Research Ideas and Outcomes 9: e105197
doi: 10.3897/rio.9.e105197

Research Article

Digital transformation strategies for applied science domains

Samuel Bentum‡, David J Wild‡

‡ Indiana University, Bloomington, United States of America

Corresponding author: Samuel Bentum (sbentum@indiana.edu)

Reviewed v 1

Academic editor: Editorial Secretary

Received: 19 Apr 2023 | Accepted: 20 Jun 2023 | Published: 09 Aug 2023

Citation: Bentum S, Wild DJ (2023) Digital transformation strategies for applied science domains. Research Ideas and Outcomes 9: e105197. https://doi.org/10.3897/rio.9.e105197

Abstract

The key hallmark of a digitally minded organisation today is seen in their rapid advancement, globalisation, innovation and resilience to change. Companies that wish to thrive must be prepared to adapt to the new digital reality. Being digitally minded does not mean implementing new technology, investing in tools and upgrading current systems. These stages are critical, but they are not the entire picture. If a company wants to remain competitive, it must not just be able to adapt to changes, but also anticipate and drive innovation. Companies must plan ahead and be proactive architects of their future in order to achieve this vision. This is where a digital transformation strategy is crucial. A digital transformation strategy assists organisational leadership in addressing challenges about their business, such as the present level of digitisation and a digital maturity roadmap. Although diverse data capturing technologies and data-generating assets exist, material/chemical science domains, such as R&D and Manufacturing groups, struggle to harness the full power of their data. A typical industry will have significant data sources generating large amounts of data stored in siloed databases with minimal to non-existent cross-talk. This in part creates scenarios for researchers to be able to perform a deep dive in one set of data, but unable to co-populate and harness the interdependences or relationships amongst the different datasets. This paper seeks to define, distinguish, aggregate and propose an integrative approach to utilising the various types of disparate data sources commonly encountered by researchers in the field of their material science research. The main focus here is defining strategies to harness insights across integrative data to aid in

efficient research in R&D organisations as these industries seek to embrace the power of digital transformation. Although the principles described here relate to industries in the applied science domain, the general strategies proposed can be applied to other industries on a case-by-case basis.

Keywords

digital transformation, digitisation strategy, applied science, data management, systems integration, KNIME

Introduction

Digital transformation is occurring across organisations, from the pharmaceutical industry to the service industry, bringing such benefits as better decision-making and faster processing as information is shared instantly. A decade ago, prospects for data-driven progress were hardly imaginable, but now they are becoming more and more possible because to falling computation costs and more accessibility to cloud-based analytics. Chemical plants create enormous volumes of data, much of which has the potential to be used to improve efficiency and raise yields, similar to the majority of large-scale industrial facilities. For instance, according to recent research, chemical producers might boost their return on investment by as much as 5% by merely digitising their product processes (Kumar 2019). Moreover, digital monitoring of an organisation's energy use may assist in lowering dependency on low-efficiency fuels and feedstocks, significantly enhancing consumption-to-yield ratios across numerous production pipelines. The materials and chemical research and development (R&D) organisations can also benefit from digitalisation and the use of data-driven tools, such as machine-learning algorithms. However, to harness these benefits, material science industries have to build and maintain comprehensive data aggregates from the disparate sources of data generated during the research phase. As has been widely reported, the challenge of disparate or siloed data sources inhibits the collective use of the full spectrum of synthesis to performance data of materials developed (Morgan 2020, Vasudevan et al. 2021, Gao et al. 2022). The US government's Materials Genome Initiative (MGI) highlights the need for comprehensive materials data infrastructure to accelerate innovation (Hernandez 2018). An initial National Institute of Standards and Technology (NIST) MGI report attributes significant cost saving worth billions of dollars in economic value in the United States alone to next-generation materials innovation infrastructure (White 2012, Hernandez 2018, de Pablo et al. 2019. To realise this benefit requires a paradigm shift from the traditional approach of data management and use in the industry. Known legacy data management approaches have led to situations in the industry where:

1. Decisions driven by experimental data are limited, based on the time and cost associated with generating said data;
2. Existing experimental data are sparse and incomplete;
3. Data analysis and statistical methods are rarely used; and

4. Access to historical data is poor, causing repetition of experiments and limited learning from past work.

To curb the above challenges requires an integrative data management approach which allows researchers to accurately analyse and understand their results as they try to optimise the properties of a new material. If simulation techniques or machine-learning are being used to reduce the number of experiments needed, then integrative data management becomes even more important, as algorithms need structured information to work.

According to the MGI programme, developing new materials for next generation application use can be a much arduous, time-consuming and very expensive process. A typical development cycle spans between 5 to 10 years (de Pablo et al. 2019, Olson 2000). The development process involves design of experiment, formulating, processing, testing and production. Each of these stages inherently produces troughs of data associated with the selected experimental design and they typically occur within different research groups or departments. Hence, each experiment tends to produce data that resides in different systems across the organisation. As an example, during material synthesis, the processing parameters, such as pressure, temperature and rate from the chemical reactors, are stored in the computer (PC) peripherals attached to the individual equipment in use or a network device that aggregates all the parameters for the reactor. This provides a rich source of data when troubleshooting reactor issues; however, it may not necessarily provide any context by itself to a researcher who is evaluating the performance of the final tested product. In the same way, data generated from the analytical characterisation of the new material developed tends to be siloed within the analytical department's databases. These example data sources tend to be segregated to their respective department, offering minimal integration for drawing data insights. A frequent challenge highlighted around these types of data is their different native formats not being compatible across the development spectrum. While that is accurate, recent advances in material informatics do highlight new technologies that can transform different data types into unique formats, consequently ingestible for machine-learning (Berthold et al. 2009, Warr 2012, Dwivedi et al. 2016).

A fully-integrated stream of R&D data sources enables researchers to ask questions that can accelerate their product development. A typical objective of a research group might be exploration of virtual product development geared towards faster material development and better understanding of certain composition and performance yields. Another area of interest resides in the in-silico synthesis of materials or chemicals, whereby in new product or process development, researchers are able to determine upfront which process, composition or catalyst are needed to produce products with desired properties and yield. In most cases, there are data available which do not cover the entire research life cycle; process/chemical structure/property/performance. Hence, it is essential to include advanced calculation constraints when attempting to optimise inputs and outputs in a predictive synthesis model. An example of such a case would be studying structure property relationships in-silico where structure is defined as an input as opposed to process data which may not be complete or accessible easily. Full utilisation of advanced

analytics and modelling tools requires transparent data access for researchers working in this industry.

