White Paper

Data Governance vs. Data Integrity

Supporting a Digitization Strategy

METTLER TOLEDO

1 Introduction and Definition

With many published articles on data integrity and ALCOA+ [1] for the regulated environment, it would seem that all data issues have been resolved. This assumption would be true if the focus is on compliance only.

However, data integrity is not concerned with how valuable the data are or how it can be used for data analytics. Valuable data or value-added data can offer new insights (e.g. into data sets from HPLC analyses, pH values, UV-Vis spectral analysis, weight values, etc.) and provide full transparency, including all additional metadata (device, serial number, data & time, operator, ID's etc.). This is what gives the opportunity to conduct data analytics. Data analytics very often comes hand-in-hand with machine learning to develop new growth areas such as drug development [2] and really is the desired level to aim for. In addition, data analytics is used to make faster decisions, increase productivity or make predictions. All of this can only be successful with defined data architectures and laboratory data that are enriched with metadata. Including all this information, we can speak of value-added data (as shown in Figure 1: The Data Pyramid of Wisdom).

Figure 1: The Data Pyramid of Wisdom.

The focus of this white paper is on "data is the most valuable asset", with data governance being the most important topic to achieve this.

From the perspective of corporate IT, data governance is an "old" discipline, as regulatory requirements (recording obligations, financial regulations such as SOX (Sarbanes-Oxley Act), etc.) have had to be fulfilled for decades. The goal needs to be to establish data governance in the laboratory environment as well.

2 Understanding of Data Governance

To enable data to be a valuable asset to your laboratory, it is important to understand and adapt best practice procedures within the organization. Involving your IT or lab IT to understand the possible strategic direction for topics such as data harmonization, metadata requirements or data architectures is a good starting point. Figure 2 illustrates many aspects of data governance [3], where several aspects overlap with data integrity best practices. The areas of "Document & Content Management", "Data Quality Management", "Metadata Management" and "Reference & Master Data Management" (light blue in Figure 2) are particularly similar to Data Integrity.

Figure 2: Key areas of data governance (DAMA-DMBOK [3]).

Let's take a closer look at three important laboratory data issues (highlighted in green in Figure 2) that should be addressed in addition to the data integrity requirements already associated with data governance. Note: the other two areas (shown in dark blue) are not discussed in the scope of this white paper, because they are more IT-related.

1. Data Architecture Management is the process of defining and maintaining specifications that:
   - Provide a standard common business vocabulary
   - Express strategic data requirements
   - Outline high level integrated designs to meet these requirements and
   - Align with enterprise strategy and related business architecture

2. Data Security Management is the planning, development, and execution of security policies and procedures to provide proper authentication, authorization, access, and auditing of data and information assets.

3. Data Operation Management is the development, maintenance, and support of structured data to maximize the value of the data resources to the enterprise.

A big picture of the laboratory data, referred to as data architecture, is required, independent of the nature of the data (either purely digital based from connected solutions or manual collected values from offline devices). Within this big picture, data classification (e.g. personal data, test result data, measurement data, etc.) with access requirements, which is a part of the data security management, must be defined.

To drive the digital transformation in the laboratory and within the Data Operation Management to maximize the value of data, data breaks with interrupting data flows, should be eliminated or strictly limited. Data breaks are defined as situations such as manual transcription of values, manual data entries inside Lab Systems or an attachment of a document, which cannot be electronically accessed. Manual transcription from a unconnected device to Lab Data Solutions are error prone and limit the number of valuable metadata because of time consumption and the limitation of visibility of metadata from the user (e.g. last routine test, internal status, SW version etc.).

