Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 11, 2025

Harmonization of Discrepant Data: A Solution to the Computational Models for Data Collection in the Tertiary Institutions

Aburuotu, E. C.
Department of Computer Science, Faculty of Natural amd Applied Sciences, Ignatius Ajuru University of Education, Port Harcourt, Rivers State, Nigeria
Nathaniel A. O.
Department of Computer Science, Faculty of Natural amd Applied Sciences, Ignatius Ajuru University of Education, Port Harcourt, Rivers State, Nigeria
Email: emmanuel.aburuotu@iaue.edu.ng, Nataniel.ojekudo@iaue.edu.ng

Abstract:
Data visualization, interoperability, analysis, and business decisions face a significant challenge as a result of the growing volumes of heterogeneous data being produced by agencies and institutions in the education sector. Stakeholders in the education sector should implement Harmonization of Heterogeneous Data-set as a critical solution. The legacy data that is analyzed and is intended to be used in decision support systems and analytical applications is imported from various data sources with different data types and database architectures and structures. All of these data need to be harmonized for the intended business solutions and growth. Therefore, the Support Vector Machine (SVM) algorithm for Heterogeneous Data Harmonization technique has emerged as the most effective method for creating high-quality data intended to enhance the governance and usefulness of its purpose across the enterprise. The goal of the research was to create and refine a support vector machine-based heterogeneous data harmonization solution for enterprise databases. A data harmonization technique was developed with the integration of harmonization tools using a support machine learning algorithm, and the work was implemented using the Java Script (JS) development environment. The study looked at existing data production techniques on various active databases. To accomplish its goals, this work used the Rapid Application Development (RAD) system methodology. The system's AI machines were tested and trained using both structured and unstructured data imported from Microsoft Excel applications, thanks to the Supervised Machine Learning procedures. There were 10,990 different data sets that were used for training and testing. Testing was conducted on 8,393 (70%) datasets, while 2,597 (30%) were used for training. The outcomes demonstrate that the system was successful in redefining the data headers and column dimensions as a means of coordinating the pull of data imported into the system.

Keywords: Data Harmonization, Analytics, Visualization, Dashboard Systems, Platform, Interoperability, Machine Learning, Algorithms.

1. Introduction

Solutions for data harmonization have evolved to be a significant problem for the administrators in the education sector, particularly when it comes to managing diverse data sets. Computational tools and techniques from artificial intelligence (AI) are essential for managing the massive amounts of data that real-world applications generate every second. Some of the real-world application areas for big data include healthcare, telecommunications, financial services, retail, law enforcement, marketing, new product development, banking, energy and utilities, insurance, education, agriculture, and urban planning. Textual data is being produced from a range of sources in a variety of formats, from structured and semi-structured to unstructured (SSU). Decision-makers have a challenge when attempting to make judgments based on the dispersed data since dissimilar data cannot be analyzed using basic tools and procedures. Data volume increases as well [1]. Because disparate data cannot be analyzed using basic tools and procedures, decision-makers face a barrier in making decisions based on the scattered data. As data volume grows, so does its variety. Investigating such heterogeneous data is one of the most difficult challenges in data analytics and information management.

Data visualization and prediction are influenced by the variety and decentralization of data sources, which in turn influences analytical results and business decision [2]. Data harmonization (DH) is a field that integrates the representation of data with such a diverse nature. Several tools and solutions have been developed over the years to reduce the variety and disparity in formats of big-data type and management. Distributed database systems run across enterprise domains and these databases are designed to store textual data from multiple sources. The textual data that is imported from various databases with different structures need to be harmonized for the intending enterprise solution. Data Harmonization has become the best approach in producing quality data that is meant to improve the governance and usefulness of its purpose across the enterprise. Anil defined Data Harmonization as a method of creating a single source of truth. It does this by taking data from disparate sources, clearing away any misleading or inaccurate items, and presenting it as a whole [3]. This means that one gets a single-window view of everything and anything that supports ongoing decision-making, including financial information and business performance. Data comes to you from different sources, but once it is harmonized, it has been cleaned, sorted, and aggregated to provide a complete picture [4]. Due to the many sources data is collected from, and the volume of data being analyzed, organized, or sorted in this 21st century for effective decision making, there is no better way of doing it than the integration of smart data tools like Machine Learning and Artificial Intelligence. These tools play key roles in the process of data harmonization. Machine Learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, Machine Learning allows computers to find hidden insights without being explicitly programmed where to look [5].

Growing volumes and varieties of available heterogeneous data, cheaper computational processing, and more powerful and affordable data storage and mining capabilities mean that it is possible to quickly and automatically produce models that can analyze bigger, more complex data and deliver faster, more accurate results even on a very large scale. To build precise data management models, organizations can explore better chances of identifying profitable opportunities or avoiding unknown risks. This concept and idea is supported by machine learning tools and artificial intelligence. There is this data wrangling problem that is growing as different types of unstructured data or data in varying formats are pouring in from sensors, online, and traditional databases. All these data must be cleaned up and organized before Data Analytics tools can be applied [6]. This is where automation tools come into play. Automation and artificial intelligence (AI) can help organize data from a variety of sources, then present the organized data in charts and graphs. Artificial Intelligence can supplement to make it easier to prepare and harmonize data, thereby speeding mainstream adoption of big data techniques. Indeed, the growing diversity of data coming from emerging networked sources like the Internet

of Things is fueling demand for more and better automation tools [7]. Data Harmonization goes one step ahead and involves cleaning of data to remove any inconsistencies and inaccuracies from various sources of data. It tries to create "harmony" between different sources of data to create a complete, cohesive picture. The Data Harmonization process is like joining the pieces of a puzzle to make sense, wherein the different data elements and variables are identified, cleansed, and processed together to create a data store that facilitates decision making.

Overview of Data Harmonization

The use of non-standard, country-specific, and/or agency-specific data is highly inefficient in terms of cost and accuracy, both for the government and the education sector has enabled the revolution of discrepant data in the education sector. Regulatory authorities are required to maintain or develop agency-specific systems, and trade must operate and maintain interfaces to meet these redundant and duplicative reporting requirements. This level of duplication is also evident in non-automated, paper-based systems where records generated in the education sector are required to provide highly redundant forms. The situation is especially critical for universities that must interact with many students and administrators and many other government agencies. The cost and complexity of meeting these requirements are staggering. Addressing this issue would benefit not only the educational stakeholders but also other partners who may intend the use the data for other purposes. Techpoint noted the issue of non-unified databases in Nigeria and offered a straightforward solution. He noted that data was collected and stored everywhere, most of which include personal data such as name, sex, address, blood type, place of birth and date of birth etc. These data are therefore collected through various parastatals and government databases, making it impossible for accurate information to be collected [8].