Impact of COVID-19 pandemic on industries

COVID-19 changed R&D work across many scientific areas. Labs were compelled to handle operations remotely and scientists relied on digital technologies to keep vital research moving forward. Some jobs were amenable to conversion to the remote world. Instead of meeting in rooms, researchers collaborated through Zoom®. They planned work schedules and carried out data analysis from their homes. Tasks that needed physical presence in the facilities to prepare and load samples, as well as study analogue data in real time, were much more complex. Researchers had to struggle with ensuring the right process parameters and reagents were fed into their synthesis tools remotely. One of the greatest challenge faced by scientists in the chemical industry space was real-time access to data generated in their labs remotely during the pandemic era. This issue was magnified during the pandemic era, though it has always been a challenge for non-digitalised R&D organisations. The effects of the pandemic's work-from-home rules in 2020 and 2021 have provided many companies the chance to re-evaluate the amount of technological debt they are carrying with legacy informatics systems and how outdated methodologies do not help the digitisation of their business. This has enabled organisational leaders (including those in the chemical sciences domain) to speed up comprehensive modernisation plans for their research laboratories and process operations.

Process flow chart of a typical R&D operation

The commercialisation of new product and chemistries is a very complex and resource-intensive process. The time taken from conceptualisation of research ideas to final product delivery to customers is an expensive journey filled with multiple iterations of development failures and small successes. Take for an example the process required to come up with a new material for the industry as depicted in the chart in Fig. 1. The process begins with an ideation step on what researchers would like to achieve with a set formulation. This step usually is derived with the goal of either making a new material tailored for a specific segment of the market and customer or for a broader commodity market. In either case, the next logical step involves several iterations in the development cycle to deliver a robust and stable formulation that is further enhanced with inputs from select commercial interest groups. The process of enhancing the new product is carried out through sampling and testing for specific properties' requirements to meet a market need. This particular stage can be quite exhaustive with repeat processing, sampling and testing batches until the desired properties are met. Having completed this cycle, the new product is scaled up commercially through manufacturing plants and after which commercial revenue streams starts flowing into the organisation.

As discussed in the previous section, the process to fully develop and commercialisea new product on the market requires extensive time commitment and resources. From Fig. 1, it

is clear that the development of any new material requires collaborative efforts from different part of the business right from the ideation stage. The most time-consuming part of the process lies within the design and development cycle. It is at this stage that most of the experimental data are generated by different labs and testing groups. Organisations with complex stand-alone systems or labs suffer greatly in fully utilising data generated from research work for which they have spent a lot of money. Unfortunately, this scenario is all too common in the chemical/material science industry where true digitalisation of assets are still at its infancy when compared to Pharmaceuticals and FinTech organisations.

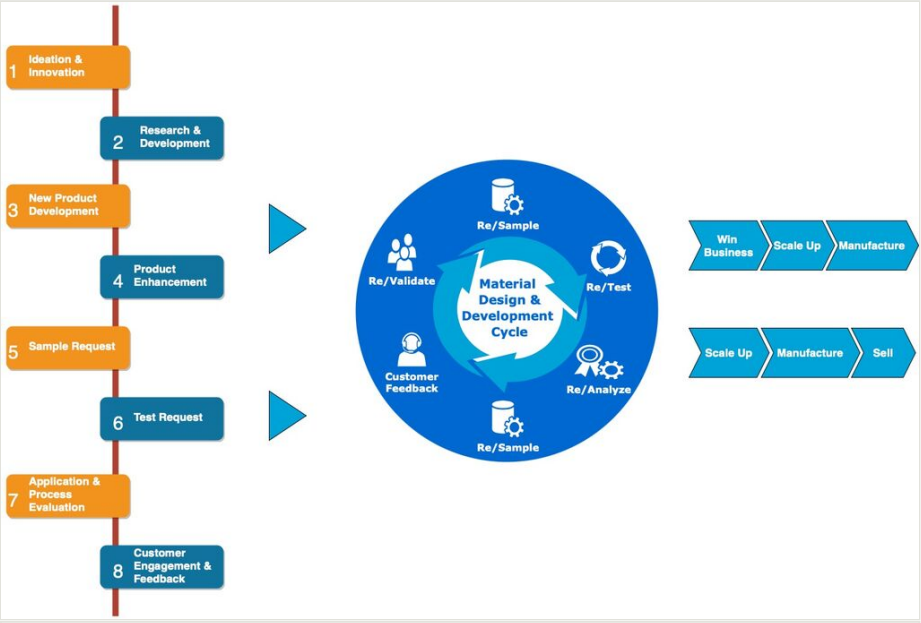

Figure 1. doi

Material commercialisation is difficult, time-consuming and expensive.

In an effort to truly harness the power of the huge trough of data generated in today's research labs, organisations require a digital strategy covering data management and lab connectivity protocols. To be clear, implementing a digital strategy requires more than merely upgrading the technology in laboratories; rather, the entire corporation must be considered. Such an initiative re-imagines how people work and engage with one another, focusing as much—if not more—on people than technology. The key to a successful transformation is a deliberate focus on talent and skill, which enables employees to integrate their scientific duties with new technology, which in turn, connects the many laboratories to one another across the company. However, it is crucial to pay great attention to each individual lab type and the difficulties particular to the people and procedures in each environment. Despite the fact that each type of lab is distinct and has its own special skills, procedures and technology, they are all interrelated and mutually beneficial. As a result, all laboratories should be evaluated and considered as equally capable of creating increased value.

Integrating technologies

Gaining access to all of your data sources is the first step in doing end-to-end data science. The process of gathering and shaping data from any source within an organisation is the core definition of integration. There are a number of unique technologies in the market now that address the different types and levels of integration (Hendler 2014; Miller 2018).

