

ARTICLES

ArQiologist: An Integrated Decision Support Tool for Lead Optimization

Atipat Rojnuckarin,* Daniel A. Gschwend, Sergio H. Rotstein, and David S. Hartsough

ArQule Inc., 19 Presidential Way, Woburn, Massachusetts 01801

Received April 6, 2004

This paper describes ArQiologist, a Web-based tool that integrates chemical, analytical, biological, and computational data to facilitate decision support for lead optimization at ArQule. It features an easy-to-use graphical query builder that allows queries to be saved, reused, and shared by researchers. Query results can be viewed with built-in data browsers or exported with structures to external applications such as Microsoft Excel or Spotfire for further analysis.

INTRODUCTION

The discovery of drugs requires balancing many properties of candidate compounds in order to attain the desired potency, selectivity, and pharmacokinetic profile. Often, the relevant property data reside in heterogeneous data storage systems developed either internally or by different commercial vendors. This is one of the main difficulties faced in the design of decision support tools.¹ Larger organizations additionally often need to support legacy database systems and/or deploy different systems for different discovery areas or geographical locations.² While vendor-supplied database systems generally include data access applications, researchers often need to spend considerable amounts of time manually consolidating the data that they need to make decisions. Consequently, an integrated decision support tool that can retrieve data from multiple heterogeneous data sources and present them to the users in a consistent manner is very valuable to an organization.

A well-designed decision support system needs to allow users to ask questions ad hoc. This requirement presents system designers with a series of challenges relating to efficiency.³ Despite recent advances in relational database systems and computing power, an inappropriate execution plan can keep a simple query, which potentially completes in seconds, running for hours. While commercial query tools are available, they tend to either work only with a particular vendor's database systems (e.g. IDBS DiscoveryChannel,⁴ Accelrys AccordHTS⁵) or seriously restrict the ability to customize how queries are constructed and optimized (e.g. ISIS/Base,⁶ DayBase,⁷ ChemCart⁸). Our inability to find a suitable commercial solution combined with our familiarity with our various data sources led us to pursue the development of an efficient and cost-effective decision support tool internally.^{1,9} The resulting tool, ArQiologist, was initially released in 2001. Since then, it has become the primary query tool for internal research data at ArQule. This paper describes in detail how ArQiologist organizes and presents data, its user interface, the data sources that are integrated by the tool, and its architecture.

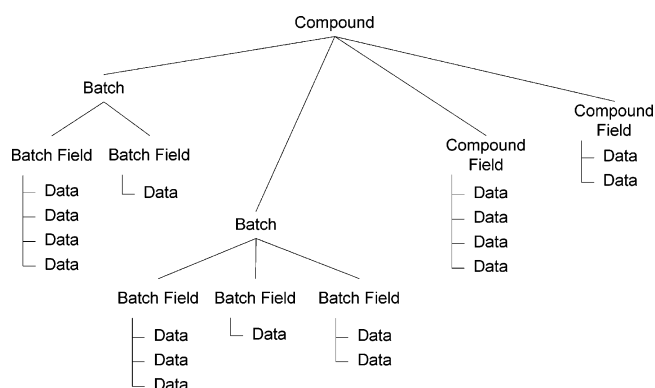


Figure 1. Data structure.

DATA ORGANIZATION

ArQiologist organizes data into a compound-centric hierarchy (Figure 1). A *compound* is defined as a unique parent chemical structure. A *batch* is defined as a unique synthesis of the parent chemical structure. Different salt forms of the same parent chemical structures are considered the same compound but different batches. Since different syntheses may result in different salt forms with different types and amounts of impurities, it is typically desirable to separate assay data performed on different batches. In Figure 1, batch-independent data fields (e.g. predicted properties, descriptors) are labeled “compound field”. Data fields that can vary in value from batch to batch are labeled “batch field”. Either of these field types can have multiple values.

When performing a search, ArQiologist first determines which compounds and batches satisfy the query and then it returns data associated with them. The system is therefore less suitable for the retrieval of data such as an individual experimental run. If users specify plate barcodes as a query criterion, ArQiologist looks up the compound batches associated with the specified barcodes and returns *all* the data for those batches regardless of whether these data were collected from the specified bar-coded plates. This design results in a more global view of compound data and is therefore better suited for decision support.

While most commercial query tools manipulate data in terms of rows and columns, the inherent hierarchical nature

* Corresponding author e-mail: arojnuckarin@arqule.com.

