WHITE PAPER

Chemistry Data Migration Challenges and Best Practices

A cross-functional team of experts as a key element for preserving chemical assets of life science and chemistry R&D

By Thomas Doerner, Gerd Blanke and Mary Beth Walsh

www.kalleid.com
www.linkedin.com/company/structurependium-technologies-gmbh
www.tdoerner.eu
www.informaticsalliance.org

Kalleid
StructurePendium Technologies GmbH
Thomas Doerner
Informatics Alliance

CONTENTS:

WHITE PAPER

Chemistry Data Migration Challenges and Best Practices

A cross-functional team of experts as a key element for preserving chemical assets of life science and chemistry R&D

By Thomas Doerner, Gerd Blanke and Mary Beth Walsh

Chemistry plays an important role in our lives and touches almost every aspect of our existence in one way or another. To enrich our quality of life, research scientists in the chemical and life science industries create, analyze and optimize vast numbers of new chemical compounds in order to create novel solutions to problems in health and materials.

The primary focus of research in the biotech and pharmaceutical industries, for example, is to identify safe chemical compounds that have efficacy against a disease and can thus form the basis of a pharmaceutical drug. The chemical structure of a compound is responsible for virtually all the factors that will determine the feasibility of it becoming part of an FDA approved drug – absorption, potency, metabolism, selectivity, distribution, excretion, toxicity, patentability, and viability of synthesis. Chemical structure and reaction data (including data for entities such as peptides and antibody-drug conjugates) form the basis for valuable intellectual property at biotech and pharmaceutical organizations.

Biopharmaceutical companies model, screen, synthesize, test, and analyze thousands of compounds every year seeking those with therapeutic value. Analysts work to identify likely drug candidates using criteria based on chemical structure, diseases/targets, and other characteristics (e.g., structure–activity–relationship (SAR) analyses). This process generates large amounts of chemistry structure and reaction data stored in data repositories. Pharmaceutical companies often add to their repositories by collaborating with external partners. Within this ever-growing body of chemistry data resides considerable predictive potential that can be used to aid identification, optimization and/or synthesis of compounds with efficacious properties.

Within this ever-growing body of chemistry data resides considerable predictive potential that can be used to aid identification, optimization and/or synthesis of compounds with efficacious properties.

Given today's highly competitive business environment, biopharmaceutical and chemical organizations are continuously modernizing their IT landscape, streamlining operations and digitizing processes to improve their discovery workflows, reduce costs, and to keep their competitive edge. As a result, IT projects that require migration (or merging) of this important chemistry data are growing more commonplace. Some examples of projects which could trigger the need to migrate chemistry data include:

- **Global harmonization initiatives** (e.g., across sites, after merger and acquisition)
- **FAIR data initiatives** - Findability, Accessibility, Interoperability, Reusability
- Building or populating **centralized data depositories** (e.g., data lakes, data warehouses)
- **SaaS / Move to cloud** - Implementation of a cloud and SaaS based IT infrastructure
- **System upgrades**
- **Application portfolio rationalization** initiatives