Data integrity

The data entry and management of an organisation's research is the dawn of their digital transformation journey and this goes on to show how far chemical and material science domains must go to attain AI-ready datasets. A fundamental step for any industry on this journey is to ensure the availability of its data in suitable formats logged electronically into a database system. However, as previously shown, the chemical and material science R&D is behind on data connectivity; even now the broad adoption of electronic lab notebooks is not widely applicable. The daunting issue of overcoming data silos have not gained much traction either. The good news is that there have been a number of solutions come on the market (Hendler 2014); however, until the chronic issues are tackled, the transformation can only go so far.

Structured vs. unstructured data

Over the last few years, data have become what is knowns as the next "oil" for organisations (Hirsch 2013) and internal data are an important component of every business' intellectual property (IP). However, just like oil, data from these organisations do not come pre-structured and ready for artificial intelligence (AI-ready). To fully harness insights from messy data will require significant iterations of data cleansing and structuring. Each stage of this exercise can yield further insights into the data and that presents opportunities to exploit early successes for decision-makers. It is important to understand that different formats of data are classified differently according to their structure stored. For example, the majority of experimental data/notes acquired in the lab are unstructured in nature. However, most time-series-based process data tend to be more structured.

Machine-learning algorithms can quickly comprehend structured data, which is often classified as quantitative data. Structured query language (SQL), is one of the most widely-used computer languages for managing structured data (Lu et al. 1993; Nagy 2017). Organisations can easily explore and manipulate structured data by utilising a relational (SQL) database.

However, unstructured data, often known as qualitative data, cannot be handled or evaluated using standard data tools and procedures. Unstructured data are best maintained in non-relational (NoSQL) databases since it lacks a specified data model. Another option for managing unstructured data is to store it in a raw form in data lakes.

Most recently, companies have hailed the advantages of AI in chemical and material science; with key focus on the speed of discovery, as well as the much simplified material compatibility assessment. The hidden truth most people do not hear: with poorly managed, unstructured data, none of these possibilities is conceivable. If organisations go into AI assuming that decades of stitched-together excel files would be the magic wand, then, unfortunately they are on a far longer (and more expensive) road than they ought to be on. It is not far-fetched for industries to start delving into their dataset only to find out significant portions of the data captured are unstructured and not as complete as they initially thought it would be. Not surprising in 2019, researchers at Deloitte (Smith et al. 2019) published an article referencing only 18% of organisations reported being able to take advantage of unstructured data. This leaves a staggering 82% of organisations underutilising their most valuable resource – data. To assist companies in overcoming the common obstacle of scarce, highly dimensional, biased and noisy data, it is imperative that the right connectivity is drawn to leverage the different types of data available. This would require an understanding of the types of data being generated and stored during the research phase.

In a typical material development workflow, researchers begin with formulation and synthesis, extrusion/molding, analytical characterisation and, finally, quality assurance release testing. This sequence highlights the different departments and data sources impacted in a development cycle. Each of these departments produce varying data types which are then housed in their specific domains and are usually read by specific software modules. As an example, the synthesis group will store data around pressure, temperature, viscosity and time during the reaction process. The formats of these data are naturally different when compared to results from an analytical test using gas chromatography or infrared spectroscopy. Coupled to these differences are also the fact that data are, most of the time, stored separately in the different departments (silo system) and there are limited to no cross-talk amongst these systems. The different formats and types of data generated through the development process in itself poses a major hindrance to researchers as they attempt to draw insights from historical data. A good practice for adoption by chemical and material science organisations is the power of leveraging systems integration. In the subsequent section, we will highlight the “as-is” situation in a typical R&D lab and propose unique ways of smartly integrating the different data siloes in research environments. This is the foundation for building an end-to-end data pipeline for all the disparate data sources needed to harness valuable insights for an organisation.

Integrating multi-data sources

The proper integration of data and technologies is critical for all downstream operations, such as information exploration and knowledge management (Adamides 2020). Organisations with effective data integration will be able to make better resource allocation decisions during the R&D process. Moreover, because there are no clear standards for integrating research data saved in electronic lab notebooks (ELN) or laboratory information management system (LIMS), integration can be done, based on business needs.

As an example, researchers working on developing new or improved materials or chemicals are typically bombarded with a host of datasets coming from all facets of the organisation about their particular project (Fig. 2). As described earlier, most of these data reside in their siloed system, with no clear and connected way for a researcher to combine and draw insights from all these datasets.

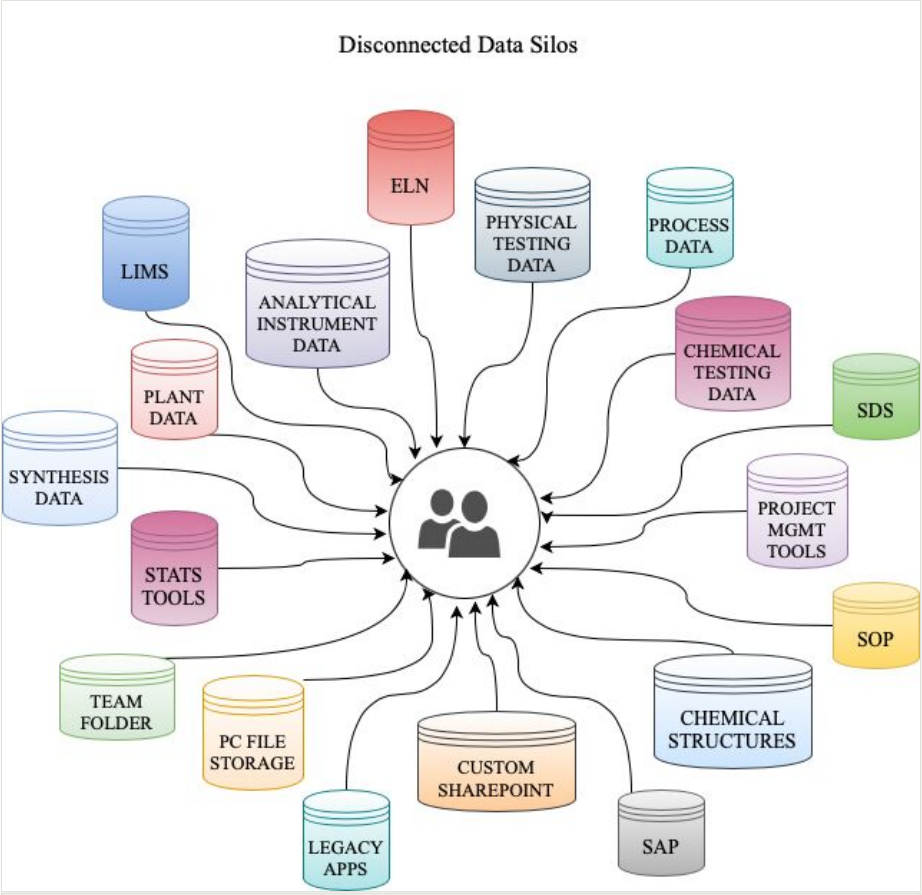

Figure 2. doi

A view of disconnected data silos typically encountered in a non-digitalised R&D lab (See Appendix for list of abbreviations).

