Capgemini

UNLOCKING DIGITAL TRANSFORMATION IN LIFE SCIENCES R&D THROUGH DATA MANAGEMENT

How focusing on data fundamentals and managing data as a strategic asset enables next-generation solutions

www.capgemini.com

BECOMING DATA-DRIVEN IS CRITICAL FOR FUTURE-PROOFING LIFE SCIENCES RESEARCH – BUT MANY COMPANIES ARE STRUGGLING TO ACHIEVE IT

Over the last ten years, the life sciences sector has been pursuing innovation through digitization at pace. Many organizations have moved from legacy, paper-based workflows or basic electronic lab notebooks to sophisticated electronic data capture systems and data stores. These new systems enable researchers to create permanent, accessible records that can accelerate scientific discovery through data analytics, data science, and artificial intelligence (AI).

But despite having these often high-profile and energetic digitization and change programs, many companies still find themselves unable to achieve their digitization goals in R&D. Data often remains in costly, siloed, and underused data management systems, and data often lacks both the quality and quantity required to enable the new kinds of insight and value generation that the data science and AI revolution promise.

The life sciences industry is now at a point where data- and AI-driven startups that were 'born digital' are seriously disrupting long-established industry norms – from identifying effective, safe drug candidate compounds *in silico* to rapidly identifying the ideal patient population for a clinical trial. Incumbents who fail to recognize and capitalize on the value of their own decades' worth of research data could soon find themselves left behind.

In this paper we explore some of the challenges that life sciences companies face in their attempts to enable data-driven R&D, and how well-devised strategies and tactics around data management can avoid the common traps that organizations fall into. From there we show how careful data management can release the incredible potential locked up in research data and establish a solid foundation for the delivery of high-quality insights from analytics, data science, and AI projects that will drive the R&D programs of the future.

CASE STUDY

Enabling next-generation automation through the introduction of sound data management principles

Our client, a global pharmaceutical company, wanted to enable their research teams to more rapidly submit regulatory filings used in initial market applications. Historically, the process involved a significant amount of manual work such as copy-pasting, manual formatting, and manual creation of tables and graphs, which required the report writers to have a deep understanding of the data itself, as well as its origin and provenance.

Improving efficiency by automating aspects of this process was an obvious way to reduce costs – but this required data to be captured much more systematically and annotated with appropriate, consistent metadata so that it was well described, in a predictable place, and in a machine-readable format so the automated system could process it accurately into the final report document.

Our solution included a digital transformation roadmap and implementation of a cloud-based platform that enabled automated, accurate regulatory reporting, reducing report authoring time by 90% and realizing annual efficiency savings of approximately 20 FTEs.

DATA MANAGEMENT IN LIFE SCIENCES R&D

What is data management?

Put simply, data management is a collection of practices, processes, and roles that treat data as an asset. In practice, this means understanding the end-to-end lifecycle of all the different data sets relevant to an organization, from their initial creation, capture, and immediate analysis for their primary purposes to their long-term storage, processing, and maintenance for secondary purposes. Each step usually needs to be supported by tools, technology, processes, and people in a way that aligns with the organization's overall data strategy.

In a life sciences R&D environment, this usually means:

- Understanding what data is generated in each type of research experiment, and how it is generated;
- Ensuring that data from those experiments is captured accurately and to the right level of quality and standardization;
- Making sure that data is stored in a sensible place, in an appropriate format, with sufficient metadata and contextual information to enable it to be found and understood by someone who was not part of the initial experimental team;
- Monitoring and maintaining data quality over time to ensure it remains as consistent as possible while keeping up to date with the latest standards and needs;
- Defining who should have access to the data (ideally keeping access as wide as possible within the organization) and implementing well-defined mechanisms to enforce that;
- Providing documented, reliable means of data access so that data can be used and reused multiple times as scientists and innovators come up with new ideas; and
- Ideally, enabling the linking and re-linking of data sets and transformation between formats to be as automated as possible, minimizing or eliminating the often very costly time spent 'data wrangling'.

If done well, these activities form a virtuous circle over time, as the needs of data consumers drive improvements in the practices and efficiency of data producers.

Why is it difficult to do data management well?

