

**Migrating Legacy Structure Data into AUSPYX<sup>®</sup>:  
A Case Study**

By *Mike Mandl and Web Homer*  
TRIPOS, Inc.

## Table of Contents

<b>EXECUTIVE SUMMARY .....</b>	<b>3</b>
<b>INTRODUCTION .....</b>	<b>3</b>
<b>REAL-WORLD EXAMPLE: DATA MIGRATION FOR A PHARMACEUTICAL COMPANY.....</b>	<b>4</b>
REQUIREMENTS.....	5
TECHNICAL ASPECTS OF THE DATA MIGRATION .....	6
<i>AUSPYX MTS</i> .....	6
<i>Normalization</i> .....	6
<i>Structure validation</i> .....	7
<i>Duplicate checking</i> .....	7
CHALLENGES IN MIGRATING THE STRUCTURAL DATA .....	7
<i>Chirality and regiochemistry</i> .....	7
<i>Converting and retaining Substance Group (SGROUP) data</i> .....	8
Super Atoms.....	8
Multiple SGROUP .....	9
Polymers.....	9
VALIDATING THE MIGRATED DATA .....	9
<b>CONCLUSION .....</b>	<b>10</b>

## EXECUTIVE SUMMARY

This case study presents considerations unique to the migration of chemical structure data from one storage system and format to another. An illustrative example of such a project within a global pharmaceutical company is presented to illustrate the challenges and typical solutions. The project involved the transfer of more than 4 million structures as part of developing a new enterprise chemical information management system to support research.

This paper shows how Tripos consultants work to analyze a project, identify problematic issues unique to the project, create solutions where needed, and bring the data migration endeavor to a successful conclusion.

## INTRODUCTION

The migration of chemical data from a legacy storage and searching system into a new storage format and medium is often a key determinant in the success of software development projects. Software tools depend on complete and valid data for their success in supporting research.

There are concerns generic to any data migration project, including preserving data integrity, minimizing data loss, and not creating data redundancy. When the data consist of chemical structures and associated information, the intricacies of capturing all of the information in the legacy system and transferring it accurately to the target system may not be immediately apparent. This is a consequence of the complexity of the information stored, which includes not only the atoms of a molecule, but also their connectivity, spatial relationships, ionization state, and more. In addition, migration can involve moving data from a homegrown system with its own idiosyncrasies, or from a proprietary structural information management system with unique fields and formats.

With these types of projects, even adherence to industry best practices in determining the scope of the migration, analyzing the design of the data source and data destination, and mapping data relationships may not be enough. Concerns specific to chemical structural data must be defined in advance and require appropriate domain expertise.

Among the key questions that should be considered in the process of chemical data migration are:

- What are the semantic differences between the way chemical information is stored in the legacy system and the new system? What tools can be used or constructed to effect accurate translation?

- Once structures have migrated from the old system to the new one, is it necessary to translate the molecules back to their original format for viewing and validation by scientists?
- Is the data migration part of a plan to replace an existing system that uses the data? If so, then translation/migration to the new storage system is only a part of the solution, and requirements of the auxiliary system must be considered.
- What is the plan for validation of the migrated data? What metrics will be used to determine if migration has been successful?
- Are there performance criteria for migration and searching, and what are the criteria?
- Are there associated data types in the legacy system that has no counterpart in the new system? Did the legacy system lead users to create unusual “workarounds” for storing data that must be retained? Are there mixtures and formulations to be considered?
- Are there new options for storing certain types of information with structures in the new storage system, and are these to be exploited?
- What are the business rules for duplicate checking?
- What is the preferred chemical sketcher for structure entry, what standards have been set by the business for use of the sketcher, and how well are those standards enforced?

### **REAL-WORLD EXAMPLE: DATA MIGRATION FOR A PHARMACEUTICAL COMPANY**

Recently Schering AG, a global pharmaceutical company based in Berlin, Germany, engaged Tripos for the design, development, and deployment of a chemical information management (CIM) system that integrates registration, inventory and ordering processes with an electronic laboratory journal. In Schering’s legacy information system, the central repository for chemical data was Oracle<sup>®</sup> coupled with a data cartridge from another vendor.

As part of making the new CIM system operational, structure data had to be moved from the legacy storage system into AUSPYX<sup>1</sup>, Tripos’ Chemical Data Cartridge for Oracle, which would serve as the central structural repository for the new system. This repository would hold all of Schering’s structures, structure related data, batch data, and container

---

<sup>1</sup> W. Homer and M. Mandl. *Unified Storage and Searching Of Structures and Relational Data in Oracle<sup>®</sup>: An Overview of AUSPYX 1.7*. Tripos Discovery Informatics Solutions, 2004.

data. Schering had 3.5 million entries in their main database, and its subsidiary had another 4 million structures in their database. There was some overlap in content between the two databases.