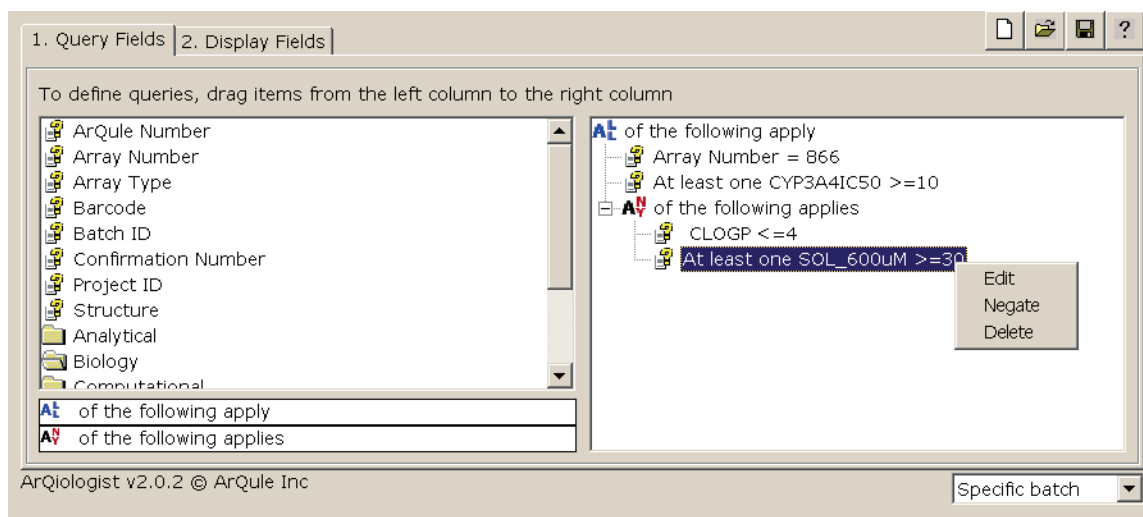


Figure 2. Query builder interface first tab (query definition).

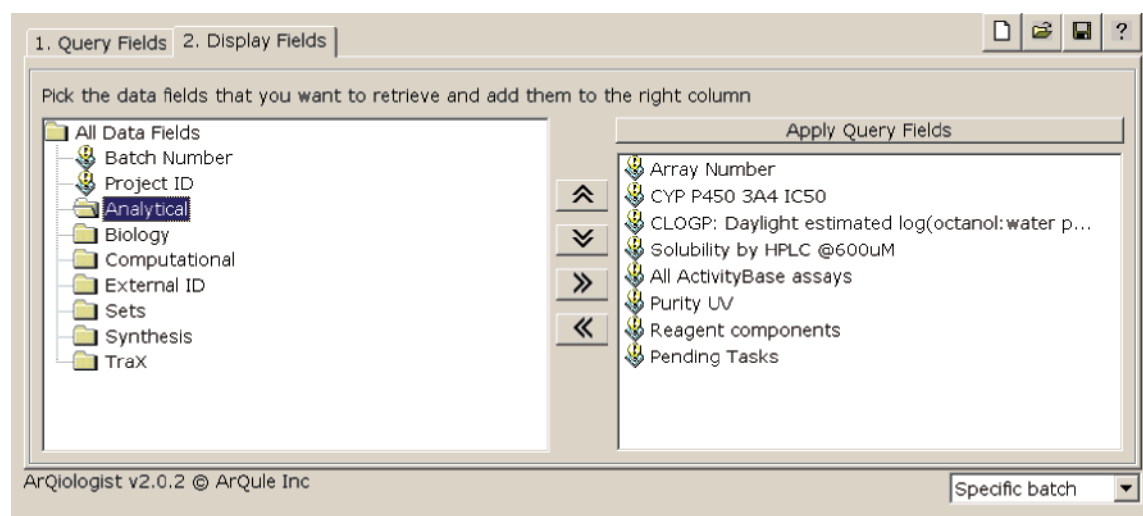


Figure 3. Query builder interface second tab (field selection).

of ArQIologist data allows it to deal effectively with multiple batches per compound and multiple data points per batch. This data hierarchy is maintained throughout the application. However, when users want to visualize the data in external applications that only accept tabular data, they are given a rich set of options for the aggregation of the data before exporting. The main disadvantage of the hierarchical data structure is the increased complexity. For example, if users want to find compounds whose inhibition against enzyme X is greater than 80%, they must decide whether they want to retrieve compounds whose inhibition of target X was greater than 80% (a) *at least once*, (b) *in all occasions*, or (c) *on average*. In lead-optimization, option (a) is usually the one of choice, as it returns more data and ensures that the user will not accidentally miss any potentially interesting results.

INTERFACE

ArQIologist is a Web-based application. Users start at a home page, in which they can administer (e.g. save, manage, and share) queries, launch the Query Builder, set a number of preferences, and find out what is new with the system.

The query builder is an ActiveX component that consists of two tabs: The *Query Fields* tab (Figure 2) allows users to define search criteria. The *Display Fields* tab (Figure 3)

allows users to define data display criteria. Each tab consists of two panels; the left panel contains palettes of available query/display items that users can select by moving them to the right panel, either by double clicking or dragging. There are also buttons on the top left corner that let users save queries to and read queries from files.