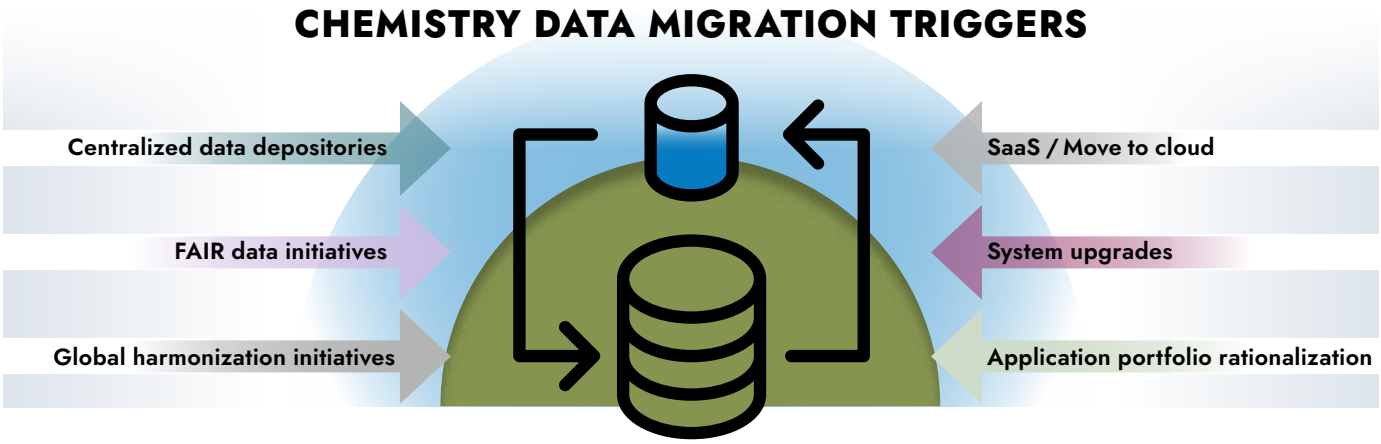


Figure 1: IT projects that require migration of chemistry data.

Affected systems include electronic laboratory notebooks (ELN), chemical registration systems, chemical inventory management systems, LIMS, chemical cartridges, and others.

Data migration involves selecting, preparing, extracting and transforming data from one computer system in order to permanently load it into another. Validation of the migrated data and decommissioning of the legacy data storage system are also considered to be part of the data migration process.

Data migration projects can be challenging and are known for high failure rates. According to Gartner, 83% of data migrations either fail completely or do not meet expected timelines or budget.[1] A recent study by Experian revealed that 44% of U.S. companies surveyed reported data quality issues caused data migration project delays.[2] The bottom line is that, without proper planning and execution, the migration process can end up being the cause of major project delays and budget overruns, not to mention the frustration from business leaders and end users as the planned go-live date for the target system slips by months or even years.

Most organizations are simply unprepared for the complexities involved in legacy data migration, but chemistry data migration is in a class all by itself. Successful migration of chemistry structure and reaction data requires a highly specialized and experienced team to ensure your valuable R&D assets and IP are protected and preserved. In this white paper, we will discuss some of the unique challenges that are involved in chemistry data migration and present an effective methodology that will help set your business up for success.

Chemistry Data Complexities

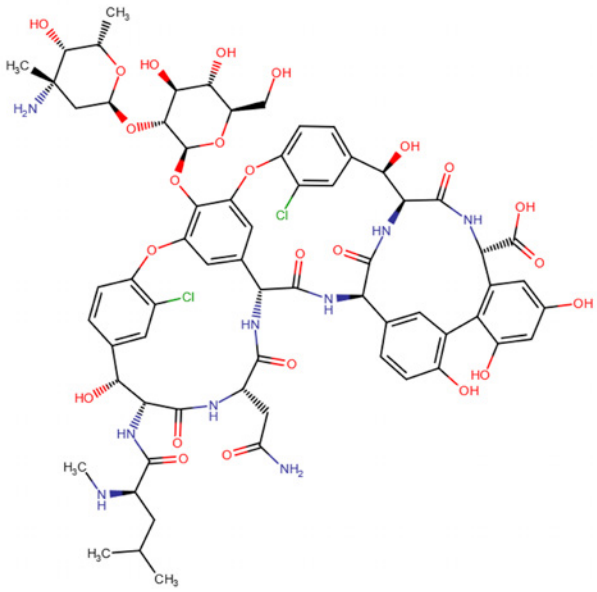
Chemistry data is not just simple alpha-numeric data. The most appropriate description of chemical compounds utilizes quantum physics and complicated wave functions. For practical purposes, however, a simpler representation that flattens the three-dimensional (3D) structure into a two-dimensional (2D) graph representation that will display on paper or a screen and can be stored in a database fulfills most of the needs of researchers. In this 2D graph, nodes represent atoms and edges represent the connections between nodes (i.e., bonds). To enable an unambiguous representation of more complex chemical structures, extra data may also be attached to an atom or bond, or to a group of atoms and/or bonds.