As a case study, let us consider an organisational research need of exploring material property predictions, based on formulation, process and analytical characterisation of a polymer material. As discussed previously, the different process outlined above generate different formats of data and tends to be siloed in a different business unit in an organisation. Hence, it is important to understand what types of data formats are at play in this scenario.

Synthesis data stream (formulation)

Formulations are chemical and/or material compositions that are homogenous. To manufacture polymer composites, formulations can be mixes of solids or solvents-based components and can comprise of oligomers, fillers, pigments and other additive materials. They can be basic unreactive chemical blends, reactive mixes where the sequence of mixing matters or blends of blends where the ultimate molar composition is determined by the original blends' composition. All of this means that there is a plethora of conceivable components, each with its own set of limitations governing its kind, number, molar percent and mixing order. Typically new formulations are usually created in response to a customer request that includes several component and process limits as well as property goals. Winning or losing major contracts depends on swiftly recognising what you already have and what can be quickly altered or expanded to meet new business prospects. Buying, storing and processing hundreds of different components is expensive and there are apparent cost reductions to be obtained by rationalising the process. Ingredient costs vary and formulations that meet goal qualities while incorporating lower-cost constituents increase profits. As global events have an influence on supply chains, the capacity to swiftly reformulate utilising materials from a new supplier has become increasingly important. It is also important to remain adaptable when new regulation on prohibited chemicals takes effect. For data management and machine-learning, the formulation space provides a number of unique issues. A typical material manufacturer will have hundreds, if not thousands of different formulations for their products. Each formulation accounts for a specific materials' property that is tied to a business need or sales order. Thus, it is crucial for manufacturers to maintain the specific ingredients and quantifies that produce the unique property of the material specified for the market need. These formulations are typically recorded as ingredients names and weights; hence, the data are mostly in a flat file format in a database that can be extracted as csv or any tabular form. However, these data by themselves are not enough to harness and improve rapid re-formulation and new material development without the right process parameters captured. Organisations can generate huge economic value by leveraging a framework that properly captures complicated process flows, understands molecular structures and utilises the deep subject knowledge of corporate specialists.

Process data stream

Imagine watching a recorded cooking show on TV, but for some reason, the network cuts out after the chef shows the ingredients and amounts needed to make the meal. The network returns after the chef has completed the meal and it is served on the plates, essentially cutting out the entire cooking steps. The question then is, can the viewer make the exact same food as shown by the chef with the given ingredients without knowing the steps? The obvious answer is, no. Similarly, as discussed above, formulation of materials are key to organisational success on delivering their products; however, the process to convert the ingredients to the product is as important as the formulation itself. For each material or chemical made by companies, there are additional tonnes of (meta-)data optimised and collected for the process. These process data points are dependent on the

synthesis approach or equipment used in making the materials. During synthesis of compounds or materials, data such as temperature, feed rate, pressure, flow rate, screw dimensions and others are routinely recorded in their native software systems of the instrument used. Historically in the manufacturing world, these data were used to monitor and optimise production process, identify any potential production issues or system troubleshooting. Engineers and scientists also used these process data to help make data-driven decisions to improve the overall material synthesis and production performance. However, with the advancement of machine-learning and visualisation tools, these forms of data can be utilised to generate more than process upsets if the data are collected and stored properly.

Process data in R&D organisations or manufacturing, in general, can be formatted in a variety of ways, depending on the specific application and the type of data being collected. Some common data formats include:

1. Numerical values: This can apply to metrics like flow rate, pressure and temperature, which are normally expressed as numbers.
2. Time-series data: This kind of information is gathered over time and may be used to monitor changes in the process. It is often expressed as a succession of numerical numbers, each with its own time-stamp.
3. Log files: This sort of data is produced by machinery and equipment and comprises information on events that occur throughout the manufacturing process, such as machine start/stop timings, alarms and other metrics.
4. Images/Videos: In certain circumstances, process data may contain photos or videos acquired by cameras or other imaging equipment, which may be used to monitor and study the manufacturing process.
5. Additionally, some data are gathered via sensors, PLC, DCS, SCADA and other IoT devices. These data may be transferred and stored in a variety of data formats, including JSON, CSV and XML.

The data format used will be determined by the application's unique needs and the type of data being gathered. It is also critical to ensure that the process data are kept in a manner in which the appropriate stakeholders can readily evaluate and comprehend. Data files, such as numerical, time-series and log files, can be saved as CSV or TXT-based flat files.

