data. Figure 5 illustrates how a user can create and access workspaces within TopBraid LSI. A workspace can then be used to retrieve, view, search and make use of diverse data. Workspaces can be saved across user sessions, so that users may go back to add more data by running new queries.

Users can also create and share worksurfaces: views onto the workspace that display the retrieved information in the way that meets users' unique requirements.

[3] VoID (from "Vocabulary of Interlinked Datasets") is an RDF based schema to describe linked datasets. For more information, see http://semanticweb.org/wiki/VoID



Workspace: NCT00001372 Secondary Analysis

Updated just a moment ago, 3 pages

Analysis of subjects in trial NCT00001372. Examining subjects 66 or older to examine variability in responses with ethnicity, sex or diet. Further analyses with SNPs rs1800497, rs1042522 and rs1805007 have been added to expand coverage of response analysis.

Workbench: NCT Clinical Trial
Created: July 29, 4:15PM

Workspace: NCT00001372 Initial Analysis

Update ust a moment ago, 3 pages

Workbench: NC
eated: July 29, 4:15PM

rkbench: NCT Clinical Trial
eated: July 25, 6:18PM

Depiction html :

label : CHEBI:103989
CHEMBL8760

Depiction : http://pubchem.ncb compound/inchikey UHFFFAOYSA-N/PNG

A butan-4-olide having a methylene group at the 3-position and two methyl substituents at the 5-position.

AlogP : 1.336
MBA : 2
HBO : 0

Smiles : CC1-C)CC(-C)C(-O)O1
inchi : InChI=1S/C7H10O2/c1-5-4-7(2,3)9-

Molecular formula : C7 H1O O2
C7H1OO2

Figure 5: illustration of a selected workspace in use based on the Workbench NCT Clinical Trial.
The example screen CHEBI 103989 *shows some data retrieved within that workspace.*

In Summary

This white paper provides a high level overview of how **TopBraid Life Sciences Insight** delivers a Logical Data Warehouse: an agile, extensible approach to querying data from diverse Life Sciences sources in a high performance, open architecture integration framework with proven enterprise scalability.

TopQuadrant has plans to offer additional domain-specific logical data warehouse solutions including TopBraid Agro Sciences Insight (TopBraid ASI), and TopBraid Oil & Gas Insight (TopBraid OGI).

To find out more, contact us at sales@topquadrant.com.

Create a semantic ecosystem,
the next step in data evolution.

About 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. Our TopBraid™ platform 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.

TopQuadrant™

For more information visit www.topquadrant.com, or contact us
at sales@topquadrant.com or by phone at +1 703 299 9330.