It should be noted that every little piece of information in a chemical representation is important. In particular, stereochemistry and stereochemical representation is crucial and can become very complex, for example in the case of only partly known stereochemistry ("advanced stereochemistry"). An up or down wedge for a bond or the presence or absence of a chiral flag can make the difference between a successful drug and a harmful compound.

In order for chemistry data to provide maximum value for research (e.g., when investigating structure–activity–relationships, when using your data as a knowledge base for chemical synthesis, when building machine learning models and artificial intelligence applications from your data, etc.) the repository where your chemical data is stored must satisfy the following:

- Accurately represent a compound's chemical structure, including stereochemistry
- Uniquely identify the chemical structure and support deduplication
- Strictly follow the business rules defined
- Permit searches of compound classes that share structural characteristics

Figure 2: Chemical structure of the antibiotic vancomycin, a last resort medication for the treatment of sepsis and lower respiratory tract, skin, and bone infections caused by Gram-positive bacteria.[3]



Chemistry Data Migration Challenges

Migrating complex chemistry data presents significant challenges that are mainly due to lack of standardization within the industry. A number of factors need to be considered when migrating chemistry data from one system to another:

Different chemical structure formats – There are several different chemical structure representation formats used to store and transfer chemical data (e.g., molfile, SMILES, InChI, CDXML, etc.). These formats are not fully feature-equivalent (see fig. 3), potentially resulting in incomplete information about the chemical structure being migrated (e.g., information about enhanced stereochemistry may be lost or even incorrectly interpreted by the new target system).

Different chemical business rules – Different companies, sites, departments, chemical systems may have different conventions for chemical structures (handling of salt forms, drawing conventions of chemical structures, definition of duplicates, etc.).

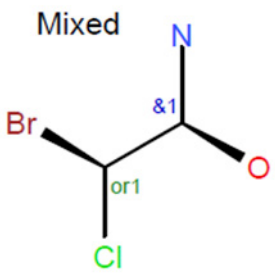


Figure 3: While the molfile format is able to capture the advanced stereochemistry information in the example structure, SMILES and InChI cannot handle this type of information.

SMILES: N[C@@H](O)[C@@H](Cl)Br

InChI=1S/C2H5BrClNO/c3-1(4)2(5)6/h1-2,6H,5H2/t1-,2+/m1/s1

Incompatibility of different chemical cartridges and/or software – Most chemical data in industry is stored in relational databases such as Oracle, with a "chemical cartridge" extending the capabilities of the relational database management system to 'talk chemistry' (e.g., run chemical searches). Chemical cartridges from different vendors are not 100% compatible with each other. Typically, when you move from one system to another, the underlying chemical cartridge that interprets the structural representation changes. Migrating chemical structures to a different chemical cartridge (or cartridge version) presents a risk that may impact structure representation and/or identity. Similar considerations apply to other chemical structure handling software, such as chemistry toolkits. In addition, each chemical software vendor has its own interpretation of chemical representation.

Inconsistent chemical data – Depending on if chemical business rules existed and to what extent they were enforced, chemical data may be heterogeneous in the sense that the same chemical feature is represented differently in different chemical structures.

Dirty chemical data – Inevitably, some stored chemical data will be dirty (e.g., errors in chemical structure, incorrect chemical structures, misuse of features, etc.) and may need to be cleaned before it is loaded into the new system.

Chemical reaction complexities – There are additional complexities to consider around how chemical reaction data was stored. Are the atoms of the reactions mapped to identify the changes in the participating molecules of a reaction? Does the data include multi-step reactions? Are the conditions, reagents and solvents stored as embedded objects in a drawing file? What if the new system does not support multi-step reactions or enumerated product libraries?

