

Realizing the potential of FAIR data for pharmaceutical R&D



Creating and using FAIR data has the potential to transform the analyses enabling drug discovery and development. Numerous resources and services can jumpstart and streamline the implementation of the FAIR Data Principles. However, adoption in the pharmaceutical industry has been slow, partially due to challenges in establishing the infrastructure. It is also critical to recognize internal hurdles in corporate organization and culture.

More realistic, multi-dimensional analyses can help understand disease and assess how chemical entities behave in biological systems. This approach promises to slash drug development times and vastly reduce late-stage failures.

Data at the center of a transformation

An unprecedented development is changing scientific inquiry worldwide. Data — the link between the subjects we study and the analyses that tell us how they work — have moved into a new spotlight for basic, applied and product-oriented research and development. Until recently, data were generated and used with a specific analysis in mind. Now, emerging technologies have unlocked the data-rich workflows needed for scientists to take large sets of existing data and apply them to new questions. The valorization of data has shifted from *owning and using data once* for a limited purpose, to *sharing and reusing data* for any number of potentially disparate analyses.

The pharmaceutical industry is no exception to this transformation. The power that computational frontiers like artificial intelligence and machine learning have unleashed in other areas brings new optimism to an industry that has suffered an unignorable decline in innovation efficiency.

The one-dimensional analyses used to date in pharmaceutical research and development have been inadequate to address the complex questions pursued. They dissect a problem to the point where critical factors that curb the efficacy of a drug candidate can go undetected.

Drug developers are now looking to bring together everything that is known about a problem to build a more accurate and nuanced picture of patients, diseases and compounds. More realistic, multi-dimensional analyses can help understand disease and assess how chemical entities behave in biological systems. This approach promises to slash drug development times and vastly reduce late-stage failures.

From disconnected datasets to relevant and complementary data

Experimental results must be analyzed according to more comprehensive models of the factors and interactions that determine drug action in patients, and the corresponding multi-dimensional analyses require large amounts of data that are housed in internal databases, commercial data sources and public repositories. Integrating datasets from disparate sources is a process that merges human understanding of how the data are connected with the computational power of machines.





Data used to answer a question must be relevant to the analysis used. That relevance is defined by the meaning of and relationships between concepts and data in a scientific domain — semantics. To use data from disparate sources for multi-dimensional analyses, they must be integrated according to these meanings and relationships.

Thus, data and their associated metadata must be curated to make these attributes unequivocally identifiable for humans and machines.

- Metadata must adequately describe how humans and machines can access and use data, including information about access protocols, licensing and authorization.
- Data and descriptions in metadata should use broadly accepted representations of knowledge to be understood by as many potential users as possible and used in established and novel applications.
- Metadata must describe data richly and transparently with accurate and relevant attributes regarding their source, origin, collection methods and conceptual context.
- Both data and metadata must be uniquely, permanently and explicitly identifiable by humans and machines.

Guidelines for success in data management

The FAIR Data Principles are guidelines for scientific data management and stewardship to make research data findable, accessible, interoperable and reusable and thus, promote their maximum use (1).

 Findable Data are richly described by metadata and have a unique and persistent identifier	 Accessible Data and corresponding metadata are understandable to humans and machines, and accessible through defined protocols.
 Interoperable Data and corresponding metadata use formal and accessible knowledge representation to guarantee reuse.	 Reusable Metadata accurately describe the provenance and usage license for the data.

1. Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 3: 160018. doi:10.1038/sdata.2016.18



What's slowing down implementation of FAIR Principles in pharmaceutical R&D?

The data stewardship objectives above are the main outcomes of FAIRification — the process by which data are prepared according to the FAIR Principles. While the ideas behind these principles have been around for some time, their implementation in the pharmaceutical industry has been slow for two overarching reasons.

First, implementation is neither finite nor predetermined. FAIRification is a long-term overhaul of how data are created and used within an organization, and this process is continuously influenced by an ever-changing knowledge landscape. A critical part of FAIRification is the mapping of data according to a

semantic model — or ontology — that describes the meaning and relevance of each data point, and its relationships to other data in the context of a knowledge domain. Gaps in our understanding of biology make this mapping a moving target. New information fills in what we don't know and changes what we thought we did know.