Analytical instrument data stream

Analytical characterisation of a material is the process of identifying the material's chemical and physical characteristics using a variety of analytical techniques. These techniques can range from Gas Chromatography (GC) to Differential Scanning Calorimetry (DSC), Transmission Electron Microscopy (TEM), InfraRed Spectroscopy (IR) and many others. The main goal of using these techniques is to help provide understanding of a material's structure and composition at the microscopic level and potentially also aid in troubleshooting any impurities or performance defects of products. Analytical characterisation of materials may be applied in a variety of ways, including the creation of new materials, maintaining the quality of already-existing materials and analysing failures

in materials already in use. It is also important in the realm of materials science, where researchers utilise it to explore the characteristics and behaviour of many types of materials. Data from material characterisation come in various formats depending on the type of analysis being carried out. Some examples of material properties measured during product development include: composition of material, crystalline structure, morphology of blends, thermal properties, mechanical properties, rheological properties and a host of other measurement types. The data from analytical techniques are usually the most challenging form of data to deal with from the onset. Such a department would have access to multiple types of instrumentations with numerous softwares that run each item of equipment. It is important to highlight here that the different data generated from test instruments take on formats driven primarily by the software output settings. As an example, output files from a Gas Chromatography Mass Spectroscopy (GCMS) instrument are mostly saved in either mzML (XML based format) or JCAMP-DX (a proprietary based format). The output data for DSC instruments are also saved in either plain-text .DSC file or CSV format. Hence, if a researcher were to ask the question “how do I compare the mass spectroscopy m/z distribution pattern to the heat profile generated from the thermal analysis of a tested material?”, one would manually have to perform data analysis on each piece of dataset independent of the other dataset because the file formats are distinctly different. Today, there is no software in the market that can automatically read and combine all the different formats of data files from lab instruments together. However, in the example above, if organisations have a truly digital integration pipeline, researchers would be able to ingest all formats of data, albeit manually and read them in a unified output language they can analyse. In this case, therefore, the researcher can parse and convert the GCMS mzML data into the CSV file format using, for example, python nodes in a pre-built integration pipeline. With both sets of data structured in the same format, the user can then evaluate the results, based on a time-stamp as both sets of data can effectively be treated as a time-series for ease of visualisation.

In general, analytical characterisation of materials is a key step in deciphering the microscopic physiognomies and behaviour of materials, which may be utilised to enhance the performance and dependability of materials in a variety of applications. Of course, no single data point from one instrument tells the entire story; hence, the need to have a comprehensive set from all measurement data types.

It is also worth mentioning here that to truly harness the power of the above described datasets in performing predictions of new materials, another key piece of data point to consider is first principle modelling of chemical reactions (Pantelides 2013, Noble 2013). This is important as it helps to solidify the boundaries of algorithms in modelling various parameters within a design space.

Another importance aspect of dealing with the data from analytical instrumentation is ensuring data are stored in a common database, such as MySQL, PostgreSQL, MongoDB and others. Having data stored in the right system formats helps for a smooth data extraction process. The process to extract the data from the instruments' data sources can be programmatically encoded through scripts in languages, such as Python, R or MATLAB (a common language amongst research engineers in the industry). For ease of use, a no-

code/low-code solution is highly preferred in the chemical/material science industry segment as the level of programming language knowledge is not mature within the research scientist skillset. If coding can be avoided in the initially phase, it will help with gradually bringing key lab scientist up to speed with what an integrated dataset can bring to their research. Numerous data integration tools exist today that can be run with no-code experience needed to extract and combine data from different instrument API sources. Examples include Talend, Informatica or MuleSoft. Certainly, organisations can also build their unique pipeline of data integration platforms using the many configurable options, such as KNIME and SciTergic Pipeline Pilot tools.

Electronic Lab Notebook (ELN) data source

Traditionally, ELN systems are meant to replace paper notebooks with digital analogue documentation platforms. They are typically used in wide array of organisation sectors, such as chemical industry, pharmaceutical and food/beverage industries (Machina 2013). ELN improves efficiency by centrally keeping any information, data or intellectual property generated by scientists in a searchable database. With time-stamps, version control and record authentication, experimental documentation required to support a patent is safeguarded, maintained and shared in a high-security location. ELN is a document-based repository for both long text-based descriptors, chemical structure information, multi-format file types, such as XLS and image files. Hence, it allows users a very flexible data capture approach mimicking document entry in MS-Excel and MS-Word. The level of data structuring in the system has a significant impact on the reporting and searching capabilities of ELN, which seem to be uneven across industry platforms. On the one hand, "free-text" ELN are adaptable and may record any type of result, similar to MS-Word or TXT document. Although the findings in free-text are searchable or available via the experiment identifier, the data are not organised. As a result, aggregating findings from several studies to generate structure-property tables is difficult. Another drawback of ELN systems is the formulation or recipe data, recorded by scientists in their notes (and synthesis scheme), is not inherently connected to any test data. Hence, interconnected data values can be low and incomplete or unstructured enough to be reused by others or ML tools. Specialised ELN systems found in the pharmaceutical research domains (chemistry, pharmacology etc.) are more structured and it is possible to search by chemical reaction or use established calculation templates, for example, to convert in vivo raw data into findings (Machina 2013). Although doing that means generating more data silos which defeats the purpose of an integrated flow of data from labs.

In practice, the primary application of ELN in R&D projects is to replace paper notes, ease information flow and comply with intellectual property restrictions. Having said that, the ancillary data created through ELN serves as a rich metadata source to augment formulation, synthesis and process data transformation to machine learning (ML)-ready datasets.

Laboratory Information Management System (LIMS) data source

LIMS is a sample workflow-driven tool which is used to monitor and record all the data generated by a process. It is widely known for its sample test data management and consolidation (Gibbon 1996, Paszko 2000). The application is built on a SQL database platform, making it fairly straightforward to query and access data. LIMS are required for GLP compliance because they enforce standard operating procedures (SOP) and can identify and document deviations from such SOP. Due to their connections to tools and procedures, LIMS can simplify work, ensuring that SOP is followed and tracks data at every stage, from beginning to completion. In addition, they produce reports for specific tests, samples or research. As a result, they shorten the time it takes from request to result, increasing data production throughput and enhancing result quality. They do, however, generate data silos and do not allow for smooth data integration because they are related to a single process. They are designed to follow a single process and are almost solely employed in regulated environments where processes do not change frequently like manufacturing environment, as opposed to R&D, where conditions and procedures are routinely updated (Prasad 2012). This creates some drawbacks for such systems used in R&D groups within the chemical industry. As an example, LIMS may or may not pull data from lab instruments depending on the system set-up within a company. For global organisations where research is done in multiple geographical locations, Information Technology (IT) teams will set up LIMS to run on remote desktop clients to make the services available across the regions. Doing that means creating virtual networks for devices to connect to and exchange information. This can be a daunting task to bring all instrument PCs into such a network; hence, IT teams will typically enable domain specific internet protocol addresses to access the LIMS client servers. This limits the ability to pull raw data from instruments for researchers, even though test data from these instruments are often more structured and complete compared to those of ELN data.