All of the above aspects need to be carefully considered for a chemistry data migration in order to preserve all the valuable information contained in the data. A low-quality chemical data migration can have a number of negative consequences for your business. Primary effects are:

- Important chemical information is lost.
- Chemical information is altered to a different meaning.
- Data of a chemical compound may be spread over several records.

Because of these primary effects, you will face even more serious downstream effects, such as (but not limited to):

- Search and retrieval are hampered.
- Reuse of data is restrained or even made impossible.
- Data mining and other kinds of data exploitation are hindered.
- Use of data for building machine learning models or to train artificial intelligence becomes unfeasible.
- Ultimately, all this will lead to poor decisions.

The key takeaway here is chemistry data migration is both very complex and very important. This data is essentially the crown jewels of your business. Any attempt to move it should be taken seriously and undertaken very carefully, as the potential exists for significant damage to your business if the information and insights contained in your chemistry data are not preserved in the migration.

Chemistry Data Migration Planning

If you want to preserve your assets and IP in a chemistry data migration, it is important to enlist a team of highly skilled and experienced data migration professionals with deep domain knowledge in chemistry, chemistry labs, and chemistry research. In addition, you need to involve your internal experts who have a deep knowledge of your local set-up, for example, your local chemical drawing rules, or the history of the data sources.

But even with a skilled chemistry data migration team in place, there will still be unforeseen issues that will arise due to the complexities involved in the process. Your team needs to be able to take these surprises in stride and improvise to get the job done in ways which preserve the integrity and meaning of your data.

Because of the complexities involved, chemistry data migrations should be treated as a proper project, with associated project management. It is important to balance quality vs. effort when migrating your chemistry data. As outlined above, quality is key, but a 100% perfect migration will be time and cost-prohibitive, if it's even possible.

Agreed-upon quality requirements will help to keep the migration project in budget and on time. As the project proceeds and you learn more about the unique obstacles and surprises involved, it will likely be necessary to adapt the project plan to keep proper balance between quality, time, budget.

A thorough planning phase is essential for success. During migration planning, key stakeholders are identified and briefed, dependencies and requirements are examined, migration scenarios are developed and tested, and a comprehensive migration plan is created that documents the what, why, how, and when of the migration.

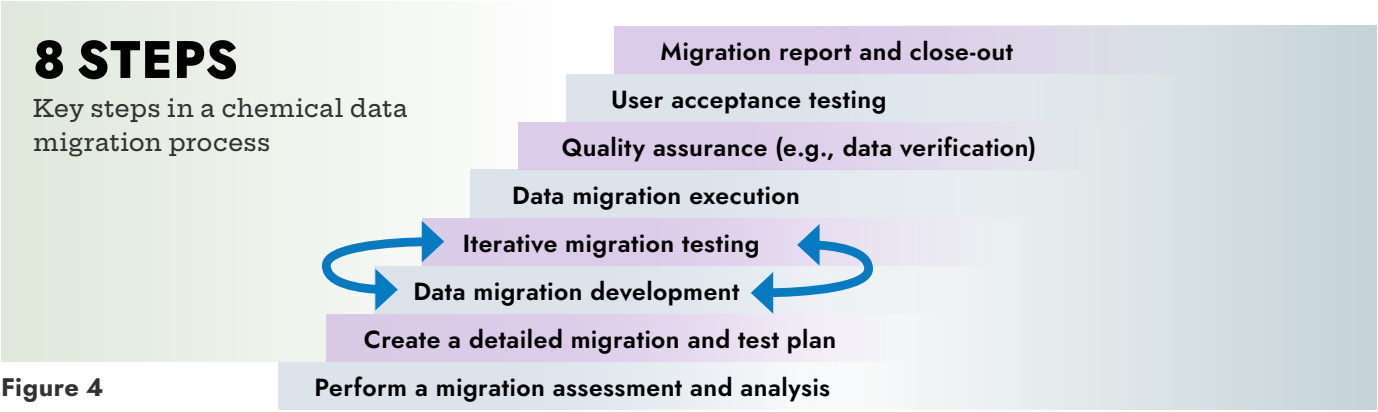


Figure 4

Migration assessment and analysis – Before migration begins, an analysis/assessment should be conducted that examines the data structure of both the source and target. There is no way to perform a quality data migration without an in-depth analysis of the data to be migrated. Assessment of the source data should include basic metrics such as the volume of to-be-migrated data but should also include aspects such as:

- Value of to-be-migrated data (e.g., relevance to current research activities and likeliness of reuse)
- Quality of to-be-migrated data (is the data normalized? errors in data?)
- What chemical entities (e.g., structures, reactions, batches) and what associated data and metadata are to be migrated
- Chemical features present in the source data; for example, enhanced stereochemistry features, or features used to describe chemical complexes and organometallic compounds
- Compatibility with target system for the features present
- Options you have for chemical features which are not directly compatible

Extra considerations may be necessary if the source data includes parallel synthesis and combinatorial chemistry data and/or large molecules (for example, proteins, biopolymers, etc.). Other, more general items that should be looked at and analyzed in this initial phase include:

- Target system – intended future use (e.g., machine learning, AI, automated synthesis planning, etc.) and business rules (e.g., duplicate handling)
- Possible migration paths – which tools are being used, and what are the limitations of each alternative?

- Formal migration requirements and acceptance criteria
- Risk analysis, including impact assessment and mitigation strategy
- Level of normalization and data cleaning required
- Is there data that should be migrated as-is or even not migrated at all?
- Archiving and long-term preservation of the original legacy data

Migration Plan – Alongside the migration analysis, a data migration plan is developed that takes into account the needs of all the different stakeholders involved (particularly the end users). The migration plan should cover the following topics:

"... effective planning and execution of a chemistry data migration will not be possible without deep domain knowledge and expertise in both chemistry and chemical data migration."

- Migration scope and timelines
- Cutover plan
- Data quality rules management process for new system
- Tools that will be utilized
- Supporting change management activities
- Testing strategy
- Validation strategy
- User acceptance and formal user acceptance testing
- Training
- Data dependencies
- Migration documentation
- Documentation for end users

Due to the complexities involved in chemical data migration, this plan will likely need to be refined and adjusted as unforeseen challenges become apparent. That said, planning effectively and getting the right people, processes, and tools in place can go a long way towards helping organizations complete a chemistry data migration on time and on budget. It should be noted, however, that effective planning and execution of a chemistry data migration will not be possible without deep domain knowledge and expertise in both chemistry and chemical data migration.

Chemistry Data Migration Execution

Once analysis has been completed and the Chemistry Data Migration Plan is in place, it is time to execute the data migration. Like any other data migration, chemistry data migration will involve three distinct steps – extract, transform, load (ETL) – to migrate the data into the new system. While other forms of data migration often utilize commercial and open-source ETL tools to automate the migration and avoid involvement from developers, chemistry data migration typically requires development of customized scripts that demand in-depth knowledge in both chemistry and IT (e.g., SQL). Best practice for this customized code is to emphasize reproducibility (i.e., consistent results with the same input data), enable validation activities to confirm expected values in each ETL stage, and to any avoid manual "ad-hoc" intervention.

Normal data migration projects are often designed in a classic "waterfall" fashion with the analyze, develop, and execute phases all performed in a linear manner. However, due to the complexity of chemistry data migration and the various types of surprises which will inevitably arise, an iterative, agile approach is a far more effective methodology. Several test migrations should be undertaken prior to the actual execution in order to validate requirements, ensure completeness, identify and analyze errors, examine results, and make improvements before the actual execution in the production environment is undertaken.

Chemistry Data Migration Quality Assurance

After the chemistry data migration has been executed, business and compliance risk demand data verification to ensure the loaded data is accurate, complete, meets requirements, and supports processes in the new system. While 100% data verification via automated tools is possible in many data migration projects, data complexities inherent in chemistry data necessitate that certain verification processes be applied manually based on the risk analysis done in the planning phase. Forensic intuition combined with the experience built in previous chemical data migration projects can be an invaluable aid in this process.