Second, implementation is likely to require significant culture change within an organization. The shift from owning and using data once to sharing and reusing data is counterintuitive for many managers and scientists in the pharmaceutical industry who have been conditioned to work in information silos.



Investing in the right people and the right approach

The infrastructure to semantically integrate and then apply data in multi-dimensional analyses demands a real upfront investment. Hardware and software for onsite data storage and processing are no trivial matter, but a more complicated task is building the underlying conceptual framework. This framework defines data attributes used for FAIRification and dictates how fragmented, unstructured and domain-specific data from internal and external sources are brought together. Part of that framework is one or more ontologies that flexibly adapt to changes in knowledge and accommodate a broad range of existing and yet-to-come data types.

Developing an ontology completely from scratch is not necessary. Examples are available for implementation and can be modified to meet the requirements of in-house data. Nevertheless, determining the best ontology for a pharmaceutical company and the data it generates requires expertise in the relevant science domains and in knowledge representation. It is rare to have one person proficient in both areas and, depending on the bandwidth of the data streams, equipping an enterprise with that know-how can mean investing in people for the right interdisciplinary task force.

Expertise and training are also needed at the practice level. Once generated, data must be furnished with the metadata that make them findable, accessible, interoperable and reusable. These curation activities consume resources and time, and benefit from knowledgeable oversight, i.e., a dedicated team for data stewardship and quality control. Of course, aspects of this curation can be automated but not without carefully planned algorithms from researchers, data specialists and knowledge engineers to ensure that outcomes are scientifically accurate.

Less visible but equally important are the updates needed in corporate governance. SOPs must be rewritten to ensure data are collected, curated and processed according to the established conceptual framework. Furthermore, changes to data management SOPs will necessarily intersect with SOPs in other areas. For example, regulatory processes covering the protection of patient data must guarantee data reuse while respecting patient privacy. A systematic review of organization-wide procedures to identify points of intersection streamlines FAIRification and makes the subsequent use of FAIR data transparent across the organization. However, even with detailed knowledge about every aspect of a company, teasing apart complex processes is labor-intensive