Query and display items are arranged into a folder hierarchy that corresponds to functional areas or disciplines. The most commonly used items appear on the top level. Users drag the desired query items to the right panel to assemble their search criteria. After dropping the query item into the right panel, a dialogue box pops up to allow users to further qualify the item. The type of dialogue box displayed depends on the query item (Figure 4). For example, if the user is querying by barcode, the dialogue allows him to enter or load the desired barcodes. If the user is searching for data from a particular assay, the dialogue allows him to specify cutoffs (e.g. percent inhibition, IC50). Criteria can be combined using either "ALL" or "ANY" operators (Figure 2), corresponding respectively to "AND" logic and "OR" logic. Criteria can further be negated, edited, or deleted through right-click menus. The Display Fields tab also has a button labeled "Apply Query Fields", a convenient shortcut to reproduce the fields selected in the Query Fields tab.

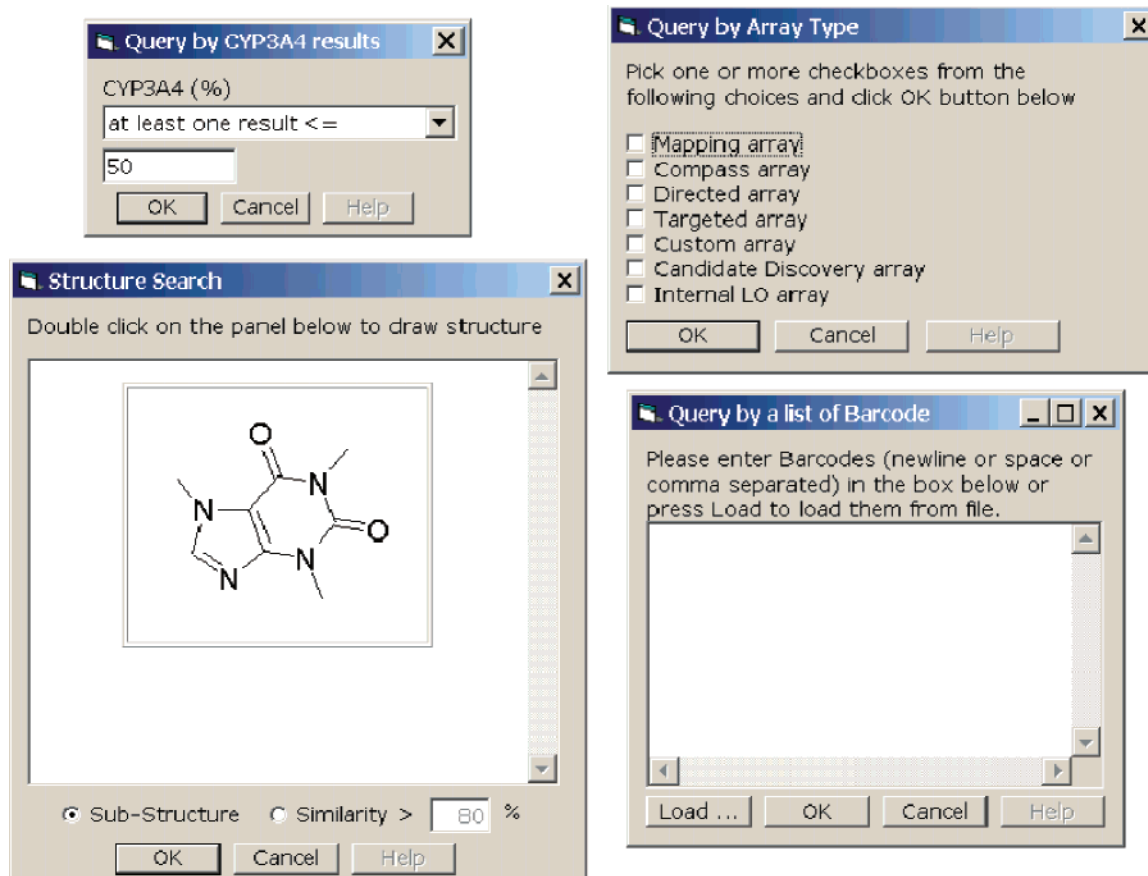


Figure 4. Example of criterion qualification dialogues.