The benefits of good data management are wide-ranging, and as more organizations recognize that data is essential to their evolution and survival, more are accepting the need to do it and do it well. But paradoxically, it can still be difficult to make a case for funding data management efforts, because:

- The return on investment is indirect – that is, data management enables value to be generated from data through reuse but doesn't, in itself, provide that immediate payback. This can lead to it being seen as a pure cost, rather than an activity that removes the unmeasured costs of data that is not optimized for reuse. In reality, the ROI of data management is often considerable, but it's essential to take a long view to be able to show this.
- Data management can be perceived as a slow and costly process – this can be true, especially if initiatives are started with too broad a remit. However, it's perfectly possible to start the process with small, targeted interventions and build out from there.
- Data management can be seen as disruptive – in R&D, workflows are often highly optimized for throughput and, consequently, fiercely defended by data producers whose time is limited and who may not see themselves as the primary recipients of the benefits of data management. A sensitive and nuanced approach is essential, and a good tactic is finding ways to pay data producers back for the time they invest in more demanding data capture processes – for example, by using data captured at a higher level of quality to automate other elements of their workflow.

What unique data challenges does the life sciences sector have?

All R&D-driven industries have their own unique data issues, but in life sciences they can be particularly challenging. Here are just a few:

- **Volume –** most commercial life sciences organisations are trying to identify compounds and substances that have a particular effect on a certain type of organism, whether that is a drug designed to alleviate disease in patients, a dietary supplement for children, or a pesticide that kills weeds while minimising environmental harm. In a typical research program, literally millions of compounds are initially tested to find the ones with promising properties, resulting in vast data sets. Adding high-volume genetic and -omics data into the mix compounds this further.
- **Complexity –** the most effective products balance efficacy, safety, cost and convenience – a pill with minimal side effects, taken once daily by patients at home, is far better than a drug that addresses a disease but causes other health issues and must be injected regularly by a health professional, for example. Balancing and optimising these factors requires many different dimensions of data to be investigated and analysed together.
- **Heterogeneity –** while much life sciences data takes the form of simple numeric properties, less structured forms of data such as medical records, images and video have also become valuable sources of information. On top of this, data systems in use across multiple labs in the same organisation may have been optimised for specialist workflows over many years, making lab processes efficient but creating an extremely heterogeneous environment of proprietary formats. Pulling together and harmonizing a data landscape like this can require a lot of hard work understanding what data is out there and developing standards and ontologies that cover the whole space.
- **Collaboration –** there are now many well-established, high-quality public life sciences data sources such as the European Bioinformatics Institute (EBI) and clinicaltrials.gov. These data sources are a treasure trove and understandably companies want to make use of them to accelerate their own research. But how can and should they be integrated with in-house data? And more worryingly, what do scientists' use of these public data stores – which may be captured in query logs – reveal about the confidential research programs of the organization they work for?

All of this often means that when busy research scientists are under pressure to deliver results quickly, they fall back on infinitely flexible data formats such as spreadsheets and documents. But this narrow view damages the overall organization's ability to make use of data as a long-term asset.