AUSPYX represents chemical structures as SYBYL<sup>®</sup> Line Notation<sup>2</sup> (SLN) strings and stores them in VARCHAR2(4000) database columns after removing all coordinate and non-structure data from the SLNs. The legacy system employed a different chemical semantic in which structures were represented as MDL<sup>®</sup> molfiles.

Schering AG initially did a quick test by running a subset of their data through the AUSPYX command line client application *udcimport* and then evaluated the migration with the results of that import. When the CIM project started, AUSPYX could translate and successfully search all but about 1000 molecules. However, Schering's requirements went beyond simple structure translation: molecules had to retain all of the associated data, but reflect changes from structural normalization occurring during import.

In order to support development of customized data conversion scripts, Schering provided a test data set of 564,000 structures from their main database of 3.5 million entries. Schering has been in the chemistry business for a hundred and fifty years and has collected a diverse compound database including polymers and many specialized reagents. The validation set contained some of the oldest structure entries in the Schering database.

For testing and eventual production purposes, MDL molfiles of the test data set were loaded into an Oracle schema to serve as a staging area for data migration into the CIM. The PL/SQL UDC\_IMPORT package used provides a rich set of features to duplicate check structure entries and perform structure validation in a native procedural interface. Tripos developed custom scripts to translate Schering's legacy MDL molfiles (stored as CLOBs) into SLNs for insertion into AUSPYX.

## Requirements

Since the old CIM system's primary molecule data entry tool was ISIS/Draw, as part of the migration to SLN, each SLN was translated back to MDL format and stored, allowing Schering's chemists to see the compound after translation and standardization. Schering required that super atoms and brackets be retained as originally drawn in the derived

---

<sup>2</sup> S. Ash, M.A. Cline, R.W. Homer, T. Hurst, and G.B. Smith. *SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation* J. Chem. Info. Comp. Sci. **37**:71-79 (1997). A full description of the language is contained in the SLN manual shipped with AUSPYX.

molfiles. The custom scripts developed by Tripos called SLN\_OPS functions to translate molfiles to SLNs and vice versa. In the process, molecules were duplicate checked, standardized and validated. A new molfile was generated from the normalized SLN and written into the CIM schema.

The new CIM system is intended to function with ISIS/Draw as the standard sketcher that all users employ for entering structures, including queries. Structure queries require different translation settings: for example, a substructure search query should not have its valences filled. In addition there are a number of molecule features that are only available in queries such as “any” atoms, “any” bonds, list atoms, and R groups. While not strictly data migration, part of Schering’s acceptance criteria was to run substructure searches against both the legacy data set and the migrated dataset. ISIS/Draw was used to create the queries, requiring that the legacy query semantics be translated into SLN.

## Technical Aspects of the Data Migration

### ***AUSPYX MTS***

The fundamental tool employed for the data migration was Tripos’ MTS, which is a second-generation tool for translating MDL’s proprietary structure file formats. Developed to replace the SDfile parsing software in UNITY<sup>®</sup>, MTS was incorporated into SYBYL and UNITY in 1996. It is a part of the *dbtranslate* application shipped with UNITY and has been used to translate billions of molecules since its creation. Since 2002 it has been used by AUSPYX, and has been enhanced to provide better support for data migration in registration systems by enabling translation of features unique to MDL file formats.

### ***Normalization***

When compounds are loaded into an AUSPYX dataset, they are run through normalization code that identifies equivalent representations of molecules and modifies the structure to a consistent form. This facilitates data integrity since this process converts equivalent chemistries to a consistent form.

AUSPYX stores aromatic molecules as aromatic, while the legacy vendor stored them with alternating single and double bonds (kekule form). Aromatic normalization is done during the translation process, the effect of which is to normalize different resonance forms. This aids in the identification of duplicate structures.

### ***Structure validation***

The configurable AUSPYX validation service identifies molecules that are ambiguous, invalid or should be reviewed by a chemist prior to registration. To this end, isotopes and atom valences are checked. Structures that violate these constraints are identified as invalid, for example, pentavalent and hexavalent carbons are rejected. Structures that are exceptions to the constraints but may actually be valid, such as a hexavalent carbon in a boron cage complex, can be registered with validation turned off for just that molecule. In addition to valence checking, AUSPYX supports database-specific rules, which when matched indicate that a molecule is not valid. In this case, Schering added a rule to reject any molecule having a net charge.