A similar shortcoming of LIMS system by itself is that test data are not inherently tied to formulations or business data. Hence, test data of samples synthesised by specific formulations documented in ELN are not integrated comprehensively for a true end-to-end research workflow. This is an inefficient way of systems set-up as companies prepare to be digitally transformed.

From the data stream formats discussed above, the next logical step is to define a workflow-based integration approach that ingests and transforms data into ML-ready assets using a pipeline tool.

Data automation and custom workflows

Today, a significant majority of cheminformatics specialists and data scientists are increasingly using web servers for data processing and automation. The use of these web-based technologies lowers the barrier of computational requirements needed to process large chunks of data. In fact, server-side scripting languages like Ruby, PHP and ASP can be utilised to automate data processing, file manipulation and database communication

even via APIs (Kannan 2018, Groth 2020). These web-server-based operations are generally not free and cannot be hosted internally. However, there are other open source programmes that make it possible to create workflows and greatly simplify the automation of data processing. One of the benefits of these processes is their versatility and customisability to match individual demands. KNIME is likely the most extensively used open access environment (Warr 2012) and it is discussed more in this section. Of course there are other pipeline-based workflow programmes available on the market.

KNIME workflow platform

Data integration is a critical step in the data mining process, in which data from multiple sources is integrated into a single, unified data repository. There are many tools available to facilitate data integration, including ETL (extract, transform and load) tools, database integration tools and data warehouse tools. As discussed earlier, the open source data integration platform KNIME is one of the most popular tools for data integration.

KNIME (Konstanz Information Miner) is a free and open-source data analytics, reporting and integration platform. It is used in a wide variety of data-driven applications, including data mining, machine-learning, data visualisation and predictive analytics (Rückert 2009, Köster 2017, Ferreira 2018, Brunner 2018). It enables users to perform a wide range of data science tasks, including data pre-processing, feature engineering, model building and deployment, without requiring programming skills. KNIME supports various data sources, including spreadsheets, databases and big data platforms and provides a vast library of pre-built nodes for various data manipulation and analysis tasks (Gollapudi 2020). The software also offers a flexible platform for creating custom workflows and integrating with other tools, making it a popular choice for data scientists, business analysts and data engineers. KNIME has achieved widespread success in a variety of analytics fields thanks to the modular workflow architecture it employs, as well as its inherent capacity to automatically align numerous tasks, free distribution and straightforward analytical pipeline communication (Warr 2012). Additionally, it is highly adaptable and enables the integration of many applications and tools (Berthold et al. 2009). See Tiwari and Sekhar's review for a thorough description of the "workflow" idea, as well as other software that uses this method (Tiwari 2007Fig. 3

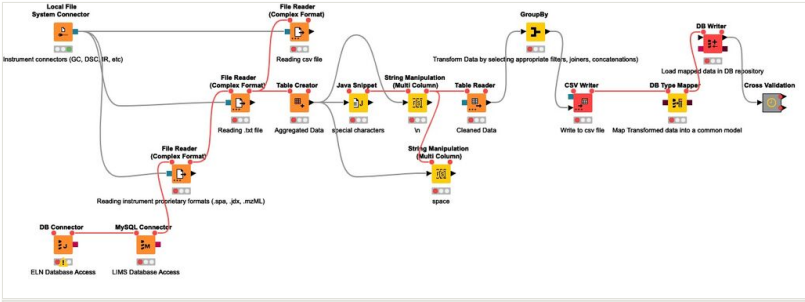

Figure 3. doi

An example of KNIME workflow to extract, clean and validate data from different sources.

From the above workflow, the key steps to consider are listed below.

1. Data Collection: Use the "File Reader" node to read data from multiple files or instruments. This can include CSV, Excel or other file formats.
2. Data Cleaning: Use the "Data Cleansing" node to remove duplicates, inconsistencies and missing values from the data.
3. Data Transformation: Use the "Data Transformation" node to convert the data into a common format that can be integrated with other data sources. This can include converting column data types, aggregating data and more.
4. Data Mapping: Use the "Data Mapping" node to map the transformed data to a common data model, ensuring compatibility.
5. Data Loading: Use the "Database Writer" node to load the mapped data into a data repository, such as a database or data warehouse.
6. Data Quality Check: Use the "Data Validation" node to verify the accuracy and completeness of the integrated data. This can include checks for missing values, data ranges and more.
7. Data Visualisation: Use the "Data Visualisation" node to visualise the integrated data and explore patterns and trends. This can include creating charts, graphs and other visualisations.
8. Data Export: Use the "File Writer" node to export the integrated data to a file format for further analysis or sharing.

These are just examples of the various nodes that could be used in a KNIME workflow for data integration. The actual workflow will depend on the specific requirements and data sources. There are over 13,000 workflows already developed by KNIME contributors on the community platform and this could serve as a point of reference for users. Further, there are thousands of nodes with descriptors from which a user can build a workflow. KNIME's use of scalable machine-learning is an intriguing feature. Some of these algorithms use naive Bayesian models or similarity searches to do virtual screening, with most options being predefined. Nonetheless, integrating scripts from programming languages with machine-learning libraries (such as R and Python) is one approach for increasing flexibility in KNIME operations.

We have designed a KNIME workflow for data integration that could serve as an example of multi-conversion nodes and data aggregator in Fig. 3.

What next after integrated data?

To develop procedures that examine various facets of chemical space, a wide variety of chemoinformatic resources are accessible. These resources are being used in custom workflows or open web servers. These technologies serve not just cheminformaticians, but also members of interdisciplinary teams inside businesses who are either non-experts or do not have the time to create their own code or procedures from scratch. Below, we provide a non-exhaustive list of tools/vendors on the market today that can help organisations in the inception of their digital awareness and transformation journey (Fig. 4).