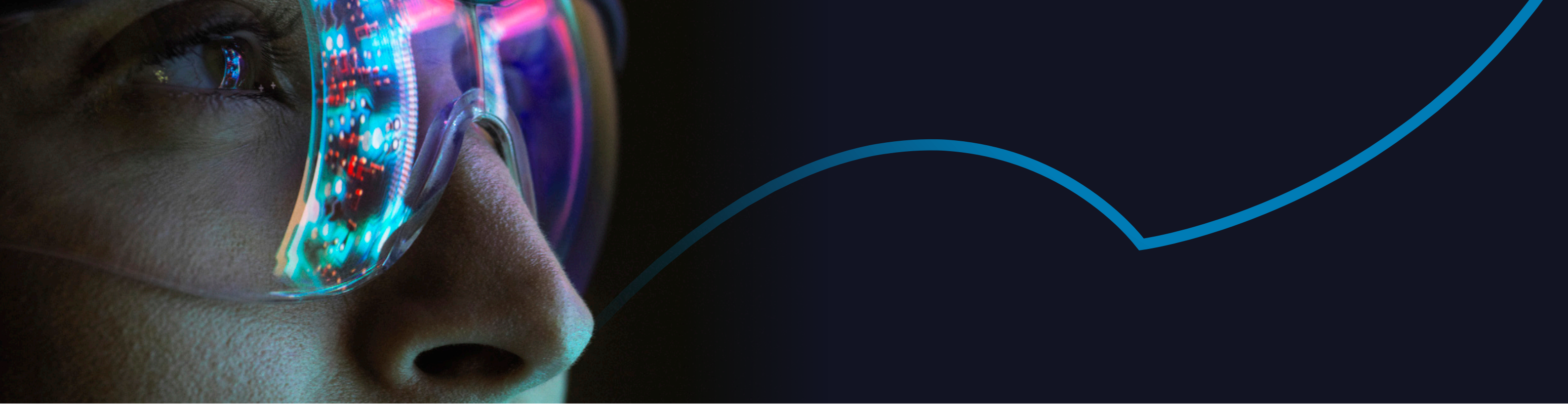
CASE STUDY

When the real problem is data management

We worked with a client in the life sciences sector who had ambitious goals for embedding modeling and simulation throughout their research pipeline. Initially, the client aimed to build or buy tools to directly support the modeling effort, but our investigation revealed early in the project that the goals were not achievable with the current state of the input data – particularly as a result of lack of consistent storage, quality, and standardization.

We advised the client to adjust the scope of the project to focus on these data management fundamentals first. Following this, we successfully defined and introduced new data standards and new data capture, storage, and management processes that balanced the needs of both data producers and data consumers. The data platform that we co-developed with the client built on these principles saved the organization thousands of hours of wasted time (and millions of dollars) annually.

Perhaps more importantly, the new platform makes previously ambiguous and hard-to-find data much more readily available and easier to understand by modelers, achieving the original project objective and unlocking the organization's ability to transform the way research is done.



THE FAIR DATA PRINCIPLES: A READY-MADE, GRASSROOTS DATA MANAGEMENT FRAMEWORK FOR SCIENTIFIC R&D

What is FAIR data?

There are many ways that good data management practices can be implemented. The FAIR principles[1] set out a framework for managing and representing scientific data sets to maximize the potential for data reuse.

FAIR data is simply data that satisfies the four FAIR principles:

- *Findable* – data (and metadata) are easy to find
- *Accessible* – once found, data are easy to access via clearly documented methods
- *Interoperable* – data can be easily transformed, linked, and integrated with other data and applications
- *Reusable* – data is optimized for intelligent, informed reuse (for example by being as clearly and unambiguously described as possible to avoid misinterpretation)

FAIR provides a popular and powerful foundation for data management in an R&D context because it originates from the scientific community itself in recognition of the many problems that poorly managed data creates for scientists. It isn't an externally imposed IT or technologist-led approach, which usually makes it more acceptable for scientists to adopt – and indeed in the life sciences industry FAIR has made a lot of headway in the last few years.

However, while easy to describe and understand at a high level, the FAIR principles are formally defined in some detail and can be challenging to implement in full in the real world.

[1] 'The FAIR Guiding Principles for scientific data management and stewardship', Wilkinson et al, Sci Data 3, 160018 (2016)

> *"FAIR provides best practice governing principles that guide your data management"*
>
> **NATALIE STANFORD**
> Senior Data and AI Strategy Consultant, Capgemini Engineering

OVERCOMING THE CHALLENGES OF IMPLEMENTING DATA MANAGEMENT IN R&D WITH THE FAIR PRINCIPLES

Findability

Challenge

A great deal of data in companies is still siloed on users' computers, domain-specific infrastructure, and in lab notebooks.

Action

Ensure interested stakeholders can search for data, enabling an efficient route to insight. This is enabled through:

- Metadata and ontology development tailored to your business so that data can be searched according to your use cases.

- Deciding whether to centralize data physically (e.g., in a data lake) or virtually (e.g., in a data catalogue that connects centralized and decentralized resources).

Accessibility

Challenge

Data is locked away in inaccessible silos, leading to wasted time in collating data for use by data scientists.

Action

Design and implement governance structures, transparent data access models, and clear, documented data access methods that ensure the right data is available to the right person at the right time. Ensure security, privacy, licensing, and regulation are all addressed. Finally, distinguish between access to data and access to metadata, so that when data access does need to be kept tight, users can at least see that the data exists and understand how to request access.

Interoperability

Challenge

Even when data can be found and accessed, it is rarely available in such a way that it can be combined with other data and reused in different models or analysis environments.

Action

Assess what levels of interoperability are currently present, and what levels need to be implemented to satisfy the business need. Levels of interoperability include:

- *Foundational*: Can it move between different systems?

