Research Article

Laboratory Instruments' Produced Scientific Data Standardization through the Use of Metadata

Nur Adila Azram[1], Rodziah Atan[2], Fahmi Ibrahim[3]

[1]Halal Products Research Institute, Universiti Putra Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
[3]School of Business, Universiti Teknologi Brunei
nuradila.azram@yahoo.com[1]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: The progression of scientific data from various laboratory instruments is increasing these days. As different laboratory instruments hold different structures and formats of data, it became a concern in the management and analysis of data because of the heterogeneity of data structure and format. This paper offered a metadata structure to standardize the laboratory instruments' -produced scientific data to attain a standard structure and format. This paper contains explanation regarding the methodology and the use of proposed metadata structure, before summarizing the implementation and its related result analysis. The proposed metadata structure extraction shows promising results based on conducted evaluation and validation.

Keywords: Data standardization, laboratory instruments, metadata structure, scientific data

1. Introduction

Laboratory instruments have been used in various research areas to conduct experiments and produce results to gains information about many things. Each of the laboratory instruments has it owns purposes and functions. Some experiments sometimes need to use more than one laboratory instrument to gains data and information. It is difficult for the scientists and researchers in managing and analyzing data and information when involving multiple laboratory instruments as the data are overflows (Hsu, et al, 2015), using different information infrastructures (Wang, et al, 2006) and been described in a varied manner with diverse level of details (Rocca-Serra, et al, 2016). There is a need to have constant structures or formats of data that can make it easier for the management and analysis process.

Currently, researchers carry out experiments from more than one laboratory instruments facing problem in managing and analyzing data as researchers need to take into account various aspects of the structures, types, or format of the scientific experimental data involved due to isolated information and various sources. Thus, there is a need for data standardization for the laboratory instruments experiment data. Data standardization is the method of changing data from various sources and systems into a uniform structure (Data standardization - Data transformation, 2015). It helps in clearing out inconsistency in the data attributes or properties to ease the data management. Various multi-disciplines research areas such as medical and food science generated scientific experimental data that have been increasing rapidly. These scientific experimental data are data made by measurement, test method, experimental design, or quasi-experiment design (Experimental Data, 2019).

Metadata standard is one of the essential elements in data standardization. Metadata facilitate in providing resource with a structure and strengthens the data acquisition and management. Metadata ensures that data management and integration can be done by particularly discovered and correctly described data to support future retrieval and reuse. Metadata is widely used in the standardized and unified management of data resources. The establishment of metadata standards is the premise and guarantee of data standardization. Issues of data management can be sorted out efficiently using metadata. Using a standards-based approach in organizing data can assist in guaranteeing adaptability between systems, as well as increase data discovery and access. The implementation of standardized metadata practices also increases data exchange possibilities and competences (Wiser, S. K., et al, 2011) and (Hussain et al, 2019).

2. Background

Data standardization converts data from different sources into a standardized format. It helps in data management by eliminating contradiction in attributes or properties of data. Agreeing on standard data format, metadata, and vocabulary standards is an essential phase to acquire the necessary data interoperability level to add value (Yeumo, et al, 2017). Proficient data standards are the basic establishment for solving scientific

discovery by empowering effective storage, management, sharing, and analysis of data (Rubel, et al, 2016).Perhaps best known theory along these lines is the most widely known differences in contemporary knowledge literature where one of the most prominent starting points toward a concept is to differentiate between data, information,and knowledge interconnected in a hierarchical system where the connection is mainly uni-directional. Basically, data is a necessary requirement for information, and information is a necessary requirement for knowledge. Nevertheless, knowledge basic structure can be obtain through data and information, which is the same as data and information that can be generate from knowledge. Hence the relationship between them in a reversed hierarchy (Tuomi, 1999) is diverse and collaborative rather than merely uni-directional. This can clarify the issue of providing 'standardized data' which may not always be feasible as knowledge can be viewed as data or information with additional layer of intellectual analysis applied, where interpreted, connected, and organized and attached to current systems of beliefs and knowledge bodies structures (Hislop, 2013).For example raw data from laboratory experiments has been analyzed using specific statistical technique, to produce some structured results.

The attempts to define metadata structures for data standardization pose problems in several ways often in the fields of knowledge management (KM). Building expert systems based on symbolic artificial intelligence (AI) research has previously been the aim of the first generation of KM work. KM systems have aimed at obtaining and storing the knowledge of experts in a database, mostly for decision-making purposes. Despite their success, there are drawbacks to these strategies, in general. There is an expectation that knowledge is a 'thing' or an element which can be produced clearly and codifiable or 'standardized data' which can result in 'information overload' where the volume of information/data is unmanageable (Ibrahim and Reid, 2010) which has caused it to lose its 'tacitness' (Hislop, 2005:p.110).