Problem Statement

The distributed databases that coordinate the storage of the datasets in the education sector do not all share a common functionality called data interoperability. This is due to the lack of an integrated single window (ISW) for data collection based on appropriate data management standards in the Nigerian education sector. Synchronization and data reuse across Nigerian agencies and organizations have been made impossible by the non-uniformity of data formats, data management software, database models, and development architectures. The data set used in decision support systems in the education sector is now imported from a variety of sources with diverse data kinds and database structures [9]. For the intended enterprise solution, the data that is imported from several databases with diverse structures needs to be harmonized. Additionally, the difficulty of database systems to synchronize data through various data transmission methods is a result of the inconsistent data from non-uniform databases. Therefore, the ideal strategy for producing high-quality data that is intended to enhance governance and its usefulness for its intended purpose across the company will be to harmonize these heterogeneous data [10].

When data is harmonized, it is taken from several sources, any false or misleading information is removed, and the data is presented as a whole. Users now have access to anything and everything that supports ongoing decision-making, such as financial data and business performance, in a single window. There is still a wrangling issue because different types of unstructured data or data in varying formats are pouring in from sensors, online portals, and traditional databases. This is true even though agencies in the education sector have resorted to building precise models in order to stand a better chance of identifying lucrative opportunities or avoiding unknown risks. Before using data analytics tools, all of these data must be cleaned up and organized. Tools for automation are useful in this situation. Data from many sources can be organized with the aid of automation and artificial intelligence (AI), which can then display the organized data in graphs and charts. Artificial intelligence can complement to make it simpler to organize and harmonize data, hastening the acceptance of big data by the general public [11].

Related Literature

Organizations that provide teaching, research and information services, like educational stakeholders work with a variety of proprietary and multilingual public sources, and for this, they must cope with enormous amounts of data, naming standards, and contexts that come from numerous sources [12]. It takes time, requires a lot of work, and is prone to error to map such data in a company's reporting and prediction master data. Such sources are impossible for machines to precisely match and map the data to the master. In order to free up their time to focus on finding insights to propel the business, it has become important to automate the labor-intensive data harmonization chores [13]. A global Market Research Major (MRM) Automation Initiative of this kind was undertaken by Deb, who used artificial intelligence (AI) techniques to great effect. A multi-step method to dictionary matching, fuzzy text similarity, and several machine learning algorithms provided the answer to automated data harmonization (ADH). It was introduced to the big data stack for improved performance and scalability. Workflows and runtime rules have been added to streamline the overall business process [14]. The proof of concept produced an average F-score ranging from 82 to 93%, depending on changes in the data gathering. The deployed version continues to offer great precision and is accepted by the entire company as a key micro-service. Business as usual (BAU) cycle time was reduced by 80% (from 14 days to 3 days). Multifunctional devices, author name and quotation resolution in journals, lead resolution in multi-channel marketing, advertising campaigns, etc. can all be used to standardize media metadata even when the solution is unique and specifically created for specific business objectives [15].

Figure 1 illustrates how data harmonization identifies and eliminates ambiguity, links records, and establishes standards.

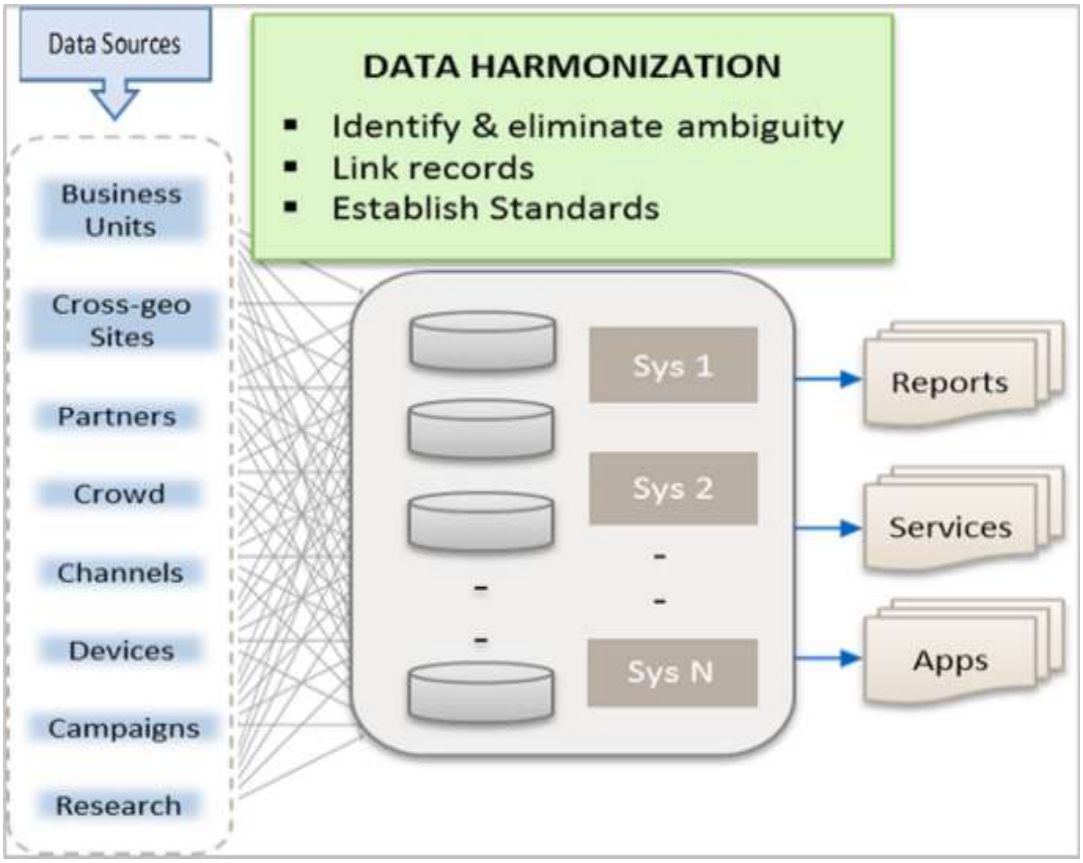


Figure 1. Data Harmonization Techniques.