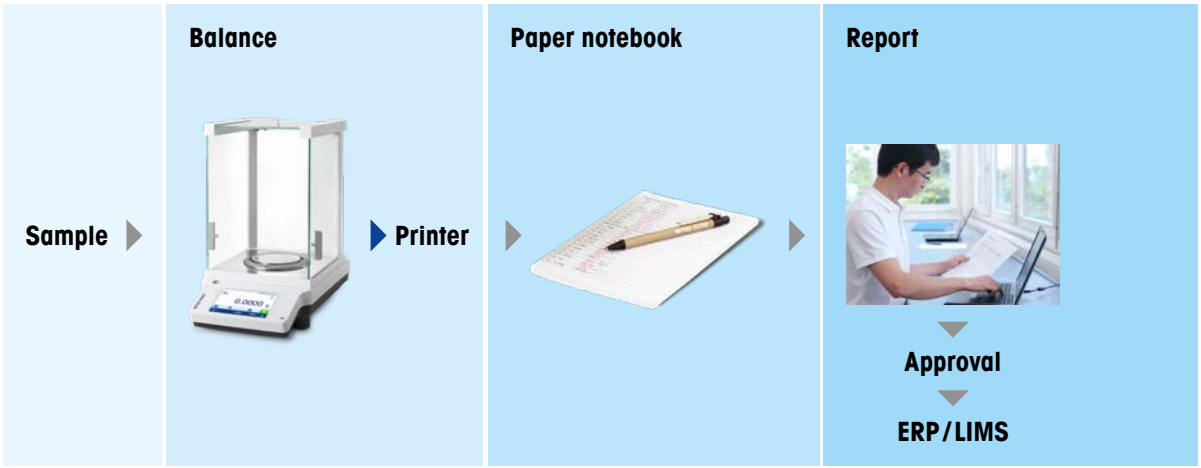

Figure 3: Data flow in a lab with weighing results documentation and storage on a print-out, eliminating one source of transcription error. However, further data processing with manual copy/paste retains the risk of making transcription errors and losing metadata.

The data architecture helps to define the data flow and workflows for the acquisition of laboratory data. In connection with the topic "Data Operation Management", the data flow between IT systems, analytical equipment and measurement data can be defined and implemented stepwise in a meaningful way. Unique identification of samples (LOT ID, sample ID, …), which forms part of the metadata, is important for the data analytics at a later stage.

Data Security Management takes care of access rights and data classification by limiting the global visibility of valuable data. In combination with the security concept for data, the protection of valuable data assets can be achieved. This is a contradiction in the area of DataLakes or BigData Analytics, because full access to all data is a requisite. Anonymizing personal or patient data for data analytics cases can solve this contradiction and help to fulfil regulatory requirements like GDPR (General Data Protection Regulation).

As explained above, a digital transformation is only possible if multiple areas of data governance come together to be implemented.

Manual data transcription has very limited opportunities for metadata, but with the next step of direct data acquisition using a "simple" data acquisition solution, this will offer new possibilities to enrich measurement data with some metadata. Automating data export from the acquisition software to an overlay data management system is part of the end-to-end data flow to meet data architecture requirements. Very often, with this type of solution, for example sample identification is limited or not possible, which is an important issue for data analytics.

The data architecture reduces the effort for the "BigData" analytics, because it is strongly related to data quality and data structures. Fully integrated digital solutions with workflow support on the devices and LES (Lab Execution System) will "Power the bench®" and are able to fulfill Data Integrity requirements for regulation purposes as well. METTLER TOLEDO's LES solution with different integration possibilities to data management systems like LIMS (Laboratory Information Management System), supports a seamless end-to-end solution.