In reality, experts have always proved quite difficult (or very time-consuming) to provide a summary of their (sometimes tacit) information that is separated from the actual operation (Greiner, 2007). Removed from the context of their practice, experts fail to express the skills, expertise, and methodologies that make up their knowledge. Drawing from the theory of KM, this viewpoint suffers from 'synoptic delusion'; the inaccurate idea that organizational information can be gathered as structured data in a single repository (Tsoukas, 2005:p.100).
Nevertheless, data standardization has been important to many research areas to facilitate the issue of data management due to the lacks of data standardization. For example, the medical science research area consists of data from many sub-areas that need to have a standard in the management and integration of medical data to ensure complete and consistent data.

Metadata is defined as "data about other data". It provides information on other data by describing the content of data to achieve data discovery, management, and sharing (Ying and Gengda, 2014). It gives structure to resources and enhances data management and discovery. Integrated metadata can handle an enormous amount of distributed data in a clear and competent manner, and help to efficiently discover, search, incorporate, and handle and apply information resources effectively (Zhang, J., et al, 2018). Metadata standard is a set of metadata elements that are organized and defined for a particular purpose. Many metadata standards have been developed to address certain information use and management prerequisites for various domains. Table 1 shows some of the metadata standards with the descriptions of each metadata standard.

Table 1. Metadata standards with the descriptions

Metadata Standard	**Descriptions**	**Domain**
Dublin Core (DC)	- Contains fifteen elements. - Identify digital libraries' resources.	Various (e.g: Cultural heritage)
Darwin Core (DwC)	- Contains seven simple DwC elements and two generic DwC elements. - Provide a common reference for sharing information on biological diversity.	Biology
Ecological Metadata Language (EML)	- Provides a framework for scientists to summarize rich semantic descriptions of data in their metadata.	Ecology
Data Documentation Initiative (DDI)	- Used to document and describe various social science research data. - Capture information and to present it in a machine-actionable format.	Social Science
Visual Resource Association (VRA) Core	- Describe works in arts and cultural objects. - Allow creating, describing, and distributing digital	Arts and humanities

objects for resources.

There are many research that proposed metadata standard to cater issues and problems on data standardization in their respective research area. For example, a study by Damerow, J., et al (2019) conducted a pilot test to decide on essential standardized metadata for physical samples in the earth and environmental sciences. A well-organized system for tenacious sample identification and tracing that is appropriate for the field, laboratory analyses, and online publication are needed as data providers to the Department of Energy's (DOE) Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) repository regularly work in huge, interdisciplinary teams and send samples to various facilities for analyses. They offer useful recommendations for effective sample data management while also maintaining and exploiting the potential value of samples into the future.

Another example is from Keller, R. M., et al (2018) that has developed a new data system with a metadata standard to enhanced Astrobiology Habitable Environments Database (AHED). The system is established as a lasting, open-access repository for astrobiology data. The improvement was done as a result of the interdisciplinary nature of astrobiology that presents some explicit challenges to data management, integration, and analysis within AHED. The new metadata was proposed for relating astrobiology datasets, with complete information about content, funding source, and scientific significance, alongside a set of relevant keywords for describing datasets. With the improvement, AHED will be able to provide better search, finding, and analysis competencies.

Brown, G., et al (2017) has proposed Atlantic Ecosystems Initiative (AEI) project to enhance the accessibility of marine species event data. The focal point is to extract content from issued articles and produce a set of standardized archives that must satisfy the Darwin Core standard that can then be shared with the greater biodiversity data community.

3. Methodology

This study design a metadata structure to standardize and gives a standard structure to scientific data from laboratory instruments. To design the metadata structure, this research proposed a few phases to be followed. Figure 1 shows the phases of the metadata structure development.

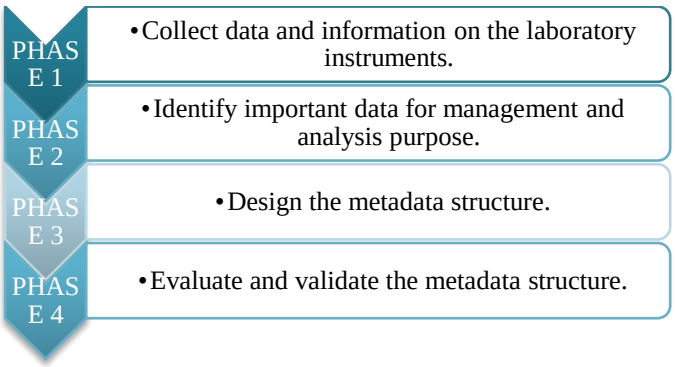

Figure 1. The phases of the metadata structure development.