According to Kipp, machine learning and artificial intelligence are prepared to have an impact on virtually every aspect of human life. Their research offers advice to physicians on key facets of AI and machine learning, analyses specific cardiological applications to date, and identifies possible AI applications in cardiovascular medicine in the future [16]. The first paper discusses

the fundamentals of predictive cardiology modelling, including feature selection and typical flaws such improper dichotomization (represent as split or opposed). Second, it looks at standard techniques for supervised learning as well as a few applications in the field of cardiology and related fields. Third, the paper discusses the creation of a wide range of knowledge and related methodologies known as non-monitored research. The article provides qualitative examples from both cardiovascular and general medicine. In order for rich data advances like genome-sequencing and the streaming of biometric mobile devices to be clinically adopted, cardiologists will need to analyse and operationalize data from numerous different biomedical domains. At the same time, in order for doctors and healthcare systems to function well, external stress is becoming more and more important in medicine [17]. Doctors in the medical field are overloaded with data that, when called for to increase productivity, requires more complex analysis. Each stage of patient care may be made better by machine education, from research to discovery through diagnosis and treatment selection. Clinical practise would become more convenient, adaptable, accessible, and efficient as a result.

Additionally, information won't be acquired solely for medical purposes. Future clinicians will be able to remotely and automatically track, analyse, and react to additional streams of biological data because to the spread of mobile sensors. In this technological coin, common ways to machine learning are brought to the medical system [18]. These approaches review numerous applications chosen specifically for cardiology and foresee the potential integration of AI by cardiovascular medicine. Unattended learning and supervised learning are two categories of machine learning approaches. These serve a variety of purposes. While supervised learning also involves categorising an observation into one or more categories or results, uncontrolled learning is based on identifying the underlying structure or connections between variables in the data set. Thus, a data set comprising predictor variables ("features") and results that have been labelled is needed for controlled learning. Predictive modeling is commonly done in medicine where findings have labels such as "events" or "checks," which are combined with related elements such as age, sex, or clinical variables [19].

Reliable Machine Learning for Data Abstraction through Automated Quality Control and Data Harmonization

Machine Learning for Data Abstraction through Automated Quality Control and Data Harmonization is a machine learning technique -Robinsons designed that can harmonize data other than student records data. He claimed that artificial intelligence robots were being used more and more for data processing, educational record extraction, decision-making, and other national data utilization requirements. Some of these analytical findings are additionally applied to clinical decision-making, from diagnosis to therapy planning. In order to prevent bias or inaccuracy from being employed in subsequent processes, machine learning models must be trustworthy and their conclusions must be thoroughly scrutinized [20]. A key component of numerous pathways for medical imaging research is picture segmentation. It is necessary to establish instances where a segment has failed for it to be correct. Robinson's study offers a thorough validation of the automated quality control (QC) method for segmenting cardiac magnetic resonance images (CMR) utilizing a reverse testing strategy modified from Reverse Classification Accuracy (RCA).

The outcomes of the UKBB imaging research are used to assess the efficacy of the strategy in the absence of Ground Reality (GT). The segmentation accuracy of the Dice Similarity Coefficient (DSC) is greater than 95%. In high-performance pipelines, pharmaceutical research, or hospitals, it may be useful to quickly analyse the segmentation quality in order to notify the user of photos acquired if they produce mediocre results. Deep learning (DL) frameworks that explicitly predict DSCs and RCA proxy scores were used in a real-time automated QC process to achieve MAE values of 0.03 and 0.14. A model might not work because it was trained on one distribution and data from another distribution was used to analyse scans from other places. For multi-site neuro-imagerial data, the paper offers and evaluates an unpaid, unattended Domain Adaptation

strategy (DA) to reduce the demographic and acquisition-related worsening of model output caused by Domain Shift (DS). Over 20% efficiency is gained in a classification job (ISTNs) when constrained appearances and space transformations are used across image and space networks [21].

Muhammad indicated that water and energy system interaction modeling is critical for infrastructure security compliance and system sustainability. Recent technological developments have enabled large amounts of data to be generated in connection with the functioning of those sectors. According to Muhammad, Khan more and more industries are beginning to realize, in terms of water availability, transport, use, and energy generation, fuel supplies, and customer demand, and interdependencies between those systems that can put these systems susceptible to cascading impacts as a result of single disturbances, that statistical and machines can contribute to the development of characteristics across the systems [22]. For the simulation and forecasting of water and energy supplies in the Nexus, different modeling techniques can be used. These can be categorized as process-based or data-driven techniques. Process-based is a mathematical technique that provides a detailed representation and interpretation through scientific principles of the underlying processes of variables within a system. Data-driven methods use data to capture the relationship between the system variables, while no physical processes are described in a system. Process-based techniques have the advantage that models are more valid and useful since they are based on scientific principles and laws to enable profound comprehension of the underlying processes [23].

Liang suggested that methods based on artificial intelligence (AI) have been identified as powerful instruments for transforming all industries involved in data production and usage. Although the classifiers of machine-learning (MLC) have already demonstrated strong image-based diagnostic performance, text-data analysis, analyzing a variety of massive electronic education-based data is still difficult. They demonstrated that Machine Learning Classifiers can consult tertiary-based data in ways that do not exist in past statistical methods similar to the hypothetic-deduction reasoning employed by data analysts and software engineers Liang's model also uses a deep learning technique to extract clinically relevant information from EHRs to use the automated natural language processing system. In total, 101.6 million data points from 1,362,559 pediatric patient visits presenting to a major referral center were examined to train and validate the system. Our model shows high diagnostic precision across many organ systems and is similar to experienced pediatricians in the diagnosis of common infancy conditions. Liang's study provides proof of the principle of the introduction of an AI system, which can help doctors handle vast quantities of data, improve diagnostic assessments and facilitate clinical decision-making in the event of diagnostic confusion or complexity. While this effect may be more apparent in areas with comparatively low health care providers, the advantages of such an AI system are likely to be universal.

Firoj have proposed that microblogging sites such as Twitter are increasingly used in natural disasters and emergencies, hence the need for a harmonized record of these users. The utility of the information on Twitter for many disaster response tasks was shown by research studies. The sensitivity of social media data, however, is a difficult task due to many factors, such as limits to available tools for analysis of high volumes and high-speed data streams, including the overloading of information. In its work, they first demonstrated that textual and social media material offers complementary knowledge that is useful for improving situation consciousness to remove these limitations. We then explore how various natural language processing and supervised learning as an aspect of artificial intelligence techniques can leverage the complementary data produced in the universities for harmonization and ensure its use across many platforms. Finally, we propose an approach based on a methodology that combines several computer techniques effectively in a unified framework to support the relief efforts of humanitarian organizations.

