Ruprecht-Karls-Universität Heidelberg
Medizinische Fakultät Mannheim
Dissertations-Kurzfassung

A standardization concept for machine actionable and reusable scientific data

Autor: Axel Wilbertz
Institut / Klinik: Mannheim Institute for Intelligent Systems in Medicine (MIISM)
Doktorvater: Prof. Dr. J. Stallkamp

The pharmaceutical industry aims to reuse scientific data for machine learning and artificial intelligence. The most critical analytical method in biologics drug development and characterization is high-performance liquid chromatography (HPLC). This thesis underscores the challenge of data integration due to the complexity of biologics and HPLC technology, seeking to bridge this gap by adapting standardization concepts like the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles to make data interpretable for machines as well as humans.

The work focuses on the development of a generic standardization concept that enables machine actionable biologics formulation data. Machine actionability refers to the complete machine-based interpretation of dataset quality, facilitating automatic comparability of datasets. Achieving this requires an adaptation of existing standardization concepts such as FAIR to enable the automation of decision-making and data set comparability evaluation. Furthermore, current semantic models, such as the semantic HPLC model from the Allotrope Foundation for chromatography devices, are only partially suitable for the intended purpose of enabling machine-actionable data. This is because they predominantly focus on the reproducibility of experiments rather than focusing on data reuse in terms of result comparability and autonomous data evaluation. To overcome this limitation, a new standardization concept was developed that consists of three consecutive steps:

1. The creation of a semantic model to convert experimental raw data into a semantic format using web technology concepts such as ontologies, making it machine-readable and giving it meaning for machines and humans. The model focuses on formulation sciences data, particularly stability screening data, and incorporates analytical chromatography results.
2. The definition of aggregated metadata as a set of rule-based acceptance criteria, thresholds, and ranges that allow for the determination of dataset quality. They are based on the subject matter expert (SME), which resembles a scientific expert for HPLC. The aggregated metadata represent the initial step beyond machine readability towards achieving machine actionability. These metadata are combined at a higher level and depend on lower-level metadata from the raw dataset, including formulation composition, analytical method, stress conditions, and experiment processing.
3. The formalization and automation of the SME comparability logic. In assessing the quality of datasets, the logic replicates the manual data assessment process of an HPLC SME. By automatically evaluating aggregated metadata expressed in the semantic model, dataset quality can be assessed without human intervention. The output of this evaluation is a comparability classification, which categorizes datasets into four groups based on their quality and suitability for machine learning. The classification and error calculations are performed on multiple levels.

In conclusion, this thesis presents a novel concept to standardizing biologics formulation data, making it machine-readable and actionable, thus enhancing data comparability and reusability in the pharmaceutical industry. The concept has been developed in a generic way, to allow for the implementation in other domain.