469 batches (425 unique AQ#) #/Pg. Viewer Sort Save Export dEvo Home

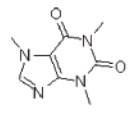
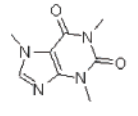
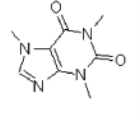
#	Structure	ArQule Number	Batch Id	CLOGP: Daylight estimated log (octanol:water partition coefficient)	Purity UV	Assay1 (uM) ▲	Assay2 (min)	Assay3 (uM)	Assay4 (Percent)	Assay5 (uM)
1		1234567	3038677	6.626	96.585 91.257	.396 .326				Result: Calc_Failure, IC50>100uM
2			3331918		100 100	0.5569	1.20	6.5228	57 52	>100.0000
		1234568	3359946	5.832	100 96.58	>10.0000	1.80			>100.0000
			3364054		100	8.4624	1.60			>100.0000
3		1234569	3038674	5.619	90.184 87.624	.579 .776				271 [WARN: Result: Calc_Failure, IC50>100uM]

Figure 5. Tabular view of data. (Structures and assay names were replaced for confidentiality reasons.)

Queries are executed asynchronously on the server, while the interface refreshes periodically to notify users of the progress. Once query execution completes, users are redirected to data browser pages. By default, results are presented to users in a tabular format (Figure 5). Users can click on most data points to drill down to more in-depth informa-

tion (e.g. IC50/half-life curves, chromatographic traces, control data). In addition, users can also switch to a structure grid view (Figure 6) in which numeric results can be color-coded to facilitate the visualization of many structures and data on a single page. Users can also create SAR tables (Figure 7) to help them understand the effects of various substitution groups.

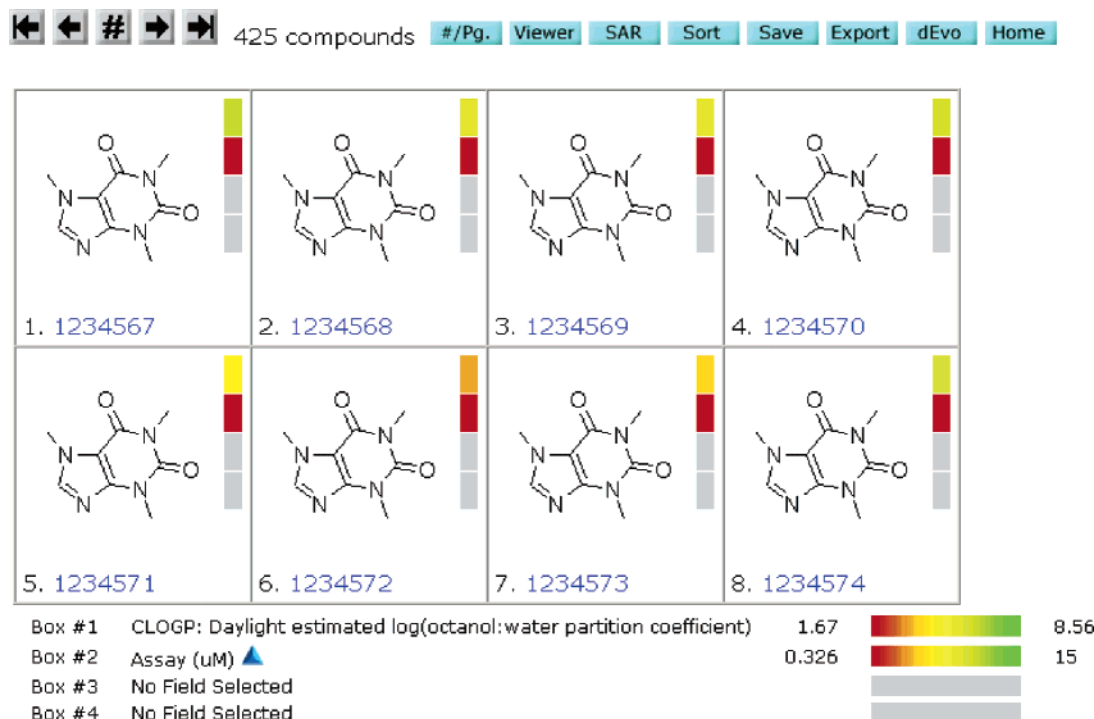


Figure 6. Structure panel view.