Several techniques and computational methods were suggested for artificial intelligence in

recent years to learn useful information from social media data to respond to disasters and manage disasters in times of critical concern. These methods are intended to address many problems including filtering, overloading, and categorization of information to summarization. During the three main disaster events of the 2017 Atlantic Hurricane season, Stieglitz and Milde performed comprehensive studies using textual and imaging contents from millions of tweets. The study shows that the distribution of different types of useful data can inform crisis managers and response personnel and facilitate the development of future automated disaster management systems [24].

Visvikis suggest that artificial intelligence techniques and more specifically (deep) learning methods were core components of the latest developments in medical imaging. They are already exploited and are designed to deal with most tasks, including re-construction of images, processing.

Hadi and Rigoberto have proven to be a powerful alternative to classical modeling techniques in the field of artificial intelligence. AI refers to the informatics industry which develops human intelligence machinery and software. AI provides benefits for addressing problems dealing with uncertainty in comparison to conventional approaches and is an efficient aid for resolving such complex issues. Moreover, AI-based solutions provide good alternatives where testing is not possible to evaluate engineering design parameters resulting in considerable savings in human spending and effort. AI can also speed up decision-making, reduce error rates and improve calculation performance. The different AI techniques have recently attracted considerable interest, including machine learning (ML) and pattern reconnaissance (PR), and profound learning (DL), and are becoming an intelligent new class of methods for the use of structural engineering. This review paper aims at summarizing the techniques related to the application in the last decade of the described AI methods in structural engineering. First, there is an overview of AI and descriptions of the role of AI in structural engineering. Thereafter, the analysis and the capability of such methods for the regulation of restrictions on traditional models will be addressed with the recent applications of machine learning (ML), and deep learning (DL) on the ground. In addition, it discusses in-depth the benefit of using these algorithmic approaches. The following are presented as well as their possible research avenues and emerging patterns for the use of ML and DL [25].

Thesmar, suggested including diverse and wealthy sources of information by integrating the data on health claims and additional information sets. These blended datasets can be dimensional and difficult to manage with traditional statistical analyses. Recent artificial intelligence advances (AI) have, however, led to algorithms and systems capable of studying and extracting complex data patterns. In combined studies such as the improving of the pipeline for insurance claim processing and the reduction of estimate bio-sets, the AI has already been implemented successfully. The use of artificial intelligence (AI) with claim data enables common health prejudices such as the doctor's and the excluded variable prejudices to be identified and reduced. Nevertheless, much more can still be said. Complex trends in highly-dimensional datasets can be detected through finding new predictors of early diseases or by providing tailored preventive services more proactively. AI's ability to identify complex trends in claims collected with other sources can contribute to improved treatment and insurance coverage. Although the use of AI is associated with possible risks and difficulties, they cannot be overcome. As with the implementation of every advancement, the use of AI approaches in healthcare needs to be thoughtful and responsible.

Mühlroth and Grottke, noted that companies apply strategic technology foresight and innovation management to identify trends early on, evaluate their expected effects and develop the future course of action that will allow superior business performance to be delivered. A growing volume of data needs to be obtained, analyzed, and interpreted for this purpose. However, a substantial part of these operations is carried out manually, requiring high investment in different resources. They introduced an artificial intelligence data mining model

that allows businesses to identify emerging topics and patterns to a higher degree of automation than before to support these processes more effectively. Its modular framework consists of query modules, data collection, data preprocessing, thematic model, subject analysis, and visualization, all of them in such a way that the initial configuration requires only a minimal amount of manual effort. The solution also includes self-adaptive capabilities to update the model automatically once new data is generated. Model parameters are based on recent research in this field, and their threshold parameter is taught through a training dataset during supervised training. To check its efficacy as an alert system, we have applied our model to an independent test data collection. In the three case studies, we illustrate, through the retrospective review which in advance of their first publication in the Gartner Hype Cycle for Emerging Technologies our model can classify new technologies. We draw both theoretical and practical consequences for companies' technology and innovation management based on our findings and suggest potential research opportunities to further advance this field [26].

Yu have proposed the progressive improvement in medical intelligence (AI). AI technologies are now spreading to areas traditionally considered to be just the province where human experts are interested in digitized data acquisition, machine learning, and computer infrastructure. They described recent advances in AI technology and its biomedical applications and highlighted the barriers to further development of AI systems, and summarized the economic, juridical and social consequences of AI in healthcare.

Manreet proposed an Artificial Intelligence (AI) machine that has the capability of performing smart tasks, and machine learning (ML) is an AI subset that describes machines' ability to learn independently and to make precise predictions. The use of AI along with "big data" from electronic health records would affect patient care. An increasing body of literature in cardiovascular health, including mechanical circulatory support, was published in recent years using ML (MCS). In diagnosis, control, and therapeutics artificial intelligence (AI) and machine learning (ML). Although conventional studies use statistics to curate and turn data into interpretable, ML does so in a manner that can either be monitored or unmonitored. Traditional statistics are usually based on assumptions and are used to deduce samples or population parameters. Since the scientific community was organized and generally recognized, the usefulness of therapies was investigated using certain techniques. ML, however, is motivated by probability, with a prediction or classification algorithm for the data structure. The first article gives clinicians a summary of the related ML and AI concepts, reviews ML predictive modeling concepts, and gives contextual reference to how AI is adopted in the area of Machine Categorization System. Lastly, it discusses how these approaches may be integrated into the practices of medicine to increase patient outcomes [27].

It is often necessary to use a conceptual and organizational structure to improve the company and maintain its desired location. The intellect artificial becomes only a curiosity. It is therefore the corporate asset that needs the attention of any entrepreneur. In the business sector, there have been rapid changes and competition is increasing in the market as a result of globalization and technological changes. This enhances entrepreneurs' need for advanced technology to achieve a competitive advantage in the market. The studies show that the advent of the latest technology sequence has become an impetus for ever-growing business activity. Artificial intelligence in all of these innovations is one of the most important technological changes that have a major positive impact on the industry because it leads to the enhancement of operations, revenues, and development of companies on the market and decreases business failure chances. Xu also explained that in the lead generation of B2B, Artificial Intelligence splashes greatly into entrepreneurs' attention. For example (ML), the reach and size of AI applications have been expanded covering many areas and aspects of the abilities and skills of entrepreneurs. These innovations help technology enhance its cooperation with multiple stakeholders and help to create confidentiality, accountability, impartiality, and confidence. Companies that use artificial information will cut their call times to about 70 percent and increase the number of leads by

about 50%. The study also indicates that up to 85% of business activities were expected to be outsourced by 2024 to computers or robots. Many business leaders assume that artificial intelligence ensures high precision and productivity of working and also ensures low-cost, profitable businesses that support companies in achieving high profits and cost-benefit consumers that ensure high profitability in the sector. This is due to many entrepreneurs and business leaders. Businesses that use artificial intelligence tend to spend less time collecting and analyzing contacts because they do so automatically with the aid of computers. This helps companies to concentrate on the actual sale. This shows, therefore, that artificial intelligence automation can change entrepreneurship [28].