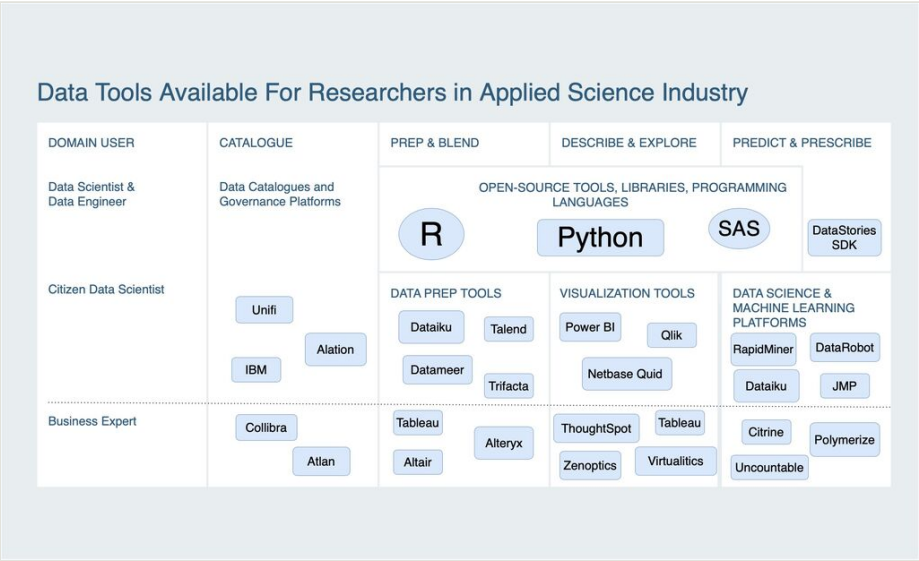

Figure 4. doi

Competitive landscape of data tools.

We classify three (3) key domain users in any organisation, based on their digital expertise level and experience. The core domain user-group will be the full-time data scientists or cheminformaticians hired and fully dedicated to programming and data governance structuring. This group will have the capacity to utilise a deep programming language like R and Python to extract, prepare, explore and build predictive models on datasets. The next related group is the citizen-data scientist, who happens to be fairly knowledgeable in data science tools as well as possessing domain knowledge of the business needs. This group can utilise the low-code/no-code platforms to build insight and push data-driven decision-making across the organisation. Merkelbach et al. provide a nicely documented approach to enabling internal organisational domain experts to become citizen-data scientists (Merkelbach et al. 2022). The final group is the business experts with roles solely based on the deep expertise in the R&D and business-related needs of the company. This group can also utilise the no-code platforms to harness business intelligence from integrated datasets.

These tools are expected to evolve and improve in the future. It is important to avoid having the user-friendly web server apps turn into unusable black boxes. To completely optimise the interpretation of the findings, it is critical that the user fully understands the computations that are performed. The user also has to be aware of the approximation and potential constraints of the application or workflow. Moreover, organisations should not shy away from approaching technical experts in the field of this and many other data tools available to them in their unique situations. The majority of these vendors offer small sand-box exercises to generate excitement and value for a use case that will be beneficial to both parties. Therefore, if an organisation is not well-versed with citizen-data scientists or data scientists in the field of AI and other ML programming languages, the key

recommendation is to engage with select domain specific vendors in a sandbox proof-of-concept to create a successful use-case story. However, of course, success is dependent on the data availability; hence, the data integration step is always going to be the first step for a truly digitalised organisation.

Conclusion

Positive change required in data management is frequently hampered by silos, whether they be operational or informational. This is especially true for data from the material and chemical science industry. Integrated data sources are critical in the research lab because they provide researchers with a comprehensive and centralised view of their research. This aids in decreasing data duplication and discrepancies, facilitating data analysis and enabling effective data administration. Furthermore, by giving access to up-to-date and correct information, connected data sources promote team communication and enable better decision-making. Thus, the usage of linked data sources can lead to enhanced research outputs, higher productivity and overall lab efficiency.

As organisations embark on a digital transformation journey, having an integrated data lake from research labs is very critical to the success of application of ML algorithms for new formulation and material improvement predictions. In order to train and validate ML models, integrated lab data sources are necessary. To make reliable predictions, ML models require a vast amount of high-quality, diversified and consistent data. The data utilised for training and validation may be more thorough, accurate and up-to-date by combining data sources, according to experts. Furthermore, an integrated data source makes it simple for researchers to contribute fresh data to the model, allowing it to continuously improve its predictions over time. Consequently, integrated data sources are essential for the success of ML in the lab since they lay the groundwork for developing new and better predictions.

Breaking down large, monolithic lab programmes that have grown into sources of technical debt and transformational roadblocks is a key step in digital transformation. An organisation may begin to recognise and appreciate the advantages of digital transformation by removing obstacles to information exchange amongst the various lab types and facilitating better flow and access to data. All laboratories, despite the fact that they may not be constructed equally, should be viewed as equally significant components of a system that can offer long-term operational and business benefits through quicker, more integrated data and processes.

Appendix

Summary of data integration steps for the lab environment:

1. Standardise data formats: Ensure that data from different instruments are saved in a standard format, such as CSV, JSON or XML.

2. Use a common database: Store data from all instruments in a common database, such as MySQL, MongoDB or PostgreSQL.
3. Write scripts: Write scripts in a programming language such as Python, R or MATLAB to extract, clean and combine data from different sources.
4. Use APIs: If the instruments have APIs, use them to retrieve data and integrate it into your system.
5. Implement data integration software: Use software tools like Talend, Informatica or MuleSoft to automate data integration from different instruments.
6. Regularly update and maintain the integration: Regularly check and update the data integration to ensure that the data are accurate and up-to-date.

