

New Discovery Data Architecture Needed to Support New Modalities

CONTRIBUTED COMMENTARY BY DAVID LOWIS | OCTOBER 22, 2021

Increasingly, the discovery and development of new medicines is embracing bioactive modalities beyond traditional small molecules, both as therapeutic agents and delivery mechanisms.

Unlocking the power of patent data is a smart, cost effective way to help mitigate this risk. Although The development of medicines from any modality—small molecules, peptides, oligonucleotides, antibodies, and antibody-drug conjugates—is inherently a data-driven activity requiring the effective representation of substances, the efficient capture of data from the research process, and the comprehensible presentation of those research data for key decision making. But the maturity of informatics systems for capture and analysis of data for new modalities, especially in antibody research where processes are different, is still catching up to that for small molecules.

Considering the rate at which additional new modality types are being developed, this scientific informatics situation will likely deteriorate in the short term with a proliferation of traditionally architected, specialized tools offering an array of modality-specific capabilities but having to duplicate base-level functionality for data access and manipulation.

An alternative architecture for pharmaceutical research is needed that allows new modality-specific scientific capabilities to be combined with existing scientific informatics platforms to handle common data needs. This architecture must embrace and support the development of standards for substance and experimental data representation and operate across a multi-vendor hosted environment.

Evolving Needs

Legacy products for small molecule development tended to be deployed on-premises, allowing organizations to bring substance and experimental data together readily and enabling scientists to build the data views required for their decision making. While some organizations are still trying to address their data access and analytics needs, this portion of the problem is soluble.

But increasingly, data are moving to the cloud and a biopharmaceutical company may have several data systems hosted by different vendors. While scientists can access the data from these systems via the vendor provided tools, to get a holistic and consistent view of

their research, scientists must pull data from multiple environments. They typically do that via web services because they can rarely access the databases directly. However, those web services are seldom of high performance or capable of filtering down to the data that the scientist or process wants. This situation precludes the direct, interactive use of data from multiple sources because only the hosted system vendor's data access capabilities can be employed on their data source. With a multiplicity of new tools for different modalities becoming available, the necessary software footprint required by a scientist will likely increase in cost and complexity and require duplication of scientific business rules.

An alternative architecture for pharmaceutical research is needed that allows new modality-specific scientific capabilities to be combined with existing scientific informatics platforms to handle common data needs.

Technology improvements are required so that all the data can be brought together in close to real-time to make research decisions. While that was difficult to achieve for small molecule drug development, the situation is exacerbated by the proliferation of new modalities and their associated informatics systems. Instead of just a small molecule registration system, sponsors may now also need an antibody, an oligonucleotide, and a peptide registration system, each of which will have its own data access capabilities. For peptide therapeutics, for example, sponsors need to register sequences and monomer libraries rather than atomistic structures. Even those modalities are evolving, from single string to branched to cyclic peptides, getting increasingly complex.

While there are now commercial scientific informatics platforms, such as Certara's D360, Biovia's Insight, Perkin Elmer's Signals, and Dotmatics' Browser that can bring

together data from multiple sources, it is still difficult for them to operate in an environment where data sources are hosted across multiple vendors.

Even when vendors do provide good data access, the assay results for different modalities are often in different systems. These assay results should be treated similarly because they are just testing a different substance. But vendors aggregate and present their assay data in many ways, leading to problematic inconsistencies that put a renewed data burden on the scientist.

Conversely, a sponsor could attempt to alleviate this situation by locking into a single vendor. While this may sound attractive initially, it guarantees that optimal solutions will not be created, significantly limits the sponsor's future agility as new technology and modalities become of interest, and introduces risk based on the vendor's financial situation.

Developing Data & Service Standards

In response, I propose that the industry create standards that describe not only how experimental and substance data should be treated but also cover the services that provide access to that data.

Ideally, these standards will be developed by an industry consortium operating in a pre-competitive environment. I envision that they will build upon the [FAIR Guiding Principles](#) for scientific data management and stewardship. Those principles focus on metadata and are designed to improve the findability, accessibility, interoperability, and reuse of digital assets. The new standards will produce vendor-agnostic FAIR data services.

FAIR data standards and capability and performance standards for data services will produce a truly effective multi-vendor environment, which allows new modality capabilities to be connected into a system without duplication of capability.

Establishing new standards will create a scientific data framework that will enable seamless data transfer between different vendors' systems, reduce system redundancies, enable more effective support of current research, and provide extensibility for the future.

These data standards will address questions such as: If an assay is run, what parameters does it need to meet, what do they mean, and how should that information be presented? Are these two assays the same? Do they measure the same scientific outcome? Is there enough metadata to support that? Are they comparable enough that public data can be blended with internal proprietary data? Standards that allow that level of analysis are invaluable.

This approach will free organizations to build their optimal informatics environment. It will allow them to choose the tools that are best suited for their research, rather than having to lock in to using all the tools from one vendor. In addition, sponsors will gain near real-time data access because they will no longer have to wait for extract, transform, load (ETL) processes to move data around. There will also be greater consistency in how data are presented. Sponsors will no longer have to deal with different business rules governing how data are aggregated, different definitions of assays that require different treatment, and the resulting inconsistencies in data presentation.

While this may seem like a major undertaking, industry members have achieved comparable goals by working together. Successful programs include the [Pistoia Alliance's](#) standards and [Allotrope Foundation's](#) findable and accessible data. We need to achieve this goal because consistency is vital for rational, scientific decision making.

David Lewis, DPhil, is Executive Director of Scientific Informatics at Certara. Dr. Lewis leads Certara's scientific informatics group and is responsible for the D360 scientific informatics platform. For the past 15 years, he has led the design and development of D360 data access, analysis, and collaboration software, expanding from small molecule discovery into biologics and pre-clinical research domains. Dr. Lewis gained his Doctor of Philosophy degree from Oxford University, and he holds a first-class honors degree in chemistry. He can be reached at david.lewis@certara.com.



Overlook Center, Suite 101,
Princeton, NJ 08540
www.certara.com
sales@certara.com