It also examined the use for domain changes to be tackled by multi-site medical imaging data using image and space transformer networks (ISTNs). Domain adaptation (DA) frequently takes place in the latent space with little regard to the explainability of the inter-domain transition. At the picture level, we use ISTNs that restrict transformations to explicit changes of aspect and form. As evidence of concept, we show that ISTNs can be trained adversely with simulated 2D data on a classification issue. For real data validation, two MRIs are constructed from the Cam-CAN and United Kingdom Biobank studies to explore domain changes due to population differences and acquisition. The result shows that age regression and sexual classification models trained on ISTN performance enhance generalization in data training on one side and on another.

System Analysis, Design and Result

Design & Implementation Methodology

Rapid Application Development (RAD) is the scientific methodology used for the design and implementation of this research. This research objectives were carefully considered, and RAD was chosen as the technique since it supports them. In order to harmonise the data, a Support Vector Machine (SVM) methodology was used. It looked at several methods of data gathering from current corporate databases, designed a system that can study the data and its structure from the existing system, and studied the data itself. The harmonised huge datasets may be viewed as a single, comprehensive source of truth, and when you have a single source of truth that is updated often or instantly, you don't have to waste time confirming, rehashing, and locating several sources of data. Rapid Application Development (RAD) is an agile software development method that places a priority on swiftly releasing and revising prototypes. RAD, as opposed to the Waterfall method, places more emphasis on user feedback and software usage than it does on meticulous planning and requirements documentation. A feature of RAD is quick project turnaround, which appeals to professionals working in a high-paced field like software development, particularly when the development requires integration into an existing system. The amount of time spent planning is decreased, while the amount of time spent creating prototypes is increased. In the planning of this task, RAD characteristics such as business modelling, data modelling, process modelling, application development and testing, and turnover are all rationally examined. The design paradigm mimics the RAD's methodology by swiftly analysing, designing, producing, showing, revising, testing, and implementing. This constitutes the prototype cycle, also referred to as the RAD model. The process-flow of rapid application development is depicted in Figure 2.

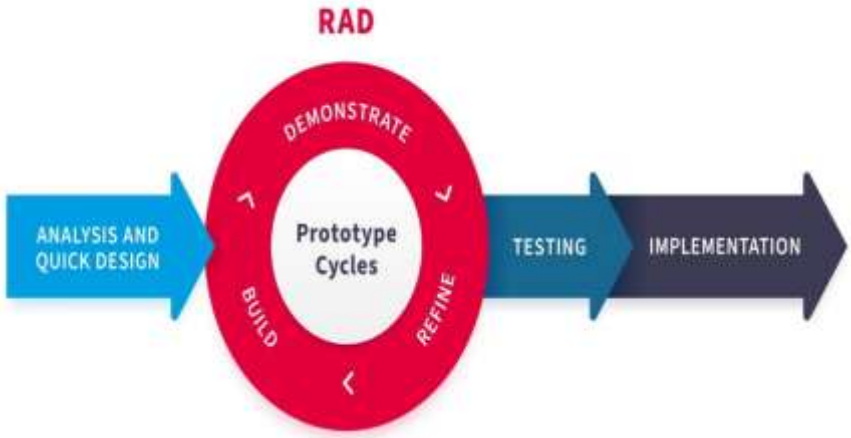


Figure 2. Rapid Application Development Process Cycle

Discrepant Data Harmonization Using Support Vector Machine Algorithm

Heterogeneous Data Harmonization System using Support Vector Machine Learning Approach (HDHS) is referred to as Discrepant Data Harmonization. The novel system in this study is known as the Supervised Data Harmonization System (SDHS). Information technology professionals are starting to understand the need to combine interdisciplinary data, metadata, and concepts from a range of relevant fields in order to solve complex and/or large-scale difficulties linked to big-data management. Information from several communities that use well-established but distinct terminologies, concepts, and strategies for naming and arranging their stuff must be combined in order to do this. The Data Harmonization System using Supervised Machine Learning Approach (SDHS) can represent, manage, harmonize, and integrate metadata and semantics for concept systems, databases, data elements, and value domains, that is, data types and sets of valid values. This includes data downloaded from portals and newly produced data.

Instead of a Data Production Platform, the new system incorporates a data upload and migration platform. Data Table (DT) is an object class that is used in the data upload interface. When data compatibility is established, data capture is specified to occur on the upload interface. For enterprise organisations who opted to import the data from office online, it studies/reads the data build, structures, and formats from Microsoft Excel Spreadsheet, XPS programme (XML Paper Specification), and Microsoft Office 365. The user's choice of either the upload and harmonisation stage or the viewer and download part will determine how the roles are defined. The potential for education sectors to exchange metadata models and terminologies could be enormous if data harmonisation tools like Data Harmonisation System using Supervised Machine Learning Approach (SDHS) are used to convey data semantics connected to terminologies. For instance, supervised machine learning is an aspect of artificial intelligence that demonstrates a computer's capacity to resolve a clearly specified issue by carrying out clearly defined operations on clearly defined data. Instead than relying on machines to read, comprehend, and process data and meta-data, it necessitates human effort by requiring people to monitor how AI applications read and process data. Humans must produce data in line with the rules that direct the actions of AI applications as a result of additional implications. The National University Commission (NUC), the Joint Admissions and Matriculations Board, the National Youth Service Corps, public and private universities, non-governmental or non-profit organisations, and others use the SDHS

to provide well-defined data descriptions for use by various cross-sectional dispensations, statistical purposes, and decision making.

Rapid Application Development Tools for Analyzing the SVM Algorithm

1. QuickBase

Quickbase is one of the few rapid application development solutions that does not necessitate any coding knowledge. Quickbase is known for the following features as shown in table 1:

2. Microsoft PowerApps

Microsoft PowerApps with Microsoft Azure is a popular choice among quick application development tools among both professional and novice developers throughout the world. It is mostly due to the enormous amount of customization that Microsoft PowerApps offers. Table 2 presents the features of Microsoft PowerApps.

3. Oracle Application Express

Oracle Application Express is primarily intended to provide you with rapid application development capabilities to help you monitor and analyze data in the organization. Table 3 shows the features of the Oracle Application Express that enable the Rapid Application Development System Methodology (Table 1,2,3).