As with the other phases in chemistry data migration, verification processes (including any necessary script development) should be undertaken by professionals with deep chemistry domain knowledge to ensure the process is effective and conclusive. Data errors identified in the verification phase may lead to an iterative process where data migration phases are repeated several times before the system is considered fully verified and deployed. All aspects of the verification process must be documented.

Completing Chemistry Data Migration

A formal user acceptance testing (UAT) should be conducted to complete the data verification process and ensure that the migration meets user requirements. Any issues that are discovered should either be addressed, or at least documented for the user community. Once UAT is complete, the original legacy data should be transferred to a long-term archive and legacy systems may be decommissioned. Finally, a migration close-out meeting should be conducted, and a migration summary report prepared, to officially end the chemistry data migration process.

"... end users should now be able to take advantage of the enhanced capabilities of the new system by searching, analyzing, mining and otherwise exploiting the migrated data."

While there may be some limitations created by balancing data migration efforts with time and budget constraints, scientists and other end users should now be able to take advantage of the enhanced capabilities of the new system by searching, analyzing, mining and otherwise exploiting the migrated data, alongside any new data they are creating in the new system. Assuming the migration was performed by a team of specialized experts with the necessary domain knowledge and experience, you will most likely have achieved your project goals (e.g., timeliness and budget) and your business goals (e.g., preservation of your IP and the ability to reuse and exploit your legacy data in new sophisticated approaches to research).

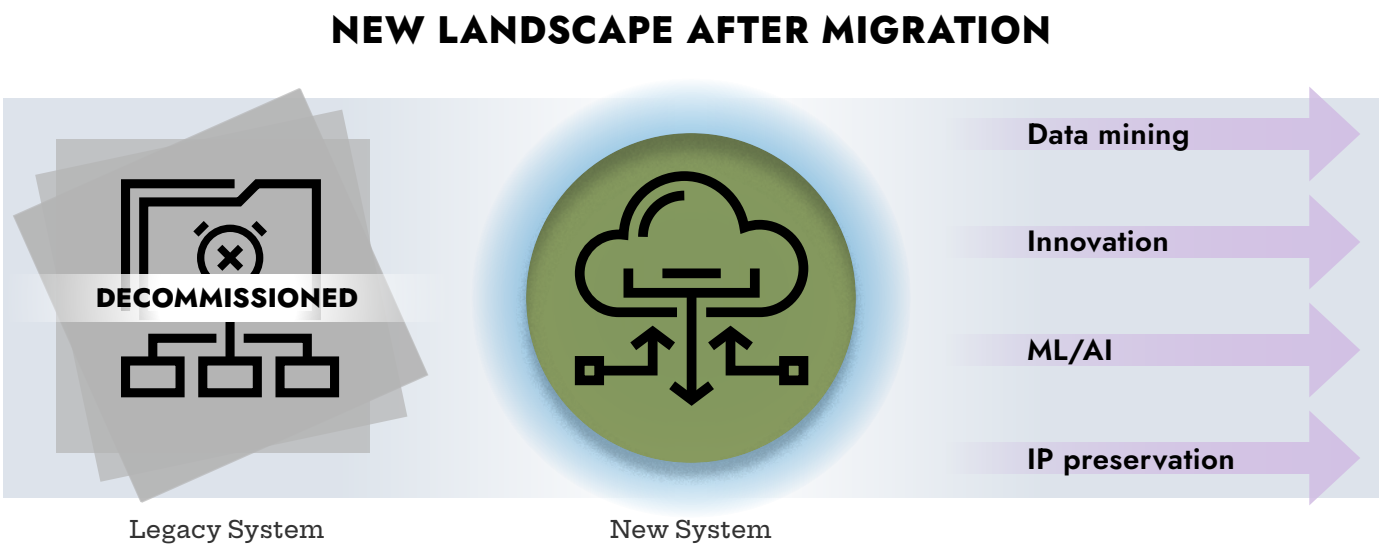


Figure 5: System landscape after completion of the data migration.