### ***Duplicate checking***

Although AUSPYX provides duplicate checking methods, Schering opted for an enhanced duplicate checker in the conversion scripts to implement their duplicate checking policy. This policy required treatment of tautomers as duplicate structures. Extensive investigation was done after the duplicate check to verify that all duplicates were valid.

## **Challenges in Migrating the Structural Data**

### ***Chirality and regiochemistry***

The legacy data was in the V2000 MDL® Molecule file format. According to the documentation for this file format, chiral molecules are specified by setting a global Chiral flag for the molecule. If this flag is set, then wedge bonds on an atom indicate the chirality of the atom. The presence of the chiral flag indicates that the compound is chiral and is pure. The lack of the flag indicates that the compound is a racemic mixture.

There are several problems with this approach to representing chirality. One problem is with the representation of meso compounds. Meso compounds are not chiral, nor are all of their different forms super imposable, as would be the case if they were a mixture. Another problem is that for most molecules there are several equivalent ways to sketch the chiral center by wedging different bonds. Adherence to drawing standards can moderate this problem but not eliminate it entirely. These issues make duplicate checking difficult for the legacy system.

In contrast, SLN syntax provides an atom attribute to indicate chirality. There is no need for a molecule global flag indicating that the structure has chiral centers.<sup>3</sup>

MTS uses the wedge bonds to determine the chirality of a center. Likewise, MTS uses the 2D coordinates of the atoms adjacent to a double bond to determine the regiochemistry of the double bond. Older structures indicated unknown double bond regiochemistry by placing adjacent atoms collinear with the double bond; MTS interprets this correctly and it will use either marking for double bonds as well.

Tripos in conjunction with Schering developed a system for marking relative stereochemistry of chiral centers based on a drawing standard using the V2000 molfile atom value flag to mark centers as being absolute, relative, or racemic. This allowed Schering to utilize older versions of ISIS/Draw to mark extended stereochemistry.

### **Converting and retaining Substance Group (SGROUP) data**

The semantic of the legacy data cartridge allowed creation of SGROUP data that served a variety of uses, such as attaching arbitrary data to atoms, bonds, or the entire molecule; or indicating that a compound is a polymer, formulation, mixture, etc. These SGROUPs have a wide variety of different uses, and their free form nature presents challenges to translation software. The types of SGROUPs are discussed below.

### *Super Atoms*

Super Atoms were used in the legacy system to label functional groups. They provided no searchable or chemically important additions to the molecule; rather they made it easier to identify common functional groups. Chemists could abuse this functionality by making up their own functional group identifiers, which in turn could make it difficult to interpret molecules created with Super Atom SGROUPs. Schering required that these be retained in the generated molfiles. MTS can retain information in the SLN to regenerate the super atoms, but also has an option to ignore Super Atom SGROUPs. When this option is used, molfiles generated from the SLN will not contain Super Atom SGROUPs.

---

<sup>3</sup> M. Mandl, R.Portnoi, J.Swanson, O.Hofmann and R.Jautelat. *Extensions to SYBYL® Line Notation for Representing Relative Stereochemistry and Isomeric Mixtures*. Tripos Discovery Informatics Solutions, 2003.

### *Multiple SGROUP*

The Multiple SGROUP is used as shorthand to indicate that a group of atoms repeat a specific number of times. This SGROUP is associated with brackets in Schering's ISIS/Draw's rendition of the molecule and does not indicate a polymer. In the molfile all of the atoms are explicitly included although they have the same 2D coordinates, and appear to overlay each other. This functionality is often used when drawing salts where the chemist needs to specify the number of instances of a molecule needed to make the stoichiometry correct.

The MTS option to ignore SGROUPs will translate the molecule as it is without the brackets. However, Schering required AUSPYX to regenerate the molfile with the Multiple SGROUP data intact.

### *Polymers*

This SGROUP represents a large number of different types of SGROUPs, which are all ways of representing different types of polymers and copolymers. It includes information about how the monomers connect to each other. Schering had ~1200 of these in their test data set. They represented legacy data that, while important, was not critical to current projects. However, Schering required that these molecules be translated as well.

MTS translates polymers into a specialized SLN extended Markush syntax. The resulting SLNs are searchable via the AUSPYX polymer operator. MTS has options to ignore SGROUPs, in which case these molecules will not be translated.