Table 1. Rapid Application Development Analysis Tool using QuickBase

S/No	Features	Description	Remark
1	Text Editor	comes with a basic WYSIWYG type editor that lets you create even the most sophisticated apps.	Compliance
2	CMS Customization	Quick Base's user interface allows you to drag and drop essential parts into the program, while other features allow you to completely customize the app's interface.	Compliance
3	Mobile Compatibility	Web and mobile apps that can represent workflows and processes are possible to create	Less Compliance
4	Data/Domain Security	t has enterprise-level security and governance capabilities, as well as quick testing capabilities.	Compliance
5	Analytical Tools	Process Improvement and Human Resources are examples of use cases.	Compliance

Table 2. Rapid Application Development Analysis Tool using Microsoft PowerApps

S/No	Features	Description	Remark
1	Office 365 Connectors	Microsoft PowerApps is fully integrated with Office365 and your Microsoft Account	Compliance
2	Registered Domain Check	All users of Microsoft PowerApps have access to over 350 data sources, as well as the ability to create custom connectors	Compliance

3	SQl Enabled	In Microsoft PowerApps, you can enable all SQL tools and queries that you prefer to use	Compliance
4	Legacy Data Upgrade	It also enables you to use Robotic Process Automation to bring legacy systems to life	Compliance

Table 3. Rapid Application Development Analysis Tool using Oracle Application Express

S/No	Features	Deccription	Remark
1	Mobile Compatibility	Oracle Application Express's universal theme user interface is easy to use on both the web and mobile platforms.	Compliance
2	Real-time Customization	The apps you create for your employees are readily customizable and deployable in real time with no lag.	Compliance
3	SQL Security Authntication/ Authorization	It has various security features, including integration with your company user repository	Compliance
4		It also works with bespoke authentication and authorization schemes built in SQL or PL/SQL	Compliance
5.	URL https protection	Parameter tampering protection for URLs and cross-site scripting prevention are also included in Oracle Application Express	Compliance

Mathematical Model for Harmonizing Discrepant Data Using Support Vector Algorithm

Mathematical modeling is the process of converting issues from one application field into tractable mathematical formulations, which can then be analyzed theoretically and numerically to provide answers and guidance for the original application. Data is migrated from a database to another database for intending harmonization purpose. Eg. MI = No of Management Infrastructures, MA = Management Applications and DS = Total Volume of Dataset used for Training and Testing (T&T). MI = MA+DA. D represents the data dimensionality = total number of data-set used for (T&T). The ability of the Supervised Data Harmonization Systems to upload a particular data-size depends on how the data is organized in the Data Source Application, such as Microsoft Excel (.xls) WPS Office Tools etc. eg. 6, 293 with 13 dataset were used for training header column. If the Column size is divided by the data-size, it gives us the total execution time. The Mathematical Model used in this research is called Linear Modelling. Infrastructures used in this research includes – Servers, Computers, Data Science applications such .xls, wps Oozie, Yarn Java React etc When modeling the behavior of a device or system that is pressed or pushed by a complicated set of inputs or excitations, we use linearity. By adding or superposing the various responses of the system to each individual input, we may get the response of that device or system to the aggregate of the different inputs. The principle of superposition is the name given to this significant conclusion.

Engineers apply this technique to forecast a system's response to a complex input by decomposing or breaking it down into a series of simpler inputs that yield predictable system responses or behaviors.

Eg. Let;

NS = No of Management Infrastructures (MI) (1)

MA = Management Applications (2)

N = Cumulative Decision-support Systems (3)

= NS + MA/N

D = dimensionality = > 6293R/13c (4)

6R/13c = 6000/13 = D (*when dimensionality = total number of row divisible by total number of column)* (5)

D = 6000/13 (R/C) = 484.1 (6)

T = upload time per Sec. (T= 20mins X 13) (7)

T = 260m /60m = 3.33hrs (8)

Check Error;

Error Check Algorithm and Time Complexity = UT/QT = 20mins (9)

**Read data entries (a1), column table (a2), row(a3), table(a4), table headers(a5), terminologies(a6), check column num(a7), row num(a8), rewrite similar terms(a9), delete remainder words(a10), check relationships between tables(a11), components(a12)–subsets (Checks if data-elements complies with SQL constructs, word reasoning and Standards. Does data-elements have SQL constructs and query elements? Y (yes); else display short notification "Hello! Copy (XYZ) and move content to (ABC), then click continue.*

If T = 260m /60m = 3.33hrs, Harmonization marked S= Successful, else return error and reload (10)

T = 260m /60m = 0

System Design

The design of the systems is organized into five prototype component parts, in compliance with the requirements for the RAD model. The components are build, demonstrate, refine, test and implementation. The design tools are drawn business modeling tools, data modeling tools, process modeling tools, application generation tools and testing and turnover tools. These tools provide basis for the design of the system's architecture, input designs, process design, and output design as outlined below:

Step 1: Define and finalize the project specifications:

Stakeholders meet to identify and finalize project requirements such as project goals, expectations, timetables, and budget during this process. You can seek management approvals once you have thoroughly defined and scoped out each part of the project's requirements.

Step 2: Start Building Prototypes: As soon as the project scope is completed, clients will work closely with designers and developers to produce and improve functioning prototypes until the final product is ready.

Step 3: Collect user feedback: Working models are created from prototypes and beta systems in this step. Users' feedback is then used by developers to alter and improve prototypes in order to build the best possible product.

Step 4: Test, test and test to ensure compliance to outlined specifications

This process entails testing your software product to confirm that all of its moving elements function as expected by the client. As the code is tested and retested for smooth operation, continue to include customer feedback.

Step 5: System Demonstration

This is the last phase before the final product is released. It entails data conversion and user education.

Data Flow Diagram of the Harmonization Solution

The Supervised computer application model uses Microsoft Excel as in input text interface for data collation. There are five user interfaces in the front-end program, each with its own function: (I) an import interface; (II) a transform interface; (III) a master data dictionary interface; (IV) an integration interface; and (V) an export interface (Fig. 3). These interfaces assist the user in importing and harmonizing the original study's data dictionary with the master

data dictionary, as well as exporting raw data from all specified variables and studies of interest into a single harmonized dataset. Figure 3 describes the functions of the five SDHS user interfaces in further detail.

High-Level Development of the Harmonization Solution

The high level development of the New System is the development stages that must be integrated into the system, so that users can use the system for business solutions. Data Harmonization System using Supervised Machine Learning Approach (SDHS) conducts harmonization in compliance with data-science policies for management of enterprise data. The systems engine starts by conducting the following operations:

i. Extracting the Uploaded Data, Transform, Load
ii. Centralization and cleansing of the data
iii. Classify and Normalize

Figure 3 demonstrates the harmonization process of the Data Harmonization System using Supervised Machine Learning Approach (SDHS).