CASE STUDY 1:

Global Pharmaceutical Company Acquires Specialty Biopharmaceutical Company

OVERVIEW

After completing an acquisition of a specialty biopharma company, a global pharmaceutical company hired an experienced chemistry data migration team with deep chemistry domain knowledge to integrate chemical compounds from the acquired company into its compound database.

BUSINESS DRIVERS

The main goal of this project was to improve R&D efficiency and collaboration by making all chemical assets in the company (including the new compound data from the acquisition) accessible to all users, in a single system.

MIGRATION CHALLENGES

The acquired company used different ELN systems to capture their chemical and biological data, and the drawing rules for chemical compounds in the different ELNs were not consistent. In addition, the chemical data had undergone multiple migrations from one system to another before, which further deteriorated data consistency. To migrate the different data sets into the common database, all chemical structures had to be normalized according to the drawing rules of the global pharmaceutical company. Consequently, this project required deep chemistry and cheminformatics knowledge to ensure that chemical meaning was not compromised in the migration and that structures were correctly loaded into the target system.

SERVICES PROVIDED

The migration team provided the following services to the client:

- Analyzed to-be-migrated compound data.
- Normalized to-be-migrated compound data to comply with business rules and ensure that chemical meaning was not compromised.
- Migrated chemical compound data, with processes in place to merge any already existing compounds to eliminate duplicates.
- Verified that chemical structures were correctly loaded into target system.

RESULTS DELIVERED

All migrated chemical assets were successfully loaded into the new system and verified to meet company chemical business rules and chemical structure quality standards. This project greatly improved efficiency and effectiveness of the company's R&D activities due to having all chemical compounds being accessible to all research scientists in a single system.

CASE STUDY 2:

Data Migration Quality Assessment for a Global Pharmaceutical Company

OVERVIEW

A global pharmaceutical company implemented a new cloud-based electronic laboratory notebook (ELN) across multiple sites in a global harmonization initiative. There were several different legacy ELN's in use for synthetic chemistry research at different sites that contained chemical reactions and associated data that needed to be migrated to the new system. The new ELN vendor performed the technical migration. Due to time and budget constraints, normalization, cleaning, or fixing of chemical data were not part of the project scope. To restrict the impact of this limitation, the company hired an experienced chemistry data migration consultant with deep chemistry domain knowledge to perform a quality assessment on the migrated synthetic chemistry data.

BUSINESS DRIVERS

There were a number of business drivers fueling this initiative:

- Provide a single system for all scientists globally.
- Increase knowledge sharing and collaboration among scientists.
- Align with the company IT strategy (application consolidation, cloud).
- Allow scientists to search legacy synthetic chemistry knowledge alongside new experiments.

MIGRATION CHALLENGES

The legacy source systems and target ELN were all different vendors with different chemical formats, which added significant complexities to the migration. Other migration challenges included:

- Considerably different 'philosophies' in chemical data handling between source and target systems.
- Major differences in supported chemical structure features between source and target systems.
- 'Complex' chemistries, such as organometallics.
- Source data locked in a proprietary system with restricted export capabilities.
- No control enforced when source data was captured, leading to high variability in chemical structure 'drawing style' and in data quality.

This project required deep chemistry knowledge and understanding of involved chemical formats and systems to perform the analysis, and also to understand subtle but important unwanted alterations of chemical structures during migration.

SERVICES PROVIDED

The consultant provided a number of services to the client:

- A detailed analysis of migrated chemical data to determine if data was inadvertently lost or altered during the migration.
- A full quantitative analysis of the different types of unwanted alterations.
- Complete technical documentation and migration-related change management for end users.

RESULTS DELIVERED

Based on the analysis performed by the consultant, the company was able to make a rational decision about whether the quality of the migrated data was acceptable. The analysis also facilitated a responsible use of the migrated data by company scientists. Scientists were able to understand the limitations of the migrated data and factor this information into decision-making, avoiding making wrong decisions based on doubtful data.