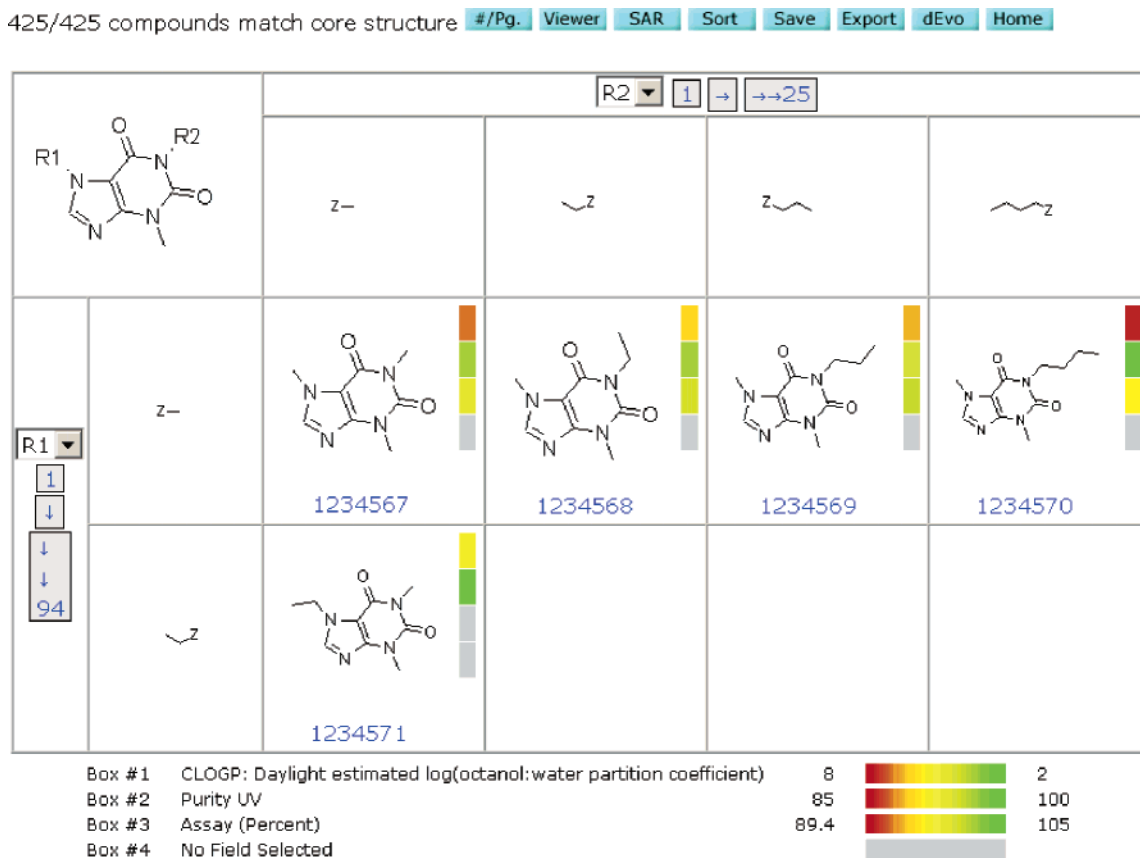


Figure 7. Structure panel view with SAR information. Users specify the core structure, and R groups are computed using Pipeline Pilot "Generate SAR Information" component.

ArQilogist results can be exported into several formats including SD file, tab-delimited, Microsoft Excel with structures, and Spotfire. Since these formats do not support hierarchical data, users are given a number of options for aggregating the hierarchical data into a tabular format (e.g. averages, minima, maxima) prior to export.

DATA SOURCES

The informatics infrastructure at ArQule is built around the ArQule United Information Repository & Exchange system (AQUIRE), a centralized database that stores information used by multiple departments (e.g. compound struc-

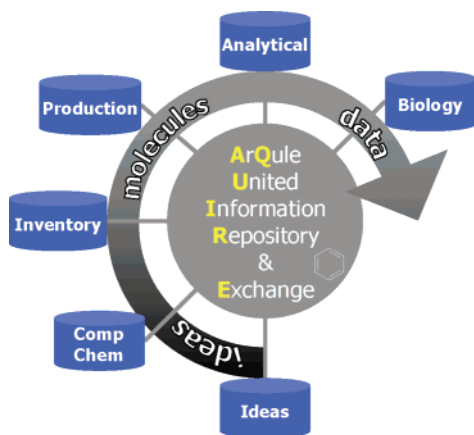


Figure 8. ArQule informatics infrastructure.

tures, batches of compounds, containers such as vials and plates, employee information). In addition, there exist a number of specialized databases that store department- and discipline-specific data (Figure 8). These distributed systems integrate with AQUIRE via exchanges of database keys and/or summary data.

Biological assay results are stored in ActivityBase from IDBS,¹⁰ which runs in a separate database instance from AQUIRE. ArQilogist retrieves biological results directly from ActivityBase. The relevant assay results are displayed as hyperlinks that allow users to further drill down to control data and IC₅₀/half-life curves when available. In addition, ArQilogist retrieves legacy assay results as needed from the legacy in-house predecessor to ActivityBase. Finally, ArQilogist can also be directed to merge data for any assay that exists in both systems into a single field.

Analytical chemistry data pertaining to compound synthesis are stored in our Array Information Management System (AIMS).¹¹ The most commonly accessed summary analytical data (e.g. purity, positive identification by mass spectrometry) from AIMS are copied into AQUIRE when the data are validated through an in-house analytical data processing software application called ASPECT.¹² The frequency of access, paired with the relatively small footprint of this summary data led us to conclude that copying the data would be more efficient than retrieving it directly from its source. However, the users can drill down to the more detailed analysis data stored in AIMS (e.g. peak information, chromatogram images, MassLynx diversity report), in a manner similar to that of the biological data.