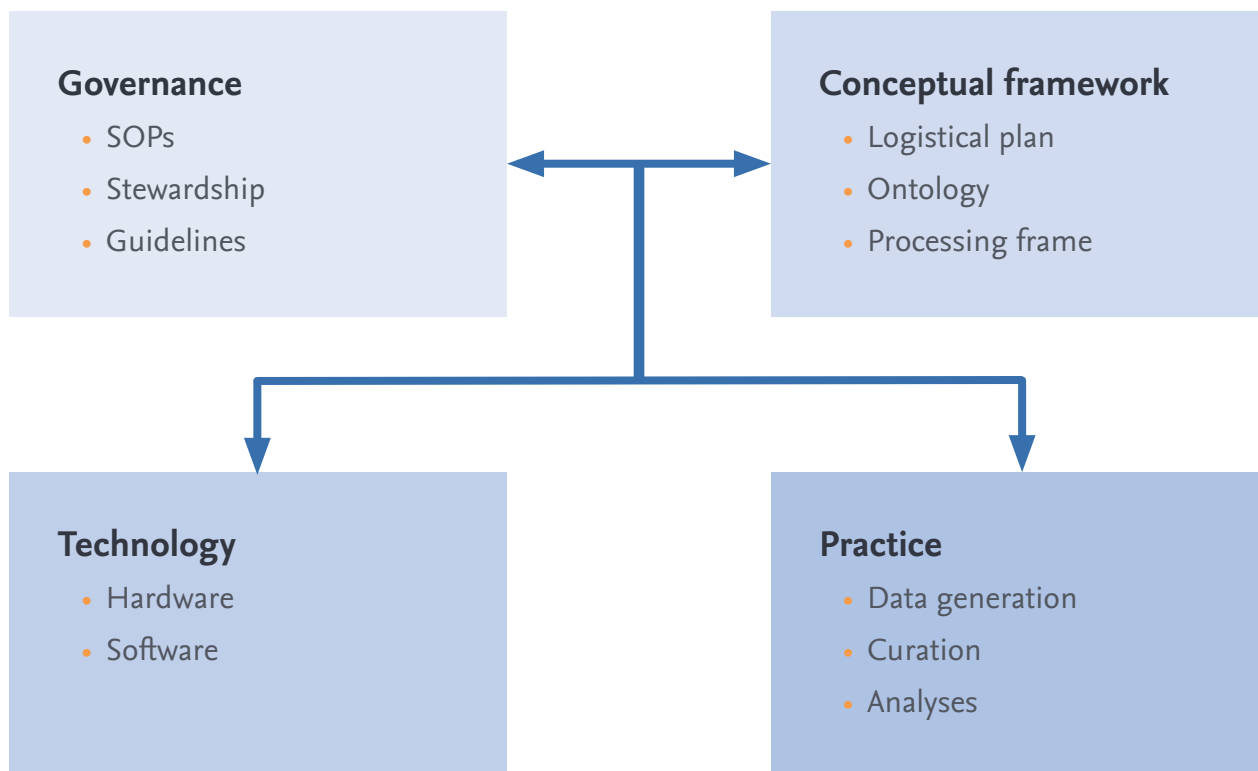


Figure 1. Upfront investment in technology, personnel and processes to implement FAIR Data Principles is needed at multiple levels. Beyond technology, the conceptual and procedural infrastructure underlying meaningful adoption of FAIR Principles requires interdisciplinary expertise and has an organization-wide impact.

A commitment to change how things are done

The adaptations and new resources needed at the conceptual, governance, technology and practice levels to implement the FAIR Principles are challenging but tractable for a motivated management. The changes required in corporate culture are more difficult to steer and may take longer to realize.

For managers, the notion of sharing and reusing data goes against decades of business acumen, which emphasizes confidentiality and isolation to protect intellectual property and secure market share.

Researchers are no different. For centuries, increasingly specialized scientific endeavors have culminated in a publication or product that differentiates the underlying research as first and unique. Scientists have worked in silos, and the data generated were used for one purpose. Now, scientists are asked to view their data as a valuable corporate asset for others to use beyond their original intent.

A shift in mentality is needed, and it includes very tangible changes to daily work routines and performance expectations. Incentivizing all parties to do their parts in generating high-quality FAIR data will require valuing efforts to that end as much as the marketable output of a drug development pipeline. Including the creation and use of FAIR data in the goals and metrics of employee performance evaluations can also help signal priorities from top management. Ultimately, development strategies, experiment plans and day-to-day work that take FAIR data into account should be an intrinsic aspect of any project.

Paving the path for implementation

FAIRification does not happen once, immediately and comprehensively. Instead, FAIRification is an evolving and progressive process. This is especially true for the pharmaceutical industry, where data production is continuous and new knowledge constantly reshapes the information landscape for research questions. However complex FAIRification may seem, it is critical to start the process now and allow for an agile, “test-and-learn” adoption.

Luckily, a large and growing network of organizations offers assistance, expertise and tools to help FAIRify data. Collaborations foster harmonized FAIR practices. International consortia define plans and create specific tools and materials to support implementation. Service providers design and customize solutions that facilitate FAIRification. These organizations exchange ideas, share lessons learned and develop best practices to help reach consensus on community-wide issues.

Many of the institutions catalyzing efforts have lengthy experience in transforming unstructured data into structured and actionable data. Their know-how can reduce hurdles to implementing FAIR Principles in terms of the necessary expertise, technology, conceptual framework and the investment in building each of these.

Current FAIR data endeavors are not the first to attempt merging data from disparate sources into meaningful information. However, this is the first time that the underlying ideas, expertise and technologies align. “We happen to be working at a time when computer infrastructure, knowledge engineering and data generation are finally where they need to be for us to transition towards powerful analytics enhanced by a semantic and more comprehensive representation of knowledge,” says Ted Slater, Senior Director of Product Management for PaaS at Elsevier. Implementating the FAIR Principles to create data environments that produce greater insights from a better understanding of biology, chemistry and their interactions is no longer a “burden of our times.” If anything, it is a guided exploration of new possibilities

Learn more about FAIRification

A system that embodies FAIR Principles at all levels facilitates FAIRification and ensures that the process remains responsive to change. To learn more, see the upcoming article *Bridging Information Silos the FAIR Way*.

Entellect

Entellect is a data integration platform that empowers data-driven R&D in the pharmaceutical industry. It helps companies achieve optimal results by delivering clean, normalized, contextualized and relevant data that are ready for predictive and exploratory analytics.

For more information about Entellect,
visit elsevier.com/entellect.