### **Validating the Migrated Data**

Once the molecules had been translated, Schering needed to verify that the molecules were correct and equivalent to the original. In order to recognize this milestone, Tripos and Schering agreed to use the test data set consisting of the first 564,000 compounds in the main database. These compounds represented the oldest compounds in their database and often the most troublesome. Schering loaded their subset into a separate Oracle schema and executed a stored procedure to translate the molfiles into SLNs and registered them along with a molfile representation of the registered SLNs. The procedure logged when a structure was a duplicate, and whether it was rejected as invalid. AUSPYX found that ~22,000 structures were duplicates and rejected 900 compounds as invalid. In-depth inspection found that most of the invalid compounds were truly invalid except for a handful of molecules containing boron cage complexes, which included hypervalent carbons and nitrogens. In some cases, Schering corrected these invalid structures and re-submitted the molecules into the system. This process detected a

large number of drawing errors and trivial duplicates such as when the same molecule was drawn with different but equivalent Kekule form.

Schering then ran a set of “exact match” queries using the original molfiles as input to verify that each molecule registered could find itself. Specific molecules were cherry-picked and used for testing. Several structures had large ring systems in which isomers differed by *cis* and *trans* double bonds. These were used to verify that regiochemistry was correctly handled. Similarly, a number of chiral and meso compounds were used to check for correctness and consistency. As a result of this validation, all but one peptide and a collection of polymer molecules matched as expected.

In similar fashion, Schering developed a set of substructure queries to validate AUSPYX’s searching capabilities as well as the migration results. Validation of this substructure search was done by Schering chemists with input from Tripos to explain the differences. One primary area where the legacy storage system and AUSPYX differed was in how they treat heterocyclic five-membered rings. For example, one of Schering’s test substructure queries was an indole ring. Many of the structures returned by this query in the legacy system should not have been returned to the user because they did not have aromatic bonds in the appropriate places. However, the same search done in AUSPYX revealed fewer hits than the legacy system, and the hits that were returned by AUSPYX were all verified to be correct.

## CONCLUSION

Through proficiency in analysis, problem-solving, and planning, Tripos consultants were able to design and develop scripts that successfully migrated all of Schering’s molecular data from a legacy system into a new AUSPYX-based chemical information management system. The challenges of this project included examples of even the most complex molecule data, such as polymers and dendromers. In the process, Schering was able to leverage the project to cleanse their chemical data: many duplicate structures were identified in the data, and a number of other structures were found to be invalid. While the scripts were developed specifically for Schering’s new CIM application, they relied upon AUSPYX’s translation, standardization, validation, and searching capabilities.

In this study, Schering benefited from Tripos’ domain experience acquired in working with many pharmaceutical, biotechnology, and life science companies through 25 years. Although each new or refurbished informatics deployment presents unique challenges, Tripos applies its breadth of understanding to every project to increase the probability of a successful data migration.

**For more information**

Please contact your Tripos Discovery Informatics consultant, call 1-800-323-2960 in the US, send email to [contact\\_us@tripos.com](mailto:contact_us@tripos.com), or visit our web site [www.tripos.com](http://www.tripos.com).

USA  
Tripos, Inc.  
1699 South Hanley Road  
St. Louis, MO 63144  
1-314-647-1099

FRANCE  
+33 1 69 59 29 49

GERMANY  
+49 89 45 10 300

AUSTRALIA  
+ 61 (7) 5439 9775

UNITED KINGDOM  
+44 1908 650000

CANADA  
+1 450 4334500

Tripos combines leading-edge technology and innovative science to deliver consistently superior chemistry-research products and services for the biotechnology, pharmaceutical and other life science industries. Within Tripos' Discovery Informatics business, the company provides software products and consulting services to develop, manage, analyze and share critical drug discovery information. Within Tripos' Discovery Research business, Tripos' medicinal chemists and research scientists partner directly with clients in their research initiatives, leveraging state-of-the-art information technologies and research facilities.

While the information presented in this white paper is believed to be current and accurate, it is presented as-is without guarantees or warranties of any kind.

Tripos assumes no obligation to update any information in this document. Tripos makes AUSPYX available subject to the terms of the Tripos Software License Agreement, and nothing in this white paper should be construed as a warranty concerning the performance of the product.

AUSPYX, SYBYL, UNITY, and the Tripos logo are trademarks or registered trademarks of Tripos, Inc. All other trademarks are the property of their respective owners.

©2006 Tripos, Inc. All rights reserved.

Printed in USA