Other array-specific information, such as array numbers, chemist names, and reagent composition, is also copied into AQUIRE through automated registration scripts and is also available through ArQilogist. Drilling-down from array numbers in result pages provides further information such as synthesis procedures, cycle time, and virtual libraries considered as well as the structures of compounds that were synthesized but did not pass purity and/or quantity requirements.

Predicted properties for compounds (e.g. polar surface area, rule-of-5, pADME) reside in another dedicated database system. In addition to retrieving existing predictions from this database, the most recent release of ArQilogist can perform on-the-fly property calculations leveraging the SOAP interface of the Pipeline Pilot application.¹⁵ Users do not need to worry about the physical location of any of these types

of data, as ArQilogist handles the virtual consolidation automatically.

SYSTEM DESIGN

ArQilogist design relies heavily on the use of interfaces to compartmentalize the system and maximize extensibility.¹³ For example

- The code that retrieves data fields from different data sources is implemented as a set of separate classes, all of which conform to a single interface. The query execution engine interacts with all data fields through this interface.
- The query builder employs the most relevant dialogue box for each data type (e.g. structure-drawing window for chemical structure fields, text box for list-based criteria). These different types of dialogues also conform to a single interface, whereby each implementation must define how to deserialize the criteria, display them, and save any changes.
- The data browser and exporter also follow a single interface so that new data browsers and exporters can be easily added as needed.

ARCHITECTURE

The architecture of ArQilogist is depicted in Figure 9. Initially, metadata are used to generate the palette of data fields that the query builder presents to users. At this point, the security policy is also applied to exclude from the palette the data fields that the current user is not authorized to access. The query builder component, an ActiveX plug-in running within Internet Explorer, retrieves the palette from the server via HTTP in XML format. When a user submits a query constructed in the builder, the query is transmitted to the server in XML format. The server then executes the query and saves the resulting data in temporary files. Finally, data browsers read the temporary files and display the results to the user.

There are several benefits from serializing queries as XML documents. Firstly, serialized documents can be saved in databases, allowing the queries to be reused, shared, and automatically executed (e.g. for the purposes of e-mailing results to users on a periodic basis). Second, XML is language-neutral; there are many software libraries that can parse it effectively.

ArQilogist's server-side code was implemented as a Perl/CGI script running on a Sun/Solaris server. The query builder ActiveX component was written in Visual Basic.¹⁴ Visual Basic for Applications is used for integration with Excel and JavaScript is used for integration with Spotfire. All database servers are Oracle8i instances running on Sun/Solaris servers. MDL ISIS/Direct⁶ is used to handle structure-based queries. SAR analyses are performed through the SOAP interface to Pipeline Pilot's "Generate SAR Information" component.¹⁵

QUERY EXECUTION

Arguably the most complex part of the system, retrieving relevant data in a reasonable amount of time, is key to the success of ArQilogist. Factors that most contribute to its complexity include the following:

- Database size. AQUIRE contains information about roughly 2 million unique compound structures and 300 000 containers

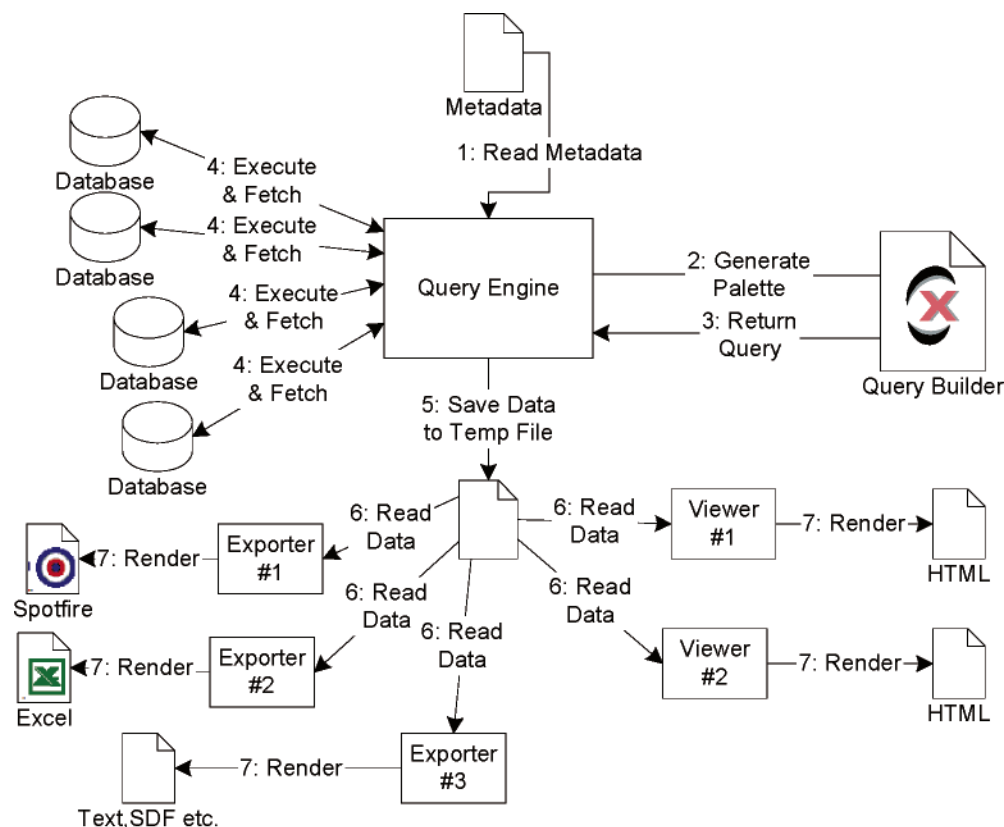


Figure 9. ArQilogist architecture.

- Multiple data sources located on multiple servers
- Access control policy. To protect the intellectual property of collaborators, our databases and software very strictly control which users are authorized to access which subsets of structures and experimental data.
- The compound-batch-data hierarchy. When a query is executed, a decision needs to be made whether data for only the batches of compounds that satisfy the given criteria should be retrieved, or the data for *all* batches of compounds should be retrieved regardless of which ones satisfy the given criteria. The correct choice very much depends on what a user is trying to learn from the query results.
- Open-range data. Results such as “compound X’s half-life in the presence of microsomes is >20 min” are very common in assay results. A user who is interested only in metabolically stable compounds may ask for a list of compounds with half-life >30 min. If she would like her criterion applied loosely, compound X described above should be included in her query result since there exists some possibility that the actual half-life of compound X indeed exceeds 30 min (i.e. is “>20” a valid result to the query “give me everything greater than 30”?). Conversely, other users may not want to include compounds unless all criteria are explicitly satisfied.

The proper behavior in situations such as the last two described above is clearly a matter of user preference. In response to this, ArQilogist sets a reasonable default option to get novice users started, but allows advanced users to override these options for their current query or, if they so choose, on all future queries. In response to user feedback, it was decided that ArQilogist should make users aware of these nuances but ultimately require them to make their own choices as to how to deal with them. This approach was taken

to mitigate the risk of incorrect interpretation of data by the users as a result of hidden assumptions made by our software.

In addition, ArQilogist also supports the use of metafields, which get replaced with one or more actual data fields in the query results. An example of this type of field is the “All Assay Results” item that retrieves all the assay result fields for which compounds in the result sets have data (Figure 3). This method allows users to explore data without knowing ahead of time what data there are.

When a user query in XML format is submitted to the server, metadata are used to convert XML elements into SQL fragments. The SQL fragments are then assembled into a “main” SQL statement that returns batches of compounds that satisfy user criteria along with any single-value data (i.e. one data point per batch or compound). Additional “auxiliary” SQL statements are also generated to retrieve data fields that may return multiple values for a given batch.

Currently, all queries are executed on the actual data sources. No additional data warehousing is used, although this could be easily implemented as our system grows. To ensure acceptable query performance, we use the following tactics:

- SQL hints are used to assist Oracle’s built-in cost-based optimizer to determine the best execution path.
- Based on our knowledge of the data, we generate either “in” or “exists” SQL subqueries depending on whether the current criterion is more restrictive compared to other criteria in the query.
- Query performance is constantly monitored. Queries that take unusually long to execute are flagged for follow-up.

Execution time of a representative set of queries is shown in Table 1.

Table 1. Execution Time of Representative Queries

query description	exec time (s)	#cpds returned
all assay results for compounds in a big project	42	2598
all analytical results, all assay results, salt forms, ClogP and IUPAC name for compounds in a small project ^a	8	200
compound lookup using a list of corporate compound Ids	<1	1
retrieve results of a given list of assays for a given list of corporate compound Ids	1	5
all compounds with IC ₅₀ < 10 nM in a given assay	1	67
all compounds that contain a given substructure and have been tested in a given assay	10	47
LogD (pH=7.4) for all compounds in a small project ^b	113	200

^a ClogP's and IUPAC names were precalculated and stored in our database. ^b LogD's were calculated on the fly using ACDLabs software packaged into a Pipeline Pilot component.

USER EXPERIENCE

The first version of ArQilogist was released in September 2001. The second version followed in December 2002, addressing a number of issues uncovered from usage of the first release:

- Assay and analytical results were averaged within the same batch of compounds from the beginning. The average was performed to force query results into tabular format where each row represents a batch. Since the averaged results could be misleading, the second version preserves the compound-batch-data hierarchy throughout and provides options to allow users to aggregate data before they are exported to external analysis applications that require data to be in some tabular form (e.g. Microsoft Excel or Spotfire).

