propharma

An Introduction to Scientific Data Standards

With automation becoming more mainstream in science, vast quantities of data are now generated by many organisations every day. This deluge of data is often managed ineffectively, limiting the possibility of use in downstream systems or of combining data for future analyses. To address these needs, a number of scientific data standards have been proposed to implement best practice recommendations for the format of the data and associated metadata.

In this article, we explore some of the different data standards available to life science and healthcare R&D organisations and the principles on which they are based.

What Are the FAIR Principles for Data Standardisation?

Two of the leading collaborative groups active in promoting and supporting the use of scientific data standards are the Pistoia Alliance and the Allotrope Foundation.

The Pistoia Alliance recommends the use of four guiding principles (known as the FAIR Principles) in the management and stewardship of scientific data. According to the FAIR Principles, data must be Findable, Accessible, Interoperable and Reusable. This aligns with similar initiatives, such as the ALCOA data integrity guidance issued by the FDA, to ensure that the context and content of the data can be trusted.

These principles are now well understood and established in the scientific community, and are key considerations when implementing any data standard. In addition, many of the new standards use ontologies (set terms approved by the scientific community) to describe data accurately and consistently. The Allotrope Foundation has been a key driving force in the implementation and promotion of such ontologies.

Which Scientific Data Standards Exist?

In order for the scientific community to get the most value out of the available data, it is vital that storage formats are optimal for sharing, archiving and reuse. Adequate description of the data (stored in the form of metadata) is also key for turning data into information.

There are currently three main options when it comes to data format standards used in the life science and pharmaceutical industries; ADF, AnIML and UDM. These data format standards are designed to be generic containers that, in principle, can be used for any type of scientific data.

Data File Format Standards

Standard	ADF (Allotrope)	AnIML	UDM (Pistoia)
Data types currently supported	**Analytical data**	**Analytical chemistry and biological data**	**Experimental information about compound synthesis and testing**
Format	**HDF5 (binary)**	**XML (text)**	**XML (text)**
Established	**2015**	**2003**	**2018**

Abbreviations: ADF, Allotrope Data Format; AnIML, Analytical Information Language; HDF5, Hierarchical Data Format 5; UDM, Markup Unified Data Model; XML, eXtensible Markup Language.

In addition to these data format standards, you may also be considering implementing an automation communication standard, such as SiLA (which is closely related to the AnIML standard, but works with any of the available data format standards).

Furthermore, many healthcare organisations are now adopting process standards already accepted in other industries, such as the S88 (or ISA-88) standard for batch processing.