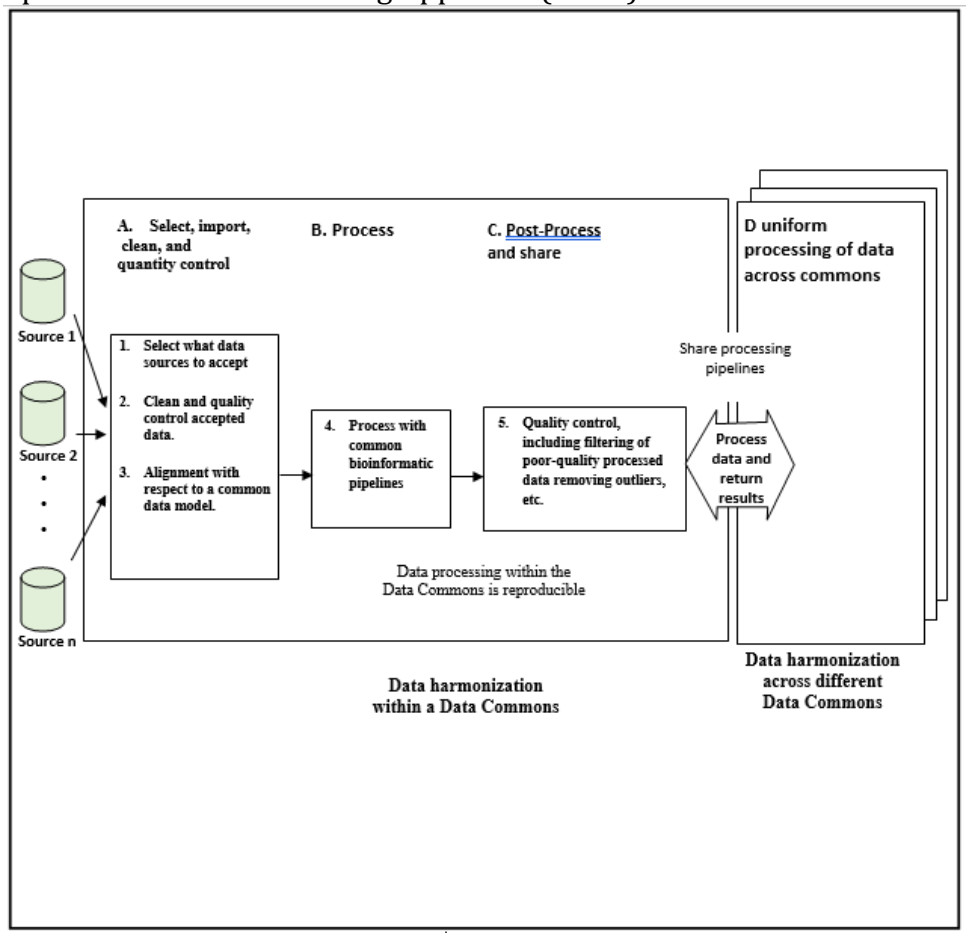
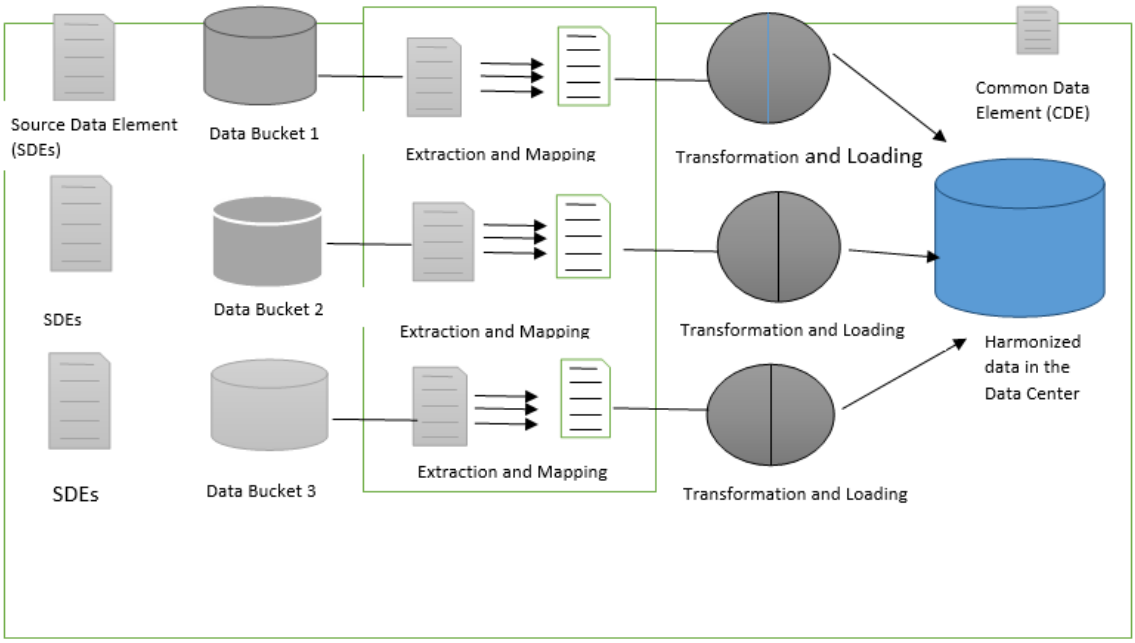


Figure 3. Harmonization process of the SDHS

Figure 4 shows the features of the data harmonization system. The features are the key



elements that supports the acting process of the Data Harmonization System.

Figure 4. Features of the Supervised Data Harmonization System

As shown in figure 3 and figure 4, the harmonization system starts with:

1. Extract, Transform, and Load:

The three database processes responsible for actually putting your data into a common database for harmonization are extract, transform, and load. Extraction pulls data from the source database, transformation transforms it into a format that can be queried and analyzed, and loading uploads the data to the destination database. Historically, this has been the most troublesome aspect of data integration, because a mistake in one phase leads to erroneous or missing data across the process. And each system has its own set of challenges that it can encounter. That's why choosing the correct technology for the ETL process is crucial, so you can spend less time micromanaging data migration and more time analyzing and making decisions.

2. Data Centralization and Cleansing:

After extraction, transformation, and loading, data centralization is the process of getting the uploaded data into a single centralized location. The harmonization system will finish its target phase before participating in application match with the target domains or platforms, even if the target domains are integrated through applications interfaces (APIs). Your platform will need a data model that blends your data together in order to harmonize it. Machine learning advances have made this business-friendly. The act of correcting or relocating faulty, broken, or erroneous data from your dataset is known as data cleansing. Consider this a makeover for your data. Most time, this can be done manually in a excel spreadsheet if it is the data production source of computer application.