- *Structural*: Can it be absorbed by systems with some context data?

- *Semantic*: Can it be exchanged with a full understanding of its origin, context, and meaning?

Reusability

Challenge

When data is available and usable it can be difficult for users to know the provenance of the data and understand it enough to know if it is valid for their use case.

Action

Define and implement the right culture and technology to ensure quality recording of data with adequate metadata and provenance, ensuring it can be reused, starting with the highest priority use cases, then cascading out to the rest of the business.

BUILDING A FAIR-INSPIRED, DATA-DRIVEN CULTURE

Organizational culture has a huge impact on how successful digital transformation projects will be. Digital transformation inevitably involves harmonizing data capture, storage, use, and reuse to enable the transformation vision. This requires strong relationships between data producers, data consumers (often in different departments), and IT departments (who often serve as data brokers between producers and consumers). An understanding and appreciation of why spending time and money to ensure data is FAIR is also important, from senior management down to the most junior scientist.

> *“Building the right culture is pivotal to implementing FAIR data and enabling data- driven R&D”*
>
> **JEROEN DE JONG**
> Senior Data and AI Strategy Consultant, Capgemini Engineering

THE FOUR KEY RELATIONSHIPS SUPPORTING A MATURE DATA CULTURE

Relationships with IT

Ensure IT and R&D have a trusting and collaborative relationship and that technical infrastructure is designed to support the scientists.

Relationships across departments

Ensure different scientific domains collaborate well and can surface visible data management successes driven by new kinds of data interchange.

Relationships with senior management

Ensure buy-in and consistency of vision from the C-suite to the bench scientist so that priorities are aligned, and data management is adequately funded.

Relationships with data itself

Demonstrate and develop enthusiasm in scientists for the benefits that data management brings to them personally, creating motivated 'data evangelists'.

CASE STUDY

Introducing data management pragmatically into an overwhelmed R&D landscape

A global pharmaceutical company was struggling with the challenge of growing in a manageable way while also harnessing the proceeds of well-established, successful drugs to facilitate future innovation. While strong and successful in their specialist disease areas, they had realized they were weaker in terms of data maturity. Data management was not a topic that featured in their plans and roadmaps for the company as a whole, which meant that individual areas within the company were either neglecting data management or had developed their own custom approaches, leading to a patchwork landscape of inconsistent and incompatible practices.

Capgemini was challenged to help set up a data management practice across the whole R&D space in a ‘lean and mean’ way - this was a fast-moving, innovative environment where results were needed quickly. Our approach was to quickly understand the current R&D data landscape and the company culture, then define the required end state. Within six weeks we proposed a pragmatic way forward that we were then asked to lead.

The company’s culture and ambitions placed a high value on quality, so our change management approach emphasized the need for high quality data. We identified local champions and started coaching them in the new data management principles. As this took flight, other departments became interested in the easy ‘playbooks’ we were creating on data management topics such as data cleansing and annotating data with metadata. Being very careful to only make ‘no regret decisions’ that were well aligned with industry-standard best practices (such as those of DAMA, the Data Management Association), our approach could easily be incorporated in other initiatives that were starting up more slowly. In about a year, data management became a top-of-mind topic of interest for many data practitioners and lab heads alike.

BRINGING DATA TOGETHER TO ENABLE YOUR DIGITAL VISION

Many of our clients come to us after trying to enable data science and AI use cases – discovering too late that their data needs considerable work before their digital goals can be achieved.

Revamping your data management processes, technologies, and systems can be a daunting task, and it is not always clear whether you will derive enough value from the enablement process.

Capgemini can support you at a number of points on your journey to becoming data-driven:

- Establishing a data strategy based on surveying your business goals, current practices, current technology, and favoured use cases;
- Supporting your business case for improved data management through Proof of Value and Proof of Concept work to demonstrate what can be achieved on a small scale before you make a big commitment; and
- Embedding data, analytics, and architecture experts into your organization to enable your vision.

> *"Fair is the foundation of AI-readiness and a critical underpinning of a data-driven organization"*
>
> **JAMES HINCHLIFFE**
> Senior Life Sciences consultant,
> Capgemini Engineering

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 325,000 team members in more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2021 global revenues of €18 billion.

Get the Future You Want | www.capgemini.com

For more details contact:

lifesciences@capgemini.com