Based on Figure 1, phase 1 is to collect data and information from the laboratory instruments. In this study, two laboratory instruments were selected to be used in the development of the metadata standard design which is Gas Chromatography-Mass Spectrometry (GC-MS) and High-Performance Liquid Chromatography (HPLC). These two instruments were selected because they were commonly used together in experiments to obtained overall characterizations of compounds as well as function under the same basic principles of compound separation, identification, and quantification techniques. The data collected were data that were generated from the laboratory instruments which contains information on the sample and result of an experiment.

In phase 2, the important data for management and analysis purposes were identified from the selected laboratory instruments. This is done by cross-checking with the researchers using both laboratory instruments on the important data they needed for managing and analyzing the data. Then phase 3 is to design the metadata structure. Figure 2 shows the elements in the metadata structure for laboratory instruments' scientific data.

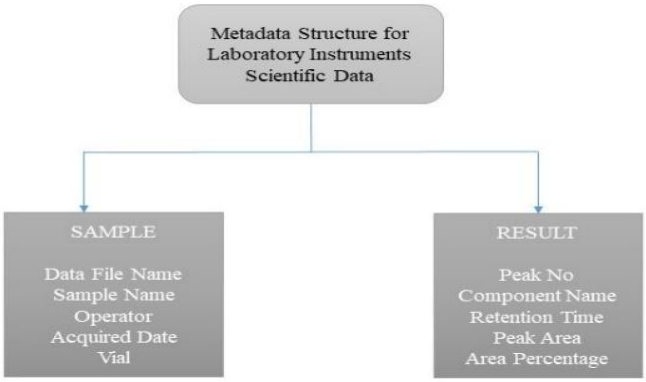

Figure 2. The elements in the metadata structure for laboratory instruments scientific data

Based on Figure 2, the metadata elements are divide into two categories named Sample and Result. Both Sample and Result comprise of five metadata elements. Table 2 shows the descriptions of Sample metadata elements and Table 3 shows the descriptions of Result metadata element.

Table 2. The descriptions of Sample metadata elements

Metadata Element	Descriptions
Data File Name	The name of the data file.
Sample Name	Name of the sample.
Operator	The name of the person that operates the instrument.
Acquired Date	Date of data file acquired.
Vial	Temporary container to hold a sample.

Table 3. The descriptions of Result metadata elements

Metadata Element	Descriptions
Peak No	The peak number of a component.
Component Name	The name of a component.
Retention Time	The time taken for a particular component to travel through the column to the detector.
Peak Area	The peak area of a component.
Area Percentage	The area percentage of a component.

Lastly, phase 4 is to evaluate and validate the proposed metadata structure. The evaluation is done through metadata extraction and validation is done through precision and recall analysis. In this study, precision is defined as extracted metadata that is relevant and recalls as relevant metadata that is extracted. The precision and recall were calculated in percentages. The calculation formula for precision is as in Eq. [1] and recall as in Eq. [2] below.

$$\text{Precision \%} = \frac{Number\ \ of\ relevant\ \ metadata\ \ extracted}{Number\ \ of\ metadata\ \ in\ collection} \times 100\% \quad [1]$$

$$\text{Recall \%} = \frac{Number\ \ of\ relevant\ \ metadata\ \ extracted}{Number\ \ of\ metadata\ \ extracted} \times 100\% \quad [2]$$

4. Result and Discussion

As mentioned in the Methodology section above, for the evaluation of the metadata structure, the evaluation was done by conducting metadata extraction to the selected laboratory instruments data files. Metadata elements based on the proposed metadata structure were extracted from the data files. The total of metadata elements extracted as well as the total of each metadata element extracted from the data files was calculated. 100 data files were selected from each of the laboratory instruments for the evaluation purpose. Table 4 shows the total of metadata extracted from each selected laboratory instrument data files. Table 5 shows the total of each metadata element extracted from the 100 data files of the GC-MS laboratory instrument and Table 6 shows the total of each metadata element extracted from the 100 data files of the HPLC laboratory instrument.

Table 4. The total of metadata extracted from each selected laboratory instrument data files.

Laboratory Instrument	Total of metadata extracted
GC-MS	9
HPLC	9

Table 4 shows that the total of metadata extracted from the GC-MS and HPLC laboratory instruments were the same with nine metadata extracted. One metadata element was not extracted from both instruments as the element was not presented in the data file.

Table 5. The total of each metadata element extracted from the 100 data files of the GC-MS laboratory instrument.