Conclusion

Chemistry data migration is a challenging endeavor, and many companies make the mistake of underestimating what is involved and at stake when conducting a migration. Software vendors often prioritize the system implementation/configuration, with little focus on data migration. In addition, vendors for the new system will usually not be familiar with the legacy technology platform and its chemical representation nuances. To move your chemistry data from one system to another and successfully preserve the meaning and information contained in the data, you need a team of experienced data migration specialists who understand both systems.

While it may be tempting to rely on a consulting company that offers generic data migration services to facilitate your chemistry data migration, this is a big mistake. The complexities of chemistry data migration make the vast majority of "generic" research informatics consultants unsuited to manage the process and preserve your R&D assets and your IP. Instead, your team needs to have deep chemical representation domain knowledge to be able to understand and preserve the nuances in your data. You can't throw bodies at this, and you can't train someone to do it. You need people with decades of experience and highly specialized knowledge. Even with an experienced team, there will still be surprises. While the time and money spent on migrating your chemistry data properly will not be insignificant, these costs are much lower than what will be experienced by negatively disrupting your innovation pipeline due to poor quality data or lost IP.

About Kalleid: **Mary Beth Walsh** is the founder and President of Kalleid, Inc, a boutique IT consulting firm that has served the scientific community since 2014. Kalleid works across the value chain in R&D, clinical and quality areas to deliver support services for software implementations in highly complex, multi-site organizations. At Kalleid, we pride ourselves in supporting the success of your IT projects and overall organizational transformation efforts with a wide range of interconnected services. We have built an extensive network of collaborative partnerships that allow us to staff projects worldwide with high quality professionals to get the job done right for our clients. For more information, visit **kalleid.com**.

About StructurePendium: **Gerd Blanke** is founder of StructurePendium Technologies GmbH. StructurePendium Technologies GmbH offers consulting services in the area of chem- and bioinformatics with a major focus on standardization and normalization of chemical structures and reactions for registration and retrieval processes. These services are provided in the context of database mergers, data transfers between different vendors, and data analytics. For more information, visit **www.linkedin.com/company/structurependium-technologies-gmbh**.

About Thomas Doerner: **Thomas Doerner** is an independent specialist for research informatics in life sciences and chemistry. Located at the interface of R&D and Informatics, Thomas helps clients in pharma, biotech, and chemistry define, design, and implement solutions that enable scientists, foster more effective R&D, and lay the foundation to achieve better outcomes faster. Since starting his own business in 2010 Thomas has contributed to several successful chemical data migration projects. For more information, visit **tdoerner.eu**.

About the Informatics Alliance: **The Informatics Alliance** is a small group of dedicated chem- and bioinformatics experts focusing on serving the life science, agro and chemical industries. Each of us brings many years of experience with research informatics projects and practical implementations. We operate independently but we know and help each other, sharing experiences and expertise, and for bigger projects we join forces, for the benefit of all our clients. Mary Beth Walsh, Gerd Blanke and Thomas Doerner are members of the Informatics Alliance. For more information, visit **www.informaticsalliance.org**.

[1]"Risks and Challenges in Data Migrations and Conversions," Gartner, February 25, 2009. Ted Friedman.
Available at: https://www.gartner.com/en/documents/897512/risks-and-challenges-in-data-migrations-and-conversions

[2]"Data Migrations Begin (and End) with Data Quality," Experian.
Available at: https://www.edq.com/globalassets/white-papers/data-migrations-begin-with-data-quality-whitepaper.pdf

[3]Chemical structure provided by ChEMBL.
Available at: https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL262777/

WWW.KALLEID.COM
WWW.LINKEDIN.COM/COMPANY/STRUCTUREPENDIUM-TECHNOLOGIES-GMBH
WWW.TDOERNER.EU
WWW.INFORMATICSALLIANCE.ORG