3. Classification and Normalization of Data:

The terms "data normalization" and "data harmonization" are interchangeable. Both require the same treatment of the core element of the organization's data. Classifications are the names of the fields in your data, or, to put it another way, the titles at the top of table columns (excel online/excel applications. These are what help you segment your data, filter it, and drill in or zoom out (e.g., from 'Names' to 'Student's Names' to 'Candidates Names') from a business standpoint. Even when referring to the same topics, the naming conventions used to describe your data can (surprise) differ widely among your education tools and teams. For example, it might be called "Students Name" on the NYSC site, "Candidates Name" on JAMB, or "Name of Students" on the institutions portal for the same institutions. To address the classification,

harmonization is required to bring these together for the same objective. Harmonization takes it a step further by allowing you to integrate that metadata across the data content, such as through your website and Customer Relationship Management systems. Data imported from various domains may not always contain the classification required by the users. However, by harmonizing your data, you can add new categories or dimensions to your centralized data model, such as generating the region "Nigeria" from all of your Nigerian country classifications.

Advantages of the Harmonization Solution Using Support Vector Machine Algorithm

1. The supervised data harmonization system contain data and metadata scrubbing functionalities.
2. Discourages data reproduction to avoid duplicate records of Nigerian students (entities) in many domains.
3. Complies with the policy of Data Science and data science applications.
4. Provides built-in constructs to create or harmonize data.
5. Contains shared view of the data, relationships, and procedures that characterize “protocol-driven research and its accompanying regulatory artifacts” as a whole.
6. Support integrated systems in data management applications.
7. The SDHS is a powerful tool that can be used to create standard constructs for registries of multiple domains, metadata, and their inter-relationships, either through an additional external ontology or annotations, and can enable new capabilities that make such registries invaluable to their users, particularly for question answering, through a combination of open and closed world reasoning.
8. Maintains uniqueness of data across many platforms and domains.

2. Conclusion

1. The research is an advanced solution for data heterogeneity problem in the education.
2. It has repositioned harmonization or integration approaches for large and dispersed data, computational techniques that can effectively process and manage textual data, and performance evaluation of tools, techniques, and models. The expertise provided by this SDHS will be useful for academics and data analysts who are working with the large sorts of industrial text files and datasets, according to the results and discussion.
3. It has developed an interface for migration, representation, and visualization of data that are in various forms would benefit from the use of the most modern machine learning, deep learning, and SVM approaches, as was previously indicated.
4. The SDHS solution also support data cutting-edge applications like zero-shot network domain data, Industry Revolution 4.0, and IIoT.
5. Alignment of non-uniform domain to manage same data-set.
6. Encourages data interoperability among individual information systems in the supply chain.

References

[1] S. Bagui and P. C. Dhar, "Positive and negative association rule mining in Hadoop’s MapReduce Environment," *Journal of Big Data*, vol. 8, pp. 10–15, 2019. doi: 10.1186/s40537-019-0238-8.

[2] O. A. Bamiro, *Enhancing the quality of leadership and governance of Nigerian universities towards sustainable management and optimal performance*, Executive Development Programme for Council Members of Nigerian Universities, Abuja, 2016.

[3] L. B. Becnel et al., "BRIDG: A domain information model for translational and clinical protocol-driven research," *Journal of the American Medical Informatics Association*, vol. 24, no. 5, pp. 882–890, 2017. doi: 10.1093/jamia/ocx004.

[4] E. Bisong, *Google Collaboratory*, in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Apress, Berkeley, 2019, pp. 1–7. doi: 10.1007/978-1-4842-4470-8_7.

[5] J. H. Boyd, P. D. T., and S. S. Saint, "Technical challenges of providing record linkage services for research," *BMC Med. Inform. Decision Making*, vol. 23, 2014.
[6] M. E. Boza, "SDG Dashboards: The role of information tools in the implementation of the 2030 Agenda," Report of research collaboration between UNDP-SIGOB and the UNDP Bangkok-Hub, 2017.
[7] C. Bryan and J. K., "Visualization of Heterogeneous Data," *ResearchGate*, 2007. https://www.researchgate.net/publication/3411507_Visualization_of_Heterogeneous_Data.
[8] T. Carneiro et al., "Performance analysis of Google Collaboratory as a tool for accelerating deep learning applications," *IEEE Access*, pp. 77–85, 2018.
[9] B. K. Daniel, "Big Data and data science: A critical review of issues for educational research," *Britain Journal of Education Technology*, vol. 50, pp. 101–113, 2019.
[10] K. Dahdouh, A. Dakkak, L. Oughdir, and F. Messaoudi, "Big data for online learning systems," *Educational Information Technology*, vol. 23, pp. 2783–2800, 2018.
[11] L. Ding, Z. Fan, and D. Chen, "Auto-categorization of Harmonization System Code using background net approach," *Procedia Computer Science*, vol. 11, pp. 1462–1471, 2015.
[12] D. Doiron, P. Burton, and Y. Marcon, "Data harmonization and federated analysis of population-based studies: The BioSHaRE project," *Emerging Themes Epidemiology*, vol. 10, no. 12, 2013. doi: 10.1186/1742-7622-10-12.
[13] D. Doiron et al., "Facilitating collaborative research: Implementing a platform supporting data harmonization and pooling," *Norsk Epidemiologip*, vol. 21, pp. 221–224, 2012.
[14] A. Dutta, T. Deb, and S. Pathak, "Automated Data Harmonization (ADH) using Artificial Intelligence (AI)," *OPSEARCH*, 2020. doi: 10.1007/s12597-020-00467-4.
[15] E. F., M. S., L. S., and Y. Z., "Semantic Web Enabled Software Engineering," 8th International Semantic Web Conference, Virginia, USA, 2009, pp. 25–29.
[16] E. O., S. E., V. E., and E. O., "Web mining: Cybermetrics analysis of the nine (9) newly established federal universities in Nigeria in 2011," *International Advanced Research in Computer Science and Software Engineering*, vol. 5, pp. 904–913, 2015.
[17] T. Fera and W. A., "Next IT Challenge: From Data Acquisition to Harmonized Information Management," *Journal of AHIMA*, pp. 42–44, 2010.
[18] H. Gatner, "Data Modelling – Understanding Tools and Techniques Involved," *Gartner-Global Research and Advisory Firm*, https://www.xenonstack.com.
[19] S. Gomatam, A. F. Karr, J. P. Reiