# Migrating chemical information to new architectures

**Andrew Lemon[†], The Edge, 27 November, 2006**

## Abstract

In this paper we shall discuss some of the drivers behind the increasing number of pharmaceutical and biotechnology companies moving from traditional proprietary chemical information systems into modern alternatives. We shall explore the business drivers for change and discuss some of the issues which must be considered before migrating chemical information, including an analysis of common issues.

New approaches such as systems biology and translational medicine demand interoperability and integration of data and systems. Oracle data cartridge technology offers the opportunity to break the dependency between a pharmaceutical company's data and the software that manages it, encouraging an increase in the utilisation and value of data.

We shall discuss the benefits of this approach and explore how to achieve migration, focusing on the benefits of JChem Cartridge and the ChemAxon products.

[†] Andrew Lemon is director and principle chemistry consultant at The Edge Software Consultancy Ltd (http://www.edgesoftwareconsultancy.com/). Previously he held positions including principle scientific programmer at MDL Information Systems and chemistry products manager at IDBS.

# Table of contents

# Introduction

Discovery research is a dynamic and changing environment driven by information. The industry is powered by data, directing and informing the scientists throughout their experimentation. The cost of developing a drug for market is well documented and is almost entirely based on the cost of generating vast amounts of information in an attempt to provide proof that the drug is a safe and effective treatment for the target disease. Information is therefore the fundamental currency of drug discovery. Most treatments are based on small molecule drugs synthesised by chemists. The chemistry is therefore a fundamental element of the intellectual property generated.

The corporate registration system traditionally manages and centralises all the information related to compounds synthesised in the organisation and is a business critical system. The compound identifiers are uniquely generated from this database and are used to link the compound to its biological test results. A failure to correctly identify a structure can lead to a missed adverse effect in testing, which can cost a pharmaceutical company billions of dollars and even threaten the company's continued existence.

For this reason the chemical information systems of most pharmaceutical companies are seen as *mission critical* databases and attract a low risk strategy. This has all too often meant that these systems have not been updated and remain housed in legacy software, restricting integration and knowledge exploitation. Ten years ago there was only one or two vendors servicing this market. Today there are around half a dozen companies offering this type of software for pharmaceutical and biotechnology companies, varying in sophistication and capability. Often the choice can appear difficult, they all seem to offer similar software and features but in the world of chemical information the devil is definitely in the detail.

# Business Drivers for Change

Before considering how and what is involved in migrating chemical information it is important to understand some of the business drivers behind the change from the perspective of a research organisation.

## Cost

The drug discovery industry is under an increasing pressure to control costs. The days of unlimited budgets and deep IT pockets are long gone. In today's industry every decision must be well thought out and the case for a return on the investment clearly communicated to the business. The discovery IT industry has been in many cases slow to adapt to this model. The majority of discovery IT departments are under an annual pressure to get more from their budgets. In contrast many larger software companies have a duty to their shareholders to increase profits and reduce operating costs. These opposing forces have radically changed the industry and led to a new breed of vendors determined to take advantage of the opportunity by offering more value for money.

### Licensing Models

A recent trend introduced to the discovery informatics industry is the *one license* subscription model. This can be seen from two perspectives:

- From the vendors perspective it is about encouraging customers to make use of more software and to remove competition
- From the customers perspective it is about paying more for the software they actually use

For small biotechnology companies with no existing systems, this can be appealing as it means they pay once for all the software, as long as they are prepared to take all the software from one vendor. This inevitably leads to compromise because no single vendor offers the best solution for all the main problems in discovery informatics. It is common practise to buy *best of breed* and to integrate the solutions to maximise their benefit. This practise of *bundling* is becoming more common but inevitably leaves the customer feeling they are paying more to access software that they do not need. This has been joined by a trend towards more named user licensing of software which has traditionally been seat or server based.

Subscription models can also be difficult for IT teams as they must re-justify their purchase decision each year, which can involve considerable effort to re-analyse alternatives and reconstruct the business case.

There is nothing fundamentally wrong with in any of these licensing models. The problem is that change is painful.

### Value for Money

Many pharmaceutical and biotechnology companies feel they are no longer receiving value for money from their existing vendors and are increasingly looking for more service orientated vendors.

This feeling of dissatisfaction is often centred on maintenance. There is an expectation that maintenance costs cover two aspects:

- Bug fixes to existing software
- Access to upgrades

Many customers pay between 20-30% of their perpetual licensing costs to receive maintenance. However an increasing number of vendors are failing to deliver real value in upgrades. This leaves the customers feeling under valued and frustrated.

Some software solutions rely heavily on consultancy services. This often covers complicated installation and upgrades, in addition to customisation and integration costs. The nature of these software systems often leaves consultancy as the only option for integration. This can be a reasonable approach, however the cost and quality of consultancy is often disproportionate leaving the customer with the impression that the cost of the solution was misrepresented and that the vendor’s consultancy is over priced. This is exasperated by the use of proprietary technology that means there is not an open market for consultancy skills.

### Buy verses Build

In many companies the above factors have changed the balance between buying and building software. Of course the majority of biotechnology companies are limited in their ability to build systems because of resource restrictions. However an increasing number of such companies are basing their business models on utilisation of modern IT technology, introducing professionalism in information management from the start to maximise the exploitation of their intellectual property.

Generally the decision to buy or build is based on the answers to a few simple questions:

- Can we buy off the shelf software that meets our business requirements? *If* yes *then buy*
- Can we customise an off the shelf application to meet our needs? *If yes then buy and customise*

If the answer to these questions is no

- Does the solution need to integrate to other internal systems? *If no then consider consultancy*
- Can we buy appropriate technology to facilitate internal development? *If yes then build in house*

Increasingly the need to accommodate integration and manage change is leading to build decisions. In our experience this is a natural cycle in the industry and is related to budgets and technology developments. Historically the hidden cost of internal development and concerns over risk management have lead to buy decisions, followed by the emergence of internal projects as commercial solutions fail to keep pace with scientific requirements and improvements in technology.

### Deployment

The true costs of maintaining software needs to be considered, including the major cost of deploying upgrades with bug fixes and enhancements. This issue is especially acute for many large pharmaceutical companies. The cost of deployment can often exceed the cost of purchasing the software by some magnitude. Many IT departments have introduced rules for software selection that increase the barrier to introduction of new software that cannot be deployed centrally. The cost of deploying a new desktop application for example can be in the order of $1,000,000 in some larger organisations. This has encouraged the use of web-based applications and technology such as CITRIX server to reduced costs.

## Support

In today's global market place support needs to be accessible regardless of location and time zone. Speedy resolution of issues is essential in a mission critical system such as the chemical registration database. This means fast access to knowledge and experienced support staff who can diagnose the problem and provide advice on how to resolve it. This goes far beyond just logging the call and assigning a support number. The key factor is *resolution time*. We would encourage any organisation considering purchasing a research information system to road test the support organisation before purchase. There are two aspects to support:

- Documentation
- Human response

An increasing number of vendors provide extensive access to documentation on-line, either on open access or using secure support websites. This documentation must provide help to solve real problems.

## Architecture and Technology

The world of software technology is rapidly changing, adapting and evolving to meet the business challenges of its customers. However, the pace of change is often much slower in the discovery informatics space. This has been especially true of the larger vendors who have the problem of legacy databases to maintain for their customers. This problem has been exasperated by the use of proprietary database technology that does not take advantage of industry proven tools for data life cycle management. Traditionally these solutions have attempted to address all the information management needs of the customer, locking them into proprietary systems and forcing them to train specialist staff. During the mid to late 1990's it was commonplace for each company to employ a specific administrator known as the "Compound Registrar" dedicated to running the corporate information system.

In today's discovery environment management of change is crucial to the reactivity of IT departments. They need to be able to service current needs and react to evolving requirements in a timely manner. The demands of these challenges need a flexible IT platform supported by enabling technology. Most of the requirements for managing chemical information are relatively simple with one notable exception, dealing with the molecular structure.

If one compares the challenges of managing biological and chemical information.

- *Biological information is a complex arrangement of simple data types*
- *Chemical information is a simple arrangement of complex data types*

In order to meet the challenges of flexibility and adaptability required by discovery, it is important to take advantage of standard industry tools and techniques. These often originate from the financial and corporate information industries. They are well practised and proven techniques for managing large scale data systems. Examples include technology from companies such as IBM, Microsoft, Sun and Oracle. Oracle for example has become the standard relational database technology for discovery. Standard industry frameworks such as Enterprise Java Beans, Microsoft .NET, IBM Websphere, and BEA WebLogic provide support for development, deployment and maintenance of enterprise scale systems.

These frameworks have a well established record for maintainability, interoperability and integration. Of course the issue has been that they do not natively accommodate the special chemistry data types. However, several vendors have recognised the value of providing specialist tools which are compatible with these frameworks.

### Tractability

The advantage of these mainstream technologies is that they come with an extensive body of information related to their use, such as forums, online documentation, best practice groups, and source code repositories. This provides an important support for developers and is in stark contrast to their vertical alternatives which rely on specialist knowledge only available from the supplier. The accessibility and cost of skills and training for mainstream technologies is much easier and there is a greater capacity for re-use outside the specialist area of chemical information.

For these reasons many IT departments are approaching their information challenges from a much more technology neutral position.

## Integration

The value of information within a drug discovery program cannot be over stated. The process of discovering a new treatment for a disease is an incredibly complex set of interdependent facts which must be understood weighed and measured, leading to a few decisions that are crucial for success or failure. The future of the whole research program may hinge on one observation in a single experiment out of many thousands, one data point from millions of measurements. The key to the success of the process is therefore in integration and accessibility of information to the scientists. Many of the new paradigms of drug discovery such as translational medicine, systems biology and personalised medicine rely on integration of information and interoperability of systems. This may include connectivity between gene expression, proteomics, chemical synthesis, biological test results, and clinical trials responses through to post approval clinical feedback. Each piece of data is important and interconnected.

There are two aspects to integration which are important:

- Accessibility: Systems need to be accessible from across the organisation, be that intercontinental or across the laboratory
- Standardised: Integration is critically dependant upon established standards for interoperability and intercommunication

**Accessibility**

There is an increasing trend towards outsourced and collaborative research in the pharmaceutical and biotechnology industry. This can range from outsourcing library synthesis or screening with contract research organisations to full scale collaborative research projects with specialist biotechnology companies. This trend increases the need for flexible IT solutions which are compatible or even accessible by collaborating groups.

Company consolidation is another aspect of the need for greater integration and accessibility. The last 10 years has seen an unprecedented number of mergers and acquisitions in the pharmaceutical industry. The process of merging chemical information systems is a special case of migration. Often merger precipitates a fresh look at the informatics infrastructure and a re-evaluation of the requirements of the merged company. This can be particularly acute when the two companies use different technology and systems. The need to consolidate multiple databases and provide accessibility across multiple time zones and locations places increased demands on the corporate information infrastructure.

## Dependency

Past experiences of using proprietary software systems and the expense and difficulty of integration and change has lead many pharmaceutical companies to feel dependant upon their suppliers. This dependency is often seen as a risk which they would like to mitigate. In the past this was impossible because there were so few vendors offering software to meet their needs. This situation has changed and today there are several companies who offer appropriate technology for dealing with chemical information. The move away from specialised information systems to more horizontal technology has reduced the barrier to introduction of new players into the market.

Another reason for this dependency has been the integration strategies used in the past. It is common for multiple systems to be integrated to the corporate registration system. In the absence of a robust integration technology, a *fragile* integration is frequently the only option. The term *fragile* means that the systems are intimately interdependent, reducing the capability of the IT department to independently maintain and enhance each system. This type of integration frequently implies that all systems need to be taken offline, creating an intolerable loss of productivity during downtime. This inevitably restricts improvements and maintenance of the system.

The need to break these dependencies is a key driver for change to more decoupled IT architectures where the dependency on individual components is reduced.

## Oracle Data Cartridge Technology

The introduction of Oracle data cartridge technology has probably been the single most important factor in the evolution of chemical information management since introduction of chemical search technology itself.

In essence the Oracle extensibility architecture allows external business logic to be integrated directly into the Oracle platform allowing custom indexing and searching methodologies to be accessed together with standard search operators for numeric and alphanumeric data. This means chemical searches can be conducted using essentially standard SQL and that Oracle takes care of the management of the data. With an appropriate data cartridge installed Oracle understands and can index chemical information in the same way that it natively understands text and numeric data types.

Whilst alternatives to the Oracle data cartridge have been developed for other relational database technology, such as DB2 extenders, and Informix datablades, Oracle remains the obvious choice in the discovery informatics industry.

## Integrating chemistry into a modern application architecture

Enabling chemistry data types allows Oracle to deal with all aspects of data storage, searching and management. This breaks the dependency between the data layer of an application and its business logic and client tiers. This is important as it means that provided appropriate technology can be sourced for each tier, it can be chosen independently to meet user requirements. System architects are free to choose their own technology in the knowledge that they only need select appropriate chemistry enabled components for each tier rather than vertically. Of course there are still issues of compatibility. The chemistry logic of each component must be consistent to enable interoperability.

For example a three tier application model may use Oracle for the data layer, an apache Tomcat web server on the middle tier and a web browser client tier. Chemical structures can be drawn by the chemist using a Java Applet such as ChemAxon's Marvin; the page will post the structure back to the web server where the JSP or ASP business logic takes care of validating the chemical structure and the rest of the data from the html form. The JSP page then executes standard SQL to insert the structure. In Oracle an Oracle cartridge such as ChemAxon’s JChem Cartridge dynamically builds an index for the structure ready for searching, whilst Oracle indexes the text and numeric data. Each tier of the application is independent and uses appropriate technology to deal with the chemistry data types. If required any part of the application stack can be replaced or upgraded independently.

Standard Oracle tools can now be used to manage the data life cycle. For example backup and recovery strategies make use of standard Oracle DBA skills without special training. The web server can be upgraded without affecting the data layer and additional web servers can be added to ensure failover or for load balancing. The application has been transformed from a proprietary system to a modern maintainable architecture, improving integration, and facilitating future development. This allows the IT teams to reduce costs and improve the service they provide.

## Performance

Introduction of new techniques for automated synthesis and screening of compounds has led to an increase in the size of many pharmaceutical company’s compound collections. There are vast collections of commercially available compounds for screening using high throughput and virtual screening methodologies. At the same time scientists are increasingly used to instant responses to searches from systems like Google and MSN. This has decreased their tolerance of software and increased the demand for responsive searching.

With the introduction of Oracle data cartridge technology there is still an expectation that the database should exhibit acceptable performance and be responsive to applications. This demand places the emphasis on performance of chemistry searches and the time taken to build databases.

Scalability and performance are critical to the success of any system

# Migrating to a cartridge-based architecture

In this section we shall discuss the issues of preparation and migration of chemical information from one system to another.

## Chemical information systems

Before considering the issues involved in migrating chemical information let's consider which systems are affected.

### Chemical Registration

This is often referred to as the corporate registration database. It contains details of the compounds synthesised and acquired by the organisation. The corporate identifier is uniquely assigned to each compound based on its chemical structure. This database frequently contains information about the chemists that synthesised the compound and which notebook the evidence can be found. It is commonly the foundational database in a drug discovery information landscape as it provides the connectivity between experimental observations and the compound itself.

### Compound Inventory (sample management)

This database contains information about the physical location of samples within the organisation and is linked to the compound registration database.

### Reagent Inventory System

This is a database which tracks the reagents used by chemists within the organisation to synthesise compounds for testing. The reagent inventory database is also used to house reagents that are commercially available for purchase.

### Electronic Laboratory Notebooks

These systems are commonly used by chemists to track their daily activities providing both a historic record of the origins of the organisations intellectual property and compliance evidence for the purposes of FDA submission.

### Reaction Databases

In the absence of an electronic laboratory notebook many organisations have historically used internal reaction databases to track synthetic methods.

### Content databases

There are a number of commercially available databases which are used by chemists to plan and execute experiments. They vary from databases of commercially available screening compounds, and literature based reaction databases to databases of patented drugs.

### Virtual Screening databases

It is common for computer aided design to generate large databases of *virtual* compounds which are used in docking and modelling experiments. Most of these compounds will never be synthesised but are crucial to the design process.

### Connected Systems

In addition to systems that directly store and manage chemical information, there are many that make use of the information.

- **Biological Data Management**: These systems are responsible for managing the process of biological testing and are often linked for the purposes of structure activity analysis. In some systems the corporate registration system is an integral part of the biological testing software.
- **Sample Management**: These are commonly integrated to the corporate registration system, linking the compound identifier of the sample to chemical information such as molecular mass and structure to facilitate sample preparation and for searching and tracking purposes.
- **Analytical Chemistry**: These are used to track the analytical evidence that the stated structure is correct such as NMR, MS and IR spectra, together with information about purity and other physical properties of the sample.
- **Laboratory Information Management Systems (LIMS)**: These systems are commonly connected to the chemical information to support reporting and calculation of the correct concentration of samples.
- **Analysis and reporting**: The need for reports is common across discovery, providing time recorded evidence of the actions of scientists together with evidence of the progress of projects and campaigns. The chemical structure is a fundamental part of any report related to the compound and is sourced centrally from the registration system.

Clearly there are many systems which are either directly implemented using chemical information software or may be integrated in a fragile fashion and therefore dependant upon it. The process of migration therefore requires careful planning and execution. We recommend careful risk assessment of each database system before migration. Although this process can be involved it is far from intractable and in many cases can be relatively straightforward.

## Evaluation

Replacing chemical information software is a major change for any organisation, it is therefore important to make a careful evaluation of the technology and its chemistry capabilities. There are four main areas which deserve particular attention:

- **Synergy with internal staff**: Internal IT teams will need to use the software, their experience is crucial to acceptance and success of the system. The evaluation period should allow them to prototype the system and address the key problem areas to their satisfaction.
- **Chemical representation**: A detailed analysis of how the software deals with chemical information is critical. This is discussed in further detail in this report.
- **Performance**: Evaluation of the operational performance of the system in a realistic environment will help to reduce the risk of problems in production.
- **Business relationship**: The evaluation is a good road test of the business relationship with your potential new vendor. Any good vendor should be willing to support your evaluation and provide access to the necessary information to help make the decision.

### Performance

The number of compounds that a corporate registration system typically houses is often more related to the scientific techniques used than the size of the company. For example many small biotechnology companies are increasingly leveraging new computational docking and computer aided design methodologies which require access to databases containing all publicly available compounds which typically contain more than five million structures. They may also make use of virtual screening, requiring databases of computer generated structures which can be huge. Conversely many medium sized pharmaceutical companies have relatively small compound collections focused around a particular therapeutic application.

No matter what size the database the requirements for interaction with the software is common, "*as fast as possible*". Software users have become used to carrying out searches across the entire internet and getting results fast. Performance is therefore crucial to the user's experience. Careful application design can maximise the apparent performance of the system.

A careful assessment of performance is vital during evaluation of any Oracle data cartridge. This should be based on a realistic database that not only represents your current database size but the size anticipated in 5 years time, given the current rate of growth. Such a test database can be readily created using publicly available sources of chemical information such as the ZINC screening database[1]. Most commercial compound suppliers make the structures of their compounds available for download in SDF format. We recommend that you seed the database with representative examples of particular use cases.

Create a representative set of queries that vary in complexity. This should cover not only the simple substructure queries but also more demanding query features:

- Exact match queries
    - Include Isotopes, stereochemistry, tautomers, complicated bridged ring systems
    - Aromatic compounds
    - Check available search options such as ignore stereochemistry, isotopes, charge, tautomers, valence and bond types
- Substructure
    - Large molecular queries
    - Very small substructures
    - Query features
        - Rgroup queries with logic
        - Bond query features
        - List atoms, atom query features such as Q, A
        - Ring bond counts
        - Valency and substitution counts

It is important to understand a little about the indexing technology when evaluating performance. There are generally two types of search algorithm:

- Fingerprint based systems
- Hyper-tree indexes

In simple terms, fingerprint systems generate a molecular fingerprint for each structure in the database which encodes its molecular connectivity. Searches are based on a bitwise comparison of the query and target fingerprints (fingerprint screening). This is followed by a graph comparison (or graph walk) between the query and target structures to evaluate if the two graphs are hits. The graph match is computationally expensive; the screening search is therefore designed to reduce the number of potential hits to an absolute minimum.

In contrast Hyper-tree indexing systems use the concept of building a highly optimised super structure containing all of the structures in the database combined into a single hyper-tree. This search algorithm works by walking the hyper-graph identifying hits by matching the query into the graph structure which contains references to the hit structures. This technique

[1] Irwin and Shoichet, J. Chem. Inf. Model. 2005;45(1):177-82

produces extremely scalable searching but at the expense of insert performance. There are two aspects to the performance of any chemistry system:

- How long does it take to add new compounds? Many companies buy libraries in bulk or need to register large numbers of compounds from virtual design or library synthesis in a single operation. The performance of insert is therefore crucial. This becomes especially important during migration. I once talked to a company that calculated that to export and re-import there molecule collection from their existing proprietary database system would take 32 years.
- How long does it take to re-index the table? It is common place during hardware and software upgrades for database schema to be moved, requiring re-indexing. This is typically achieved over a weekend or over night. The Cartridge must therefore be capable of re-indexing the table within this operational time frame.

In systems beyond registration it will be important to be able to construct more complex queries with chemical information across several tables with more complex joins.

Another often overlooked aspect is the relationship between hardware and software performance. We suggest that you discuss hardware requirements carefully with your vendor and explore if and how adding more processors or a cluster of servers can help performance. This may be related to ensuring high availability using Oracle RAC or GRID computing platforms to take care of intensive operations. In theory fingerprint based search technology is more suited to parallelisation although few vendors seem to be addressing this issue at present.

## Preparing to migrate

Before starting to plan any migration it is important to take a careful inventory of your current systems, identifying interdependencies and highlighting where the chemical information is stored and visualised, together with any specific business logic that requires chemical intelligence.

For example in a registration system:

- Compound structures, average molecular masses and formula stored
- Calculation of molecular mass and formula
- Compound structure are visualised along with the molecular mass and formula
- Salt structures, formula and molecular mass stored
- Salt stripping is conducted
- Chemical business rules are applied
- Structures are checked for duplicates including tautomer matching
- Chemical searching by
  - Structure
    - Substructure
    - Exact Match
    - Similarity
    - Tautomer
    - Stereochemistry
  - Molecular mass
    - Monoisotopic
    - Average molecular mass
  - Formula

Integrated systems

- Biological test database (ID, structure and molecular mass)
- Sample management database (ID, structure and molecular mass)

- Virtual compounds database (structure used to avoid duplication)

This inventory helps to map out the requirements for the new system and can be used to identify a strategy for migration together with an operational plan for moving systems over to the new architecture.

## Plan to migrate only once

As we have discussed migration can imply changes to a large number of systems. For this reason it is important to plan to do this level of change only once. By this I mean that if you plan your new system carefully the process of moving from one supplier's data cartridge to another's can be made as simple as possible. This will break the dependency of the data from the indexing technology and leave you free to make architectural choices that suite your business needs rather than being dictated by your choice of chemical information software.

There are some key techniques that can be used to ensure future migrations are straightforward.

### Stay agnostic on data formats

Chemical information has traditionally used a few data formats which have evolved to accommodate transportation and integration of information. These formats were established by early vendors in this space. For example MDL Information Systems established the CTFile format[2]. Other vendors such as DayLight, and Tripos established linear notations such as Simplified Molecular Input Line Entry Specification (SMILES[3]), and Sybyl Line Notation (SLN[4]). Over the years each vendor has tended to create specialist formats to accommodate additional chemical features. This lack of an established public standard has been and to some extent remains a limiting factor for integration and interoperability between software systems. There have been some efforts to establish public standards such as CML[5] and INCHI[6]. Whilst these have been the catalyst for some interesting discussion, they are yet to be established as standards in that even today most commercial software systems do not take advantage of them.

In the absence of an established public standard two of the proprietary formats have emerged as the most common currencies for chemical information.

- MDL CTFile format
- SMILES

These are both ASCII based systems for representing chemical structures. The connection table file format includes information about the depiction of the structure (its display coordinates), whilst SMILES is a linear notation containing the only the connectivity, and relying on depiction to be imposed externally.

The chemical representation capabilities of each format vary. For example the MDL Molfile format has a rich set of chemical representation features, which have recently been further extended with the release of the "V3000" format including a new enhanced stereochemistry system. This format also contains some more general mechanisms for extending the

[2] Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L. et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited, *J. Chem. Inf. Comput. Sci.*, 1992, 32, 244-255.

[3] D. Weininger, et al, SMILES 1. Introduction and Encoding Rules, *J. Chem. Inf. Comput. Sci*., 28, 1, 31 (1988)

[4] Ash, S., Cline, M. A., R. W. Homer, T. Hurst, G. B. Smith, SYBYL Line Notation (SLN): a versatile language for chemical structure representation, *J. Chem. Inf. Comput. Sci*., 37, 71-79 (1997)

[5] http://www.xml-cml.org/

[6] http://www.iupac.org/inchi/

information that is stored. It is important to remember that CTFile formats contain extra information that is not necessarily essential for indexing and searching of structures.

Clearly conversion of these formats can imply some loss of information. For this reason we recommend that the richest representation is retained, but that the native index format for the chosen cartridge is used for indexing and searching. This allows future enhancements in capability of the search technology to extract extra data for searching from the legacy formats.

### Design for migration

Careful application design can facilitate migration in future. As we have discussed Oracle data cartridge technology allows chemical capabilities to be integrated into the Oracle database itself and accessed directly from SQL. Use of SQL resource files rather than hardwiring the application can facilitate very straightforward migration to different cartridges. Another technique is to encapsulate all the chemistry functions behind a single interface, allowing the system to be ported by simply re-implementing the necessary parts of the interface. This technique can even accommodate variation in the storage format by inline conversion of structures during operations. One must however keep in mind issues of data lost during conversion as described above.

These techniques mean that the underlying search technology can be replaced or upgraded without implying a complete reimplementation of the application.

### Exotic features

Some chemical technology suppliers have concentrated on more exotic and less used chemical representation use cases such as organometallic structures or higher level stereochemistry. Whilst these are potentially very useful advances in representation they can cause real issues for integration and interoperability with other software, not to mention migration. Consider carefully if they are business critical and think of the consequences of living without them.

# Migration

Migrating chemical information raises some special issues which need careful consideration. We shall discuss some of the issues of migrating chemical information focusing on data specifically related to the chemical structure. These issues also affect reactions but in this discussion we shall concentrate on chemical structures and particularly registration systems.

As we have already discussed there are three areas which must be considered when migrating to a new architecture and system.

- **Data layer**: How to move the data into a new cartridge based design?
- **Business Logic**: How can we replicate and improve the business logic of the system?
- **Visualisation**: How can we present a consistent view of the chemical information to our chemists and maintain their understanding of the system?

## Data Migration: Chemical Structures

Migration of the data requires careful attention to detail. Most of the textual and numeric data should be relatively straightforward to migrate, however the chemical structure is more complex. There are a number of specific molecular and chemical representation features which need special treatment.

### Stereochemistry

The representation of stereochemistry is an area that has seen change over the past 5 years. Traditionally systems treated this as a molecular level property indicating chirality as a flag. There has been an increasing trend towards chiral drug development[7]. Registration systems must therefore track each stereochemical entity separately for regulatory purposes. This means that the chemists need to be able to designate the precise stereochemical state of the compound. This can range from a single enantiomer to a racemic mixture. Some systems will facilitate the evolution of this information from registration of a compound with unknown stereochemical state to a fully resolved sample.

### Molecular Mass

You may be surprised to see this as a topic. However for many companies particularly those dealing with fine chemicals the periodic table may not be as certain as we have come to expect in the pharmaceutical world. There are also some subtle issues related to historical treatment of data that can affect migration in the area of molecular mass. There are two types of molecular mass which are important in a chemical information system:

- Average molecular mass
- Monoisotopic mass (sometimes referred to as the exact mass or most abundant peak)

Traditionally chemical information systems have not used a fixed periodic table. Frequently the periodic table was stored in a text file that could be customised by the client. This was done for a number of historical reasons:

- Reduced representation: Many chemists wanted to avoid drawing every explicit part of a molecule to save time and effort. For example they could define a custom atom type Ph and assign a molecular mass of 77.112 and formula C6H5 (taking account of its single valency). This was also used for protecting groups such as Tertiary butyl which could then be drawn using a single atom (*tBut*). This practice has been replaced by the use of abbreviated groups.
- Custom values for exotic atoms: This also allows the molecular mass of any atom to be changed either because the official version has changed or the company believes it has a better source. This practice is generally reserved for specialist chemical companies and is not recommended in discovery research.
- Updated measurements of relative abundance

Use of file based periodic table customisation can lead to poor data integrity if local and server files are not synchronised. This can mean that a compounds molecular weight varies between desktop and server for example.

For these reasons it is important to check the molecular mass values before migration and ensure that they meet you're requirements. This is particularly important for biological testing where small variations in concentration can mean a compound is identified as active or inactive.

### Molecular Formula

The generation of correct full molecular formula is often an area where many systems have problems. During migration we recommend that data is carefully validated to ensure that historical problems are identified and removed. The use of salts and solvates is a typical area where issues can arise. Many systems struggle to correctly construct the full molecular formula taking into account the equivalence of salt and solvate. This can cause issues in

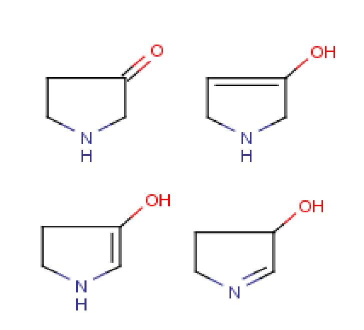
[7] Meeting the challenges of chiral drug development: Advanced stereochemical descriptors, Lemon, Andrew, *PharmaChem*, April 2003

integration where many connected systems make use of the formula to calculate molecular mass.

### Tautomers

Support for tautomers varies between different software systems. Some support full tautomer matching during duplicate checking, supporting registration of tautomers as batches of the same corporate identifier. The most common option is that tautomers are registered as separate identifiers but that a special tautomer matching option allows the chemist to find tautomeric structures in the database. Less sophisticated software supports transformation of any tautomer into an agreed single tautomeric form for registration. This means that all tautomers are treated as chemically equivalent. There are arguments for and against each approach.

In terms of migration the key is to understand any changes to the business rules for registration and to process the compound collection according to these rules. Changes will imply a migration of corporate identifiers. The old corporate identifiers must be maintained for historical records as many systems will be dependant upon them. For example biological test results will probably be referenced by the old corporate identifier unless a sample based screening paradigm has been used.



Generation and searching of tautomers

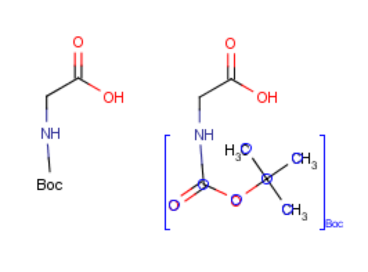
### Aromaticity

Another area which can cause problems is in the perception of aromaticity. The definition of aromaticity[8] is crucial to the process of both duplicate checking and also structural searching. The treatment of aromaticity varies from a simple treatment of Hückel's rule to more extensive algorithms that take into account empirical evidence for aromaticity. A clear understanding of the differences between systems is important. We recommend that you select a vendor who can demonstrate a clear understanding of the issues raised in this paper before migrating to them.

### Explicit Hydrogens

The use of explicit hydrogens can be a problem for some systems. Generally explicit hydrogens are removed during registration and assigned implicitly by the chemical information system, with the exception of hydrogens at asymmetric centres which define stereochemistry. Removal of unnecessary hydrogens keeps structures clean and facilitates visualisation and analysis.

### Abbreviated Groups



Abbreviated groups

Another technique used to reduce the need to draw every atom and bond in the structure is the *abbreviated group*. This is essentially a shorthand form of a structural fragment. The classic example is a protecting group. The abbreviated group is visualised as a single atom label when displaying the structure and yet represents a whole connection table. During migration the integrity of the overall structure must be maintained. This includes its molecular mass, formula and chemical graph. It is also important to maintain the visualisation of the structure and the protecting group as a single atom. This is

[8] http://en.wikipedia.org/wiki/Aromaticity

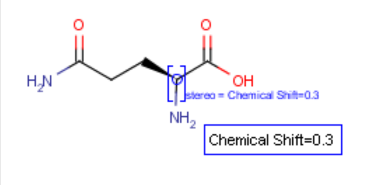
particularly important for the chemists conducting SAR analysis as it hides the complexity of common groups, allowing them to concentrate on the important features of the molecules for analysis.

### No Structure

This is a special type of structure which means that the molecular structure is unknown. This is important as it usually means something different from a NULL. Not all systems accommodate no structures.

### Data Sgroups

Data Sgroups are a particular mechanism of tagging parts of a molecular graph with extra textual data. This data can be used to restrict chemical searches in some software. This is a very flexible metaphor for extending the information stored. In the past many large pharmaceutical companies made use of Data Sgroups to encode structural features such as stereochemical flags identifying particular enantiomers and diastereoisomers during registration.



Example Data Sgroup

However, the use of Data Sgroups can cause migration issues. Very few indexing systems take into them into account and the choice of storage format is critical. If your existing system makes use of specific Data Sgroups, we recommend that they are maintained within the current structural format. In general, where possible it is best to encapsulate all structure related information within the structure itself. This helps to reduce migration problems and the need to translate formats.

As we have discussed one example is stereochemistry. It is possible to use additional tables and columns in Oracle to indicate particular stereochemical states of registered molecules. These flags must be incorporated into the application logic in order to modify the results from a structure search. The problem with this approach is that the chemical information is no longer encapsulated within the structure format.

## Business Logic

### Chemical Business Rules

During registration of a new compound it is important to check that the structure meets some standards of chemical representation. These rules usually take care of issues such as charge state of the molecule, inclusion of counter ions and waters of crystallisation, together with rules about how the chemists should draw the structure. Considerable intellectual effort is often invested in defining, agreeing and implementing these business rules. They are inevitably implemented in a proprietary scripting environment or against a particular vendor's tool kit. Depending on the sophistication of the system, this can represent a considerable effort to migrate to a new vendor's software. For this reason we recommend that these systems are isolated by converting the current implementation into a service. This can be achieved using a variety of technologies. For example converting the business rules engine into a web service, effectively partitions the logic behind an interface, allowing the data to be migrated and the implementation of the business rules to be replaced at a later date.

There are a number of alternative methods of implementing business rules which vary in their sophistication and maintainability. The key is to make sure that the business rules are robust and reusable, avoiding an explosion of similar business rules. It is important to make sure that your new software does not rely on business rules to make up for inadequacies in the business logic of its chemistry. For example there should be no reason to check for valency in a chemical business rule, this should be inherent in the systems logic, not an added feature.

### Duplicate Checking

As we have described earlier checking for structural uniqueness is the single most important operation in a chemical registration database, because it is responsible for ensuring that the corporate identifier is unique. It is important to carefully check the comparison logic and algorithm to ensure that it meets the required business rules for the system. A number of systems provide flexibility in the logic of duplicate checking which can be used to facilitate handling of compounds which only vary in charge or isotopic state, together with various types of isomerism such as tautomerism and stereoisomerism.

### Search logic

All chemistry systems facilitate searching and mining of chemical information using a rich set of chemical searches. The commonest searches include substructure, exact match and similarity searching. All chemical cartridges support these features, but vary in the sophistication of the modifiers that can be used to refine the searches. They will also vary in their ability to handle query features such as list atoms, link nodes and ring bond counts. It is important to think clearly which of these features are going to actually be used by the chemists. All too often whilst the cartridge may be able to conduct a powerful search the chemists cannot phrase the necessary query to achieve their results. For this reason most applications tend to implement fairly simple search interfaces and hide the *advanced* features for expert users. Usually migrating from one cartridge to another is simply a matter of understanding the mapping between search operators and how to construct the correct SQL. This makes migration very straightforward.

## Presentation of Structures

The visual representation of structures is very important to the chemists. The understanding of organic chemistry is uniquely associated with the ability to see and understand the chemical structure. A poorly drawn structure forces the chemist to visually analyse the structure to understand its chemistry where as a structure drawn in a familiar format is instantly recognisable. This is especially important in the pharmaceutical and biotechnology industry where medicinal chemists are required to visually analyse and understand many thousands of structures related to a particular therapeutic target. This is particularly important in structure activity analysis where visual analysis of a few hundred variants of the same structure is required against their biological results. Consistent and recognisable structure presentation is critical.

Another aspect of presentation is the drawing or editing tool. In most cases chemists will stick to a familiar tool for their career. For this reason it is important that the visual interface can make use of structures drawn in a variety of drawing tools such as ChemDraw™, ISISDraw™, Marvin™ and ChemSketch™.

## Methods of migration

There are many methods of migrating from one solution to another, which really depend on the nature of the source and target systems.

In general, where possible it is best to use industry standard methods of migration. These techniques are based on extraction, transformation and loading (ETL).

### Extraction

Depending on the current system, extraction can usually be achieved manually through a command line interface or from a standard database tool. The extraction process must be to a format that can be used to facilitate transformation and loading. Good examples include tab or comma delimited files, or XML. For chemical structures RDF or SDF files can be used or tab and comma delimited files can contain SMILES codes. This process can be very slow and potentially take a considerable time to achieve. For this reason it is best to partition the database into blocks and process these asynchronously from source to target database.

### Transformation

Transformation of the chemical information may take several forms:

- **Structure conversion**: As we have discussed the data formats used by some cartridges can be different. The current format may therefore need to be transformed into that required by the new cartridge for indexing. Some cartridges support multiple formats for indexing which can reduce the need for translation of data. In such case all that is required is a new index type to be added to the table column.
- **Structure Checking**: In a similar way to standard ETL tools migration can be an opportunity to validate the data. This typically includes checking the structures themselves and making sure they adhere to corporate business rules, and checking that dependent information such as mass and formula match the structure.
- **Data cleansing**: In common with standard ETL tools chemical information can also be cleansed. For systems that take advantage of coordinates for display purposes, structures can be subject to cleaning or layout. This is the process of creating a new 2D structure from the connectivity of the molecule.

Data transformation can be achieved using scripting languages such as Perl or Python to manipulate the data. Many Perl scripts have been created that even deal with specialist formats such as RDF and SDF.

### Loading

Depending on the file format used for extraction, the loading process is frequently achieved using standard oracle tools such as SQL*Loader. When the file format used for chemical information is a specialist file type like RDF or SDF, specialists tools are required which can read each record and generate appropriate SQL to insert data into the required data model and particularly into the structure table for indexing with the Oracle data cartridge.

Migration from one cartridge to another can frequently be achieved directly in Oracle and may even not be required. With careful planning migration can be as simple as creating a new index on the table and altering some of the SQL resource files.

```
SELECT cd_id FROM scott.jchemtable WHERE
jc_contains(smiles, 'CN1C=NC2=C1C(=O)N(C)C(=O)N2C') = 1;

SELECT id FROM scott.dcitable WHERE
contains(smi, 'CN1C=NC2=C1C(=O)N(C)C(=O)N2C ') = 1

SELECT cdbregno FROM scott.cttable WHERE
sss(ctab, '<chime-string>')=1
```

### Other tools

Specialist tools such as Scitegic Pipeline Pilot[9], Inforsense[10] or Knime[11] support direct access to database tables and include support for transforming data and loading into the next

[9] Scitegic Pipeline Pilot http://www.scitegic.com/

[10] Inforsense KDE http://www.inforsense.com/

[11] Knime http://www.knime.org/r

database. Pipeline pilot is particularly useful when dealing with chemical information although care must still be taken to ensure data is faithfully transformed.

## Managing risk

As we have discussed the registration system is a mission critical system, it is responsible for generating the corporate identifier and is central fact table for many queries including inventory and biological results. It is the key to the intellectual property of the company. Whilst transitioning from one system to another it is therefore important to keep chemists working and the information accessible. For this reason we recommend that the new system is run in parallel with the old for a suitable time period to allow everyone to be comfortable with the change. This will allow any teething problems to be identified and fixed, ensuring the new system is functioning correctly before decommissioning.

There are often many systems that need to be changed at once, a parallel approach helps to manage the level of change that must occur in one internal release. This is more than a technical issue. In some ways the more pressing issue will be related to retraining staff to use the new systems. This requires careful planning and is not to be underestimated.

## Getting help

Before migration of any chemical information system it is important to learn from the experience of others. The new vendor should be able to supply reference customers who have migrated, and recommendations for how to achieve the move. You may find your previous supplier is willing to help you as well. A mature supplier will often provide assistance; in the recognition that you will be buying other software from them in future and that you may return if they improve their offerings.

There are also independent consultants with specific experience of migrating chemical information. These can often provide impartial advice and even conduct the migration for you, using specialist tools and knowledge. This has the added benefit of reducing the burden on your IT staff, after all migration is not a routine task. Make sure the migration procedures are carefully documented and that you understand all the operations that have been done, as this will be essential information for the future and stop you becoming dependent upon your consultant. A good consultant will always provide detailed documentation.

# Migration to JChem Cartridge

In this part of the paper we shall focus on ChemAxon as a supplier and review the issues related to migration highlighting some of the advantages of their approach.

> *"We've had a really positive experience of dealing with Chemaxon. Their collaborative approach has been refreshing"*
> Alistair Sedwell, Evotec

An increasing number of pharmaceutical, biotechnology and chemical companies are migrating to ChemAxon software. Let's examine how they are addressing the business drivers for change and what makes ChemAxon different.

## Products

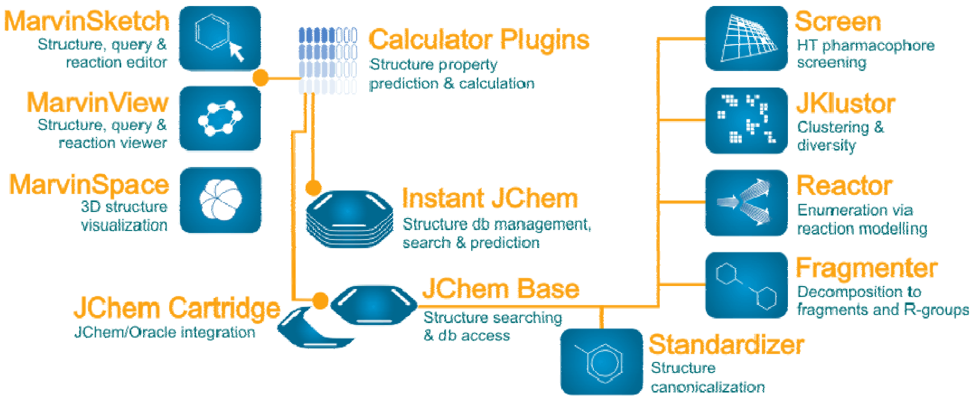

MarvinSketch
Structure, query & reaction editor
MarvinView
Structure, query & reaction viewer
MarvinSpace
3D structure visualization
Calculator Plugins
Structure property prediction & calculation
Instant JChem
Structure db management, search & prediction
JChem Cartridge
JChem/Oracle integration
JChem Base
Structure searching & db access
Screen
HT pharmacophore screening
JKlustor
Clustering & diversity
Reactor
Enumeration via reaction modelling
Fragmenter
Decomposition to fragments and R-groups
Standardizer
Structure canonicalization

Figure 1 The ChemAxon product range

ChemAxon products include application programming interfaces (API), command line tools and end user applications for chemical information management, property prediction, virtual synthesis, screening and drug design.

### Marvin

Marvin is a collection of Java tools for drawing, visualising and characterising chemical structures, substructures and reactions. This includes both end user applications such as MarvinSketch for drawing structures, MarvinView for displaying sets of structures from common file formats, and MarvinSpace for high performance 3D molecule and surface visualisation. Marvin can be used for conversion of chemical information to and from an extensive set of chemical file types. These can be deployed as applets or JavaBeans depending on your application architecture.

### Calculator Plugins

A range of tools are available as plugin calculators for Marvin, JChem Base and JChem Cartridge or directly from the Java API. These provide access to a selection of chemical and molecular properties:

pKa
logP, logD
Polar Surface Area (PSA)
Molecular Surface Area
Charge Distribution
Polarizability
Topology Analysis
Tautomer
Stereoisomer
Geometry
H-bond Acceptor/Donor
Conformer
Major Microspecies
Huckel Analysis
Refractivity
Elemental Analysis
Resonance
Molecular Dynamics
Isoeletric Point
Orbital Electronegativity

```
select jc_evaluate(<molecule>,
'(mass() <= 500) && (logP() <= 5) &&
(donorCount() <= 5) &&
(acceptorCount() <= 10)') from dual;
: Lipinski Rule
```

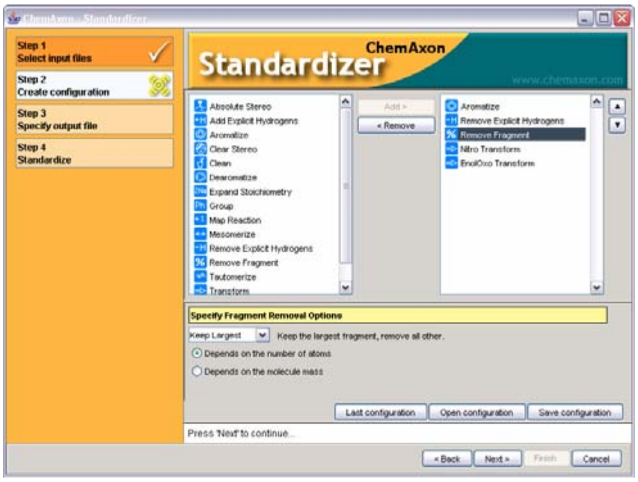
These chemical properties can be accessed from the *Chemical Terms language* which allows chemists to specify structure related chemical properties as a query in JChem Cartridge for filtering the results of a search. They can also be used to calculate properties during migration or data loading. This facilitates a powerful range of chemical evaluations directly accessed from SQL.

### JChem Base

JChem Base is a database independent access layer for chemical information. This can be used to develop chemistry aware websites or applications with structure and reaction searching. JChem Base is compatible with MySQL, Postgres, DB2, Oracle, MS SQL Server and other relational databases. This can be useful when integrating with non-Oracle information sources such as bioinformatics and systems biology data which are traditionally not housed in Oracle. It provides connectivity to the ChemAxon suite of cheminformatics tools. JChem can be used with Java and .NET development environments.

### JChem Cartridge

JChem Cartridge provides access to the full range of ChemAxon chemistry functions directly in Oracle through the SQL language. This facilitates fast indexing and searching of chemical structures and reactions in Oracle, together with extensive processing and chemistry logic. This allows application groups to build sophisticated chemistry applications that deal with complex domain issues such as structure analysis, all from within Oracle.



Step 1
Select input files

Step 2
Create configuration

Step 3
Specify output file

Step 4
Standardize

Specify Fragment Removal Options

Keep Largest — Keep the largest fragment, remove all other.

● Depends on the number of atoms
○ Depends on the molecule mass

Press 'Next' to continue...

### Standardizer

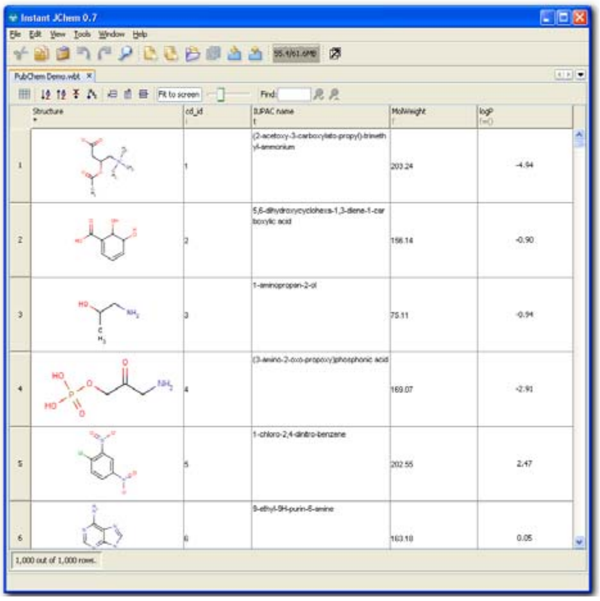
Standardizer is a structure transformation tool, which can be used to implement chemical business rules such as salt stripping, structure cleaning, transformation to a preferred tautomeric form and hydrogen manipulation. These can be accessed directly from JChem Base and JChem Cartridge. This tool is powerful enough to accommodate existing chemical business rules, and has the advantage of being deployable at any tier of your application. During migration existing business rules can be straightforwardly translated into standardizer and accessed from within Oracle using SQL.

### Advanced features

- **Screen**: Supports virtual screening of compounds using various descriptors and metrics. For example, topological or pharmacophore topology fingerprints can be used to analyse large database of compounds to identify targets for biological screening.
- **JKlustor**: Another analysis technique often used with large databases is clustering. This can be used to assess diversity and compare libraries of compounds for overlap and maximum common substructure.
- **Reactor**: Combinatorial chemistry has become a routine part of modern drug discovery used to prepare libraries of compounds based on a common chemical reaction applied to a set of reagents. Reactor is a virtual reaction engine which can be accessed from JChem Cartridge and an intuitive graphical user interface. Generic reactions can be combined with reaction rules to add selectivity and ensure the products are chemically feasible.
- **Fragmenter**: This can be used to analyse compound series identifying common structural fragments and perform R-Group decomposition to identify key fragments responsible for high or low activity. Fragment statistics can be used to discover structure activity relationships and evaluate potentially active compounds.

These modules can be used to replace or enhance existing business logic and extend the analysis capabilities of existing systems during migration.

### *Instant* JChem



| | Structure | cd_id | IUPAC name | MolWeight | logP |
|---|---|---|---|---|---|
| 1 | | 1 | (2-acetoxy-3-carboxylato-propyl)-trimethyl-ammonium | 203.24 | -4.94 |
| 2 | | 2 | 5,6-dihydroxycyclohexa-1,3-diene-1-carboxylic acid | 156.14 | -0.90 |
| 3 | | 3 | 1-aminopropan-2-ol | 75.11 | -0.94 |
| 4 | | 4 | (3-amino-2-oxo-propoxy)phosphonic acid | 169.07 | -2.91 |
| 5 | | 5 | 1-chloro-2,4-dinitro-benzene | 202.55 | 2.47 |
| 6 | | 6 | 9-ethyl-9H-purin-6-amine | 163.18 | 0.05 |

1,000 out of 1,000 rows.

The process of analysing and sharing chemical information requires ready access to chemical information including structures and data in a rich database environment. *Instant* JChem is a new solution from ChemAxon and Informatics Matters[12] supporting chemists in building and searching databases of chemical information. It provides access to many of the advanced features of JChem including access to property calculators, and chemical searching, all from with a standard application interface. This product can be used as a local database which scales to hundreds of thousands of structures or as a client to enterprise relational databases which utilise JChem Cartridge. This application is designed to replace other less scalable local chemistry database solutions bringing the advanced features of JChem within easy reach of chemists.

These cheminformatics tools provide ready access to a comprehensive set of chemical components which can be deployed at any tier of your application to address an extensive set of chemical business logic, indexing and searching of chemical information.

[12] Informatics Matters http://www.informaticsmatters.com/

### Cross platform

ChemAxon's focus on maintaining portability between various operating systems (Windows, Mac, Linux, Solaris, etc) and support for web based integration is an important aspect of their development strategy and is a key differentiator of their products.

ChemAxon provide a variety of technology for integration including Java for system development, HTML/XML for web interfaces, OLE for seamless integration with desktop applications, and communication with SQL database systems using JDBC for data storage and handling.

## Cost

> *"We have been able to expand our usage and reduce maintenance costs by approximately 50% by migrating to Chemaxon software",* Mats Kihlen, iNovacia (Biovitrum)

ChemAxon's focus on delivering value for money, through well designed focused chemistry components is one of the major reasons more and more companies are looking to ChemAxon for chemistry technology. Benefiting from a low cost base they are able to offer greater value for money and flexibility than many traditional suppliers in this area who are forced to fund extensive infrastructure costs and large sales organisation. ChemAxon passes these benefits on to their customers.

> *"Chemaxon software is great value for money; we have simply got more for our budget",* Alistair Sedwell, Evotec

### Licensing models

In common with all ChemAxon products the JChem Cartridge can be licensed in a variety of flexible models to suit your budget and buying cycles. In contrast to some licensing schemes their licensing model does not preclude high availability approaches such as Oracle RAC or deployment across multiple servers. This means discovery informatics groups are able to make design decisions that meet business needs without the constraints of restrictive licensing. Cost is an important factor in selecting chemical information software but is not the only criterion which is important.

> *"Our chief motivation to migrate was cost. We have found a lot of added bonuses from moving",* Mats Kihlen, iNovacia (Biovitrum)

### Deployment

The JChem suite provides a flexible set of components that are compatible with a range of architectures, from traditional client server to more deployment friendly architectures such as web applications. The JChem Cartridge focuses on delivering searching and indexing in Oracle freeing architects to make their own choices. With a range of high and low level API's building business advantage into internal IT systems that address real chemistry issues is both accessible and straightforward.

## Support

> *"Support was excellent, quick, helpful and engaged"* Alistair Sedwell, Evotec

> *"I'm always happy to pay our maintenance",* Matt Pustelnik, Takeda San Diego

This is an area in which ChemAxon excel. The ChemAxon support forums are a lively and well used resource providing both a knowledgebase of tips and ideas, and a source of timely support. Support is provided directly by responsible developers removing the communication barrier that is so often experienced between customer request, support and development. This enables them to respond quickly from a deep understanding of the products and deliver answers which speak right to the issue. A frequent complaint of many IT departments is that whilst they get to report issues quickly their resolution times are often extended.

> *“The cost of purchasing and maintaining ChemAxon’s products is very competitive and ChemAxon’s customer service and technical support are among the best that FMC has found”,* Ann Orth, FMC

Another aspect of this support relationship is the willingness of the development group to issue Beta release versions to address any specific problems. This flexibility means that IT staff experience a team working environment with the support organisation in which the support staff are engaged and interested in resolving their problems.

The support group are highly experienced and have helped many companies to migrate from traditional chemical information systems to JChem Cartridge. This means you can have confidence that they both understand the request and have experience of dealing with it. This type of knowledge is traditionally only available through high cost consultancy services.

> *“We’ve found ChemAxon support second to none…”*, Daniel Butler, Inhibox

## Architecture and technology

In common with all data cartridges the JChem Cartridge does not impose any architectural restrictions. The whole JChem and Marvin suite are designed to deliver chemistry functionality in a focused and capable set of components that can be deployed within the appropriate tiers of your application.

### Integration

> *“Our web applets translate structure drawings into query parameters for the search engine. It is critical that these two tiers can talk in the same language.”*, Yingyao Zhou,

Each tool can be used independently and in combination with other technology. They all support a variety of structure formats which facilitate integration with other chemistry systems. Migration to JChem Cartridge can be as simple and a few SQL changes and creating a new index on the table, helping to reduce the cost of migration. Its support for a range of popular structure formats facilitates integration and interoperability of the software. This reduces the cost of integration by removing the need for translation and associated information loss. JChem supports more chemical representation standards than probably any other comparable toolkit, reducing the loss of information during migration and ensuring valuable data is maintained. Many of the advanced features of the ChemAxon products can be accessed directly from within an Oracle database facilitating business logic such as standardisation of tautomers, salts and other molecular features, together with calculation of chemical properties and interpretation of Chemical Terms for searching.

### Performance

> *“We found performance equivalent to our old system, and the chemistry logic identical”,* Mats Kihlen (iNovacia) Biovitrum

The JChem Cartridge is increasingly recognised as providing enhanced performance over its competitors. Its architecture provides several advantages:

- Use of a highly optimised structure cache makes searches fast
- The Oracle database and the JChem compute server can be run on different machines for load balancing and optimisation
- Its efficient multithreaded design optimises processor usage, maximising the affect of additional processors and scaling proportionally better than other search systems

> *“Faster searches enable larger scale data mining.”* Yingyao Zhou GNF

- It facilitates use of high availability technology such as Oracle RAC
- Resource sharing allows it to scale to larger databases whilst maintaining performance and availability

Performance improvements can deliver a real benefit to chemists using your applications, securing buy in and enhancing the user's experience.

## Chemical Representation

> *"The chemistry logic was very familiar for our chemists which was important for acceptance",* Mats Kihlen (iNovacia)

As we have discussed chemistry logic is crucial to a successful migration. The JChem Cartridge provides one of the most comprehensive chemistry systems available on the market. The JChem search engine is compatible with many of the leading chemistry search engines, reducing the requirement for application changes during migration. An increasing number of chemistry organisations consider the ChemAxon search logic more accurate and chemically intuitive than alternative systems. JChem and Marvin handle all of the chemical representation standards including Rgroup queries, Data Sgroups, link nodes, abbreviated groups and enhanced stereochemistry options. This means the JChem Cartridge can natively index and search more of the chemistry whilst avoiding information lost during migration. JChem maintains the original structure format verbatim eliminating the possibility of information loss during conversion. This significantly reduces the risk of migration.

The JChem Cartridge also contains advanced features such as:

- Flexible matching operators
    - Ignore: isotope, charge, radical, stereo, valence, and bond types
    - Tautomer searching
- Extensive reaction support
- Reaction based library enumeration
- Standardisation of structures
- Integrated property calculators
- Chemical Terms for search modification

ChemAxon are actively developing new features and extending their chemical representation and processing capabilities. They have an enthusiastic and active user group where many of its new features and enhancements originate. Their development plans include Markush structure searching, retrosynthetic analysis and other tools to help improve the access and use of chemical information. All of these features are accessible from across the range of ChemAxon products including desktop applications like MarvinSketch, and Instant JChem to JChem Cartridge and the Java and .NET APIs. This means that chemical intelligence is available at all tiers of your application architecture in a flexible set of components and technology.

As testament to this architectural flexibility ChemAxon products are used in a variety of commercially available products including SciTegic Pipeline Pilot, InforSense KDE, Spotfire DecisionSite, Genedata Sarileo, and Electronic Laboratory Notebooks from Symyx, KLEE Group, EKM, KineMatik, Contur, Rescentris and DeltaSoft.

# Conclusion

The challenges of modern drug discovery are placing an increasing demand on pharmaceutical and biotechnology IT departments. Modern scientific approaches require a high degree of integration and interoperability between systems. Meanwhile scientists and managers tolerance for poorly maintained and under performing legacy IT systems is reducing. Other industries such as finance and manufacturing regularly deal with data of greater scale and complexity using IT technology to address the needs of their customers.

The introduction of Oracle data cartridge technology and focused components that deal with chemistry data types is giving systems architects the freedom to take advantage of advances in modern IT technology to address business needs. The use of horizontal technology benefits from ready access to skilled staff and reduced consultancy costs. The skills learned are more transferable and reduce the dependency on specialist high cost vendors.

Dealing with specialist chemistry data types no longer implies the use of proprietary database technology. The extensive capabilities of the ChemAxon tools facilitate migration to modern application architectures by delivering chemistry capabilities focused at each tier of the application. Their collaborative approach and service ethic has raised their profile as one of the leading vendors in chemical information software. This is backed by knowledgeable and engaged support services and an innovative and responsive development organisation.

Clearly the migration of any chemical information system is an involved and careful process but through the introduction of Oracle data cartridge technology and modern software components like ChemAxon, this is a far more tractable task than ever before. The benefits for moving can far out weigh the cost involved, freeing data from the restrictions of proprietary systems and breaking the dependency between chemistry data and the software that stores it.

*If you would like to understand more about Chemaxon products please visit http://www.chemaxon.com/*

*To discuss the migration of chemical information contact the author, andrew@edgesoftwareconsultancy.com*