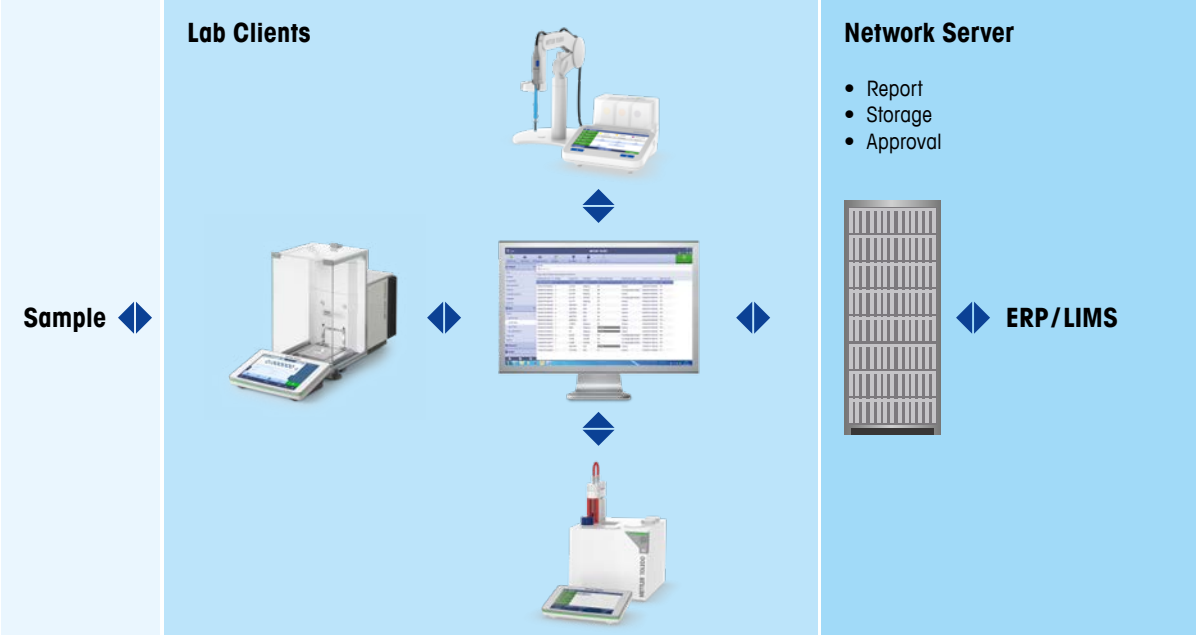

Figure 4: Data flow in a digital lab with automated transfer of data from the balance to the overlaying data system.

3 Data Integrity vs. Data Governance

The best way to achieve data integrity is to follow the ALCOA+ principles. In a regulated laboratory, it is mandatory to meet compliance requirements, such as those of the FDA. Data integrity is dealing with quality processes and result quality, but it has no focus on adding value to the results. Extending data integrity through data governance principles with a strategy for data lakes and data analytics will strongly support the growth of value-added data. These principles take care of access rights management, backup and archiving, whilst also having a strong focus on security aspects, which are increasingly becoming a key issue. METTLER TOLEDO's LES solutions such as LabX, iC Suite and STAR^{e} strongly support data integrity, but also data governance with a focus on value-added data with full metadata support.

4 Summary

METTLER TOLEDO's solutions support your digitization strategy, regardless of whether you have stand-alone devices with manual data transfer or fully integrated solutions. With respect to the fact that METTLER TOLEDO products are part of your laboratory equipment, we deliver important pieces of the puzzle for the big picture of a fully digitized laboratory.

5 Terminology

Data quality – ensuring data is generated without errors through use of proper, calibrated equipment, following SOPs, identifying and training users, and using the right materials. Systems should be designed in a way that encourages compliance with the principles of data integrity.

Data integrity – ensuring data is transcribed without errors and cannot be manipulated through proper set-up of data flow, integration of data and metadata, proper archiving, and ensuring accessibility. Good Manufacturing Practice (GMP) defines data integrity through the acronym ALCOA, which stands for attributable, legible, contemporaneous, original, and accurate. The original ALCOA principles have since been improved to ALCOA+. The original principles remain with four additions: Complete, consistent, enduring and available.

Metadata – data that provides information about other data, considered attributes of the measured values (e.g. sample identification, date, time, study number) and technical properties (e.g. instrument, calibration history, SOP, method version, etc.).

Audit trail – ensures traceability of electronic data. An audit trail is a complete historical record of who did what, when and why. Computer system design should always provide for the retention of full audit trails to show all changes made to the data without obscuring the original data. It should be possible to associate all data changes with the persons who made those changes, for example, by use of timed and dated (electronic) signatures. Reason for changes should be given.

Data Architecture – is composed of models, policies, rules or standards that govern which data is collected, and how it is stored, arranged, integrated, and put to use in data systems and in organizations. Data is usually one of several architecture domains that form the pillars of an enterprise architecture or solution architecture. (Wikipedia)

Data Governance – is a data management concept concerning the capability that enables an organization to ensure that high data quality exists throughout the complete lifecycle of the data. The key focus areas of data governance include availability, usability, consistency, data integrity and data security and includes establishing processes to ensure effective data management throughout the enterprise such as accountability for the adverse effects of poor data quality and ensuring that the data which an enterprise has can be used by the entire organization. (Wikipedia)

Data Lake – a data lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics and machine learning. A data lake can include structured data from relational databases (rows and columns), semi-structured data (CSV, logs, XML, JSON), unstructured data (emails, documents, PDFs) and binary data (images, audio, video). (Wikipedia)

6 References

[1] ALCOA Data Integrity poster, Document number 30417794A, October 2017
https://www.mt.com/global/en/home/library/guides/laboratory-division/lab-data-integrity/Data-Integrity-ALCOA-Poster.html

[2] Drug development get big data analytics boost, Novartis, July 2018
https://www.novartis.com/stories/discovery/drug-development-gets-big-data-analytics-boost

[3] The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK), April 2009
https://www.oreilly.com/library/view/the-dama-guide/9781935504009/

METTLER TOLEDO Group
Laboratory Weighing
Local contact: www.mt.com/contacts

www.mt.com/lab-software
For more information

30567498A
Group MarCom RITM625920 JK