Metadata Element	Total extracted from the 100 data files
Data File Name	100
Sample Name	100
Operator	100
Acquired Date	100
Vial	100
Peak No	100
Component Name	98
Retention Time	98
Peak Area	0
Area Percentage	98

Table 5 shows that most metadata elements were extracted from the one hundred data files of the GC-MS laboratory instrument. However, three metadata elements only extracted from ninety-eight data files which were Component Name, Retention Time, and Area Percentage. This means that the metadata elements were not obtainable in some of the data files. One metadata element was also not extracted from all of the one hundred data files which is Peak Area which means the metadata elements were not obtainable at all in the data files.

Table 6. The total of each metadata element extracted from the 100 data files of the HPLC laboratory instrument.

Metadata Element	Total extracted from the 100 data files
Data File Name	100
Sample Name	100
Operator	0
Acquired Date	100
Vial	100
Peak No	100
Component Name	100
Retention Time	100
Peak Area	100
Area Percentage	100

Table 6 shows that nine metadata elements were extracted from the one hundred data files of the HPLC laboratory instrument. However, there was one metadata element that was not extracted from all of the data files which is Operator due to its absence in the data files.

Based on the results of the evaluation, the validation of the metadata extraction was done through precision and recall analysis. Based on the precision and recall analysis, the result of precision was 90% for both selected laboratory instruments, and for recall, the result was 100% for both selected laboratory instruments. The results were in the acceptable range of validation which means that all relevant metadata was extracted and only relevant metadata was extracted based on the proposed metadata structure.

5. Conclusion

Scientific data standardization from laboratory instruments is important as these data are rapidly increasing nowadays. Researchers involved with laboratory instruments scientific data in many research areas such as biomedical need standardized data to facilitate in the management and analysis of data especially when involving data from different laboratory instruments that hold different structures and formats of data. This study has taken an effort to make scientific data from laboratory instruments in a standard structure. A metadata standard for the laboratory instruments scientific data has been proposed to give a standard structure to represent the scientific data which can be use by researchers to compare and evaluate the results from multiple laboratory instruments for analysis as the data would be in the same structure.

References

1. Damerow, J., Agarwal, D., Boye, K., Brodie, E., Cholia, S., Elbashandy, H., ... & Jones, M. B. (2019). Community use of persistent sample identifiers and metadata standards: supporting efficient data management in the field, laboratory, and online. AGUFM, 2019, IN32A-05.
2. Data standardization - Data transformation. (2015, September 10). Retrieved from Experian: https://www.edq.com/data-quality-management/data-standardization/
3. Experimental Data. (2019, 6 January). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Experimental_dataFachinger, J. (2006).
4. Hsu, L., Martin, R., McElroy, B., Kim, W., & Litwin Miller, K. (2015). Data management, sharing, and reuse in experimental geomorphology: Challenges, strategies, and scientific opportunities. Geomorphology, 180-189.
5. Hussain, A., Surendar, A., Clementking, A., Kanagarajan, S., Ilyashenko, L.K. (2019). Rock brittleness prediction through two optimization algorithms namely particle swarm optimization and imperialism competitive algorithm. Engineering with Computers, 35 (3), pp. 1027-1035.
6. Keller, R. M., Detweiler, A. M., Lafuente Valverde, B., Blake, D. F., Bristow, T. F., Cooper, G. W., ... & Parenteau, N. (2018). Steps Toward Improved Integration, Search, and Analysis of Heterogeneous Data in the Astrobiology Habitable Environments Database.
7. Rocca-Serra, P., Salek, R., Arita, M., Correa, E., Dayalan, S., Gonzalez-Beltran, A., Neumann, S. (2016). Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics.
8. Rübel, O., Dougherty, M., Prabhat, Denes, P., Conant, D., Chang, E., & Bouchard, K. (2016). Methods for Specifying Scientific Data Standards and Modeling Relationships with Applications to Neuroscience. Frontiers in neuroinformatics.
9. Wang, F., Pearson, J., Liu, P., Azar, F., & Madlmayr, G. (2006). Experiment Management with Metadata-based Integration for Collaborative Scientific Research. 22nd International Conference on Data Engineering. IEEE.
10. Wiser, S., Spencer, N., De Cáceres, M., & Kleikamp, M. (2011). Veg-X – An exchange standard for plot-based vegetation data. Journal of Vegetation Science, 598-609.
11. Yeumo, E., Alaux, M., Arnaud, E., Aubin, S., Baumann, U., Buche, P., . . . Quesneville, H. (2017). Developing data interoperability using standards: A wheat community use case. F1000 Research.
12. Ying, Y., & Gengda, J. (2004). Metadata-based information organization and ontology-based knowledge organization. Journal of Academic Libraries, 43-47.
13. Zhang, J., Chen, H., & Wang, K. (2018, November). Study on Standardization of Detection Data of Atmospheric Microparticle Lidar Based on Metadata. In 2018 14th International Conference on Computational Intelligence and Security (CIS) (pp. 481-484). IEEE.