Abbreviations

AI – Artificial Intelligence

API - Application Programming Interface

R&D – Research and Development

ML – Machine Learning

NIST – National Institute of Standards and Technology

CSV – Comma Separated Value

mzML – XML-based format for Mass Spectroscopy output files

XML – eXtensible Markup Language

JSON – JavaScript Object Notation

KNIME – Konstanz Information Miner

API – Application Programming Interface

MGI – Materials Genome Initiative

IP – Intellectual Property

SQL – Structured Query Language

LIMS – Laboratory Information Management System

ELN – Electronic Lab Notebook

PLC – Programmable Logic Controller

DCS – Distributed Control System

SCADA – Supervisory Control And Data Acquisition

IOT – Internet of Things

TXT – Text

GC – Gas Chromatography

DSC – Differential Scanning Calorimetry

TEM – Transmission Electron Microscopy

IR – InfraRed Spectroscopy

GCMS – Gas Chromatography Mass Spectroscopy

JCAMP-DX – Joint Committee on Atomic and Molecular Physical Data

m/z – Mass to Charge ratio

RDP – Remote Desktop Protocol

IT – Information Technology

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Adamides E, et al. (2020) Information technology for supporting the development and maintenance of open innovation capabilities. Journal of Innovation & Knowledge 5 (1): 29-38. [In English]. https://doi.org/10.1016/j.jik.2018.07.001
- Berthold MR, Cebron N, Dill F, Gabriel TR, K&uml o, T. M, T. O, P. T, Wiswedel B (2009) KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter 11 (1): 26-31. https://doi.org/10.1145/1656274.1656280
- Brunner L, et al. (2018) An introduction to KNIME-workflows for data science. The European Physical Journal Plus 133 (11): 589.
- de Pablo JJ, Jackson NE, Webb MA, Chen LQ, Moore JE, Morgan D, Zhao JC (2019) New frontiers for the materials genome initiative. npj Computational Materials 5 (1): 41. https://doi.org/10.1038/s41524-019-0173-4
- Dwivedi S, Kasliwal P, Soni S (2016) Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). 2016 Symposium on Colossal Data Analysis and Networking (CDAN) https://doi.org/10.1109/cdan.2016.7570894
- Ferreira F, et al. (2018) KNIME: A tool for data management and analysis. Journal of Information and Data Management 3 (2): 63-68.

- Gao C, Min X, Fang M, Tao T, Zheng X, Liu Y, Huang Z (2022) Innovative materials science via machine learning. Advanced Functional Materials 32 (1): 2108044. https://doi.org/10.1002/adfm.202108044
- Gibbon GA (1996) A brief history of LIMS. Laboratory Automation & Information Management 32 (1): 1-5. https://doi.org/10.1016/1381-141X(95)00024-K
- Gollapudi S, et al. (2020) Data integration using KNIME: A tutorial. Computing and Information Technology 3 (2): 4-17.
- Groth A (2020) What Is a Web Server and How Does It Work? https://www.lifewire.com/what-is-a-web-server-3426831
- Hendler J (2014) Data integration for heterogenous datasets. Big data 2 (4): 205-215. https://doi.org/10.1089/big.2014.0068
- Hernandez P (2018) Materials Genome Initiative. NIST URL: www.nist.gov/mgi
- Hirsch DD (2013) The glass house effect: Big Data, the new oil, and the power of analogy. Me. L. Rev 66: 373.
- Kannan PR, et al. (2018) Automation of web servers using Python programming language. International Journal of Network Security & Its Applications 10 (6): 133-143.
- Köster J, et al. (2017) KNIME: The Konstanz Information Miner. In Knowledge Discovery, Knowledge Engineering and Knowledge Management. *Springer*, Cham463-466.
- Kumar P (2019) How Digital Transformation Can Boost ROI for Chemical Producers. https://www.industryweek.com/technology-and-iiot/how-digital-transformation-can-boost-roi-chemical-producers
- Lu H, Chan HC, Wei KK (1993) A Survey on Usage of SQL. ACM SIGMOD Record 22 (4): 60-65. https://doi.org/10.1145/166635.166656
- Machina HK, et al. (2013) Electronic laboratory notebooks progress and challenges in implementation. SLAS Technology 18 (4): 264-268. https://doi.org/10.1177/2211068213484471
- Merkelbach S, Enzberg SV, Kühn A, Dumitrescu R, et al. (2022) Towards a Process Model to Enable Domain Experts to Become Citizen Data Scientists for Industrial Applications. In: IEEE (Ed.) 2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS). https://doi.org/10.1109/ICPS51978.2022.9816871
- Miller RJ (2018) Open data integration. Proceedings of the VLDB Endowment 11 (12): 2130-2139. https://doi.org/10.14778/3229863.3240491
- Morgan D, et al. (2020) Opportunities and challenges for machine learning in materials science. Annual Review of Materials Research 50: 71-103. https://doi.org/10.1146/annurev-matsci-070218-010015
- Nagy C, et al. (2017) A static code smell detector for SQL queries embedded in Java code. In 2017 IEEE 17th International Working Conference on Source Code Analysis and Manipulation (SCAM). *IEEE*147-152.
- Noble BB, et al. (2013) First principles modelling of free-radical polymerisation kinetics. International Reviews in Physical Chemistry 32 (3): 467-513. https://doi.org/10.1080/0144235X.2013.797277
- Olson GB (2000) Designing a new material world. Science 288 (5468): 993-998. https://doi.org/10.1126/science.288.5468.993
- Pantelides CC, et al. (2013) The online use of first-principles models in process operations: Review, current status and future needs. Computers & Chemical Engineering 51: 136-148. https://doi.org/10.1016/j.compchemeng.2012.07.008

- Paszko C, et al. (2000) Considerations in selecting a laboratory information management system (LIMS. American laboratory 32 (18): 38-43.
- Prasad PJ, et al. (2012) Trends in laboratory information management system. Chemometrics and Intelligent Laboratory Systems 118: 187-192. https://doi.org/10.1016/j.chemolab.2012.07.001
- Rückert H, et al. (2009) KNIME: The Konstanz Information Miner. *In SDM* 9: 76-77.
- Smith T, Stiller B, Guszcza J, Davenport T (2019) Analytics and AI-driven enterprises thrive in the Age of With. Deloitte Insights
- Tiwari A, et al. (2007) Workflow based framework for life science informatics. Computational biology and chemistry 31 (5-6): 305-319. https://doi.org/10.1016/j.compbiolchem.2007.08.009
- Vasudevan R, Pilania G, Balachandran PV (2021) Machine learning for materials design and discovery. Journal of Applied Physics 129 (7): 070401. https://doi.org/10.1063/5.0043300
- Warr WA (2012) Scientific workflow systems: Pipeline Pilot and KNIME. Journal of computer-aided molecular design 26 (7): 801-804. https://doi.org/10.1007/s10822-012-9577-7
- White A (2012) The materials genome initiative: One year on. Mrs Bulletin 37 (8): 715-716. https://doi.org/10.1557/mrs.2012.194