- No data were retrieved for "run-but-failed" assay results. If ArQilogist v1.0 showed a blank result, users could not distinguish whether (a) the compound had not yet been tested or (b) the compound had been tested but had not produced a valid result. The second version of ArQilogist provides additional information by showing error codes (e.g. "no fit", "insoluble", etc.) for extra clarity. In addition, with the integration with our recently released synthesis-to-assay tracking and scheduling system,¹⁶ users can now further distinguish between compounds that are *scheduled* to be tested and compounds whose testing has yet to be scheduled.

ArQilogist is used as a stand-alone query tool and as a drill-down from external applications, which preload the builder interface with relevant queries. Its architecture has proven to be robust and easily extensible. For example, the integration between ArQilogist and ActivityBase, a relatively complex chemical and biological data management system, took 3 FTE-weeks (the first of which was spent studying ActivityBase data model). Integration with simpler systems can usually be completed in days. The integration not only allows biological data to be easily accessible together with other data not in ActivityBase but also saves a large percentage of our users from having to learn additional software tools. Currently over 850 unique data fields are accessible through ArQilogist.

REUSING THE ARQIOLOGIST FRAMEWORK

We recently had the opportunity to reuse the query builder and query engine of ArQilogist to implement ad hoc query capabilities in a project scheduling and tracking database system¹⁷ for ArQule's Chemical Technologies business unit. The data in this system describe a hierarchical relationship between chemistry projects and the chemical libraries synthesized for each project. Data fields at the chemistry

project level consist of chemical design and development information as well as compound core availability tracking data. Data fields at the chemical library level consist of synthesis schedules, numbers of completed syntheses, and passing rates. Although the lead optimization data and project tracking data are very different in nature, the modular architecture of ArQilogist allowed us to reuse the query builder interface and query execution component by simply substituting a new set of metadata. This not only significantly reduced development time but also provided users of the new system with a query builder interface they were already familiar with.

CONCLUSION

ArQilogist is a powerful ad hoc query tool for decision support in lead optimization. It combines a flexible yet easy-to-use interface with an extensible architecture to provide a platform that can efficiently and seamlessly integrate multiple sources of data and present them to users in a consistent manner. In addition, ArQilogist also provides consistent and logical treatment of the oft-neglected nuances unique to hierarchical compound-centric discovery data, such as the treatment of multipoint data, the stringency of application of criteria to retrieve borderline compounds, and the treatment of open-range results. The feedback from the ArQule user community, combined with the Web-based nature of the tool allows us to add and deploy a steady stream of enhancements to the system with minimal impact on IT staff and minimal requirements for end-user training.

ACKNOWLEDGMENT

The authors would like to thank Dr. Mark Ashwell and Dr. Ken Drake for insightful discussions that inspired many of the features in ArQilogist.

REFERENCES AND NOTES

- (1) Claus, B. L.; Underwood, D. J. Discovery informatics: its evolving role in drug discovery. *Drug Discovery Today* **2002**, 7, 957–965.
- (2) Kreusel, D. From raw data in the laboratory to information availability in the enterprise. *Drug Discovery World* **2001/2002**, winter, 70–74.
- (3) Haas, L. M.; Rice, J. E.; Schwartz, P. M.; Swope, W. C.; Kodali, P.; Kotlar, E. DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems J.* **2001**, 40, 489–510.
- (4) <http://www.id-bs.com>
- (5) <http://www.accelrys.com>
- (6) <http://www.mdli.com>
- (7) <http://www.modgraph.co.uk>
- (8) <http://www.deltasoftinc.com>
- (9) Walters, P. VERDI – An Extensible Cheminformatics System, Daylight User Group Meeting '02, 2/27/2003.
- (10) <http://www.id-bs.com/products/abase>

- (11) Hartsough, D. S. AIMS: Array Information management system, 225th ACS National Meeting (New Orleans) 3/26/2003.
- (12) Deneau, B. R. ASPECT: A LIMS system for characterization of combinatorial libraries, 225th ACS National Meeting (New Orleans) 3/26/2003.
- (13) Gamma, E.; Helm, R.; Johnson, R.; Vlissides, J. *Design Patterns: Elements of Reusable Object-Oriented Software*; Addison-Wesley: Reading, MA, 1994; pp 185–194.
- (14) <http://msdn.microsoft.com/vbasic/default.aspx>
- (15) <http://www.scitegic.com>
- (16) Hartsough, D. S. TraX: An integrated system for drug discovery workflow management, MDL User Group Meeting (Baltimore) 5/3/2004.
- (17) aXis: A project scheduling and tracking system, manuscript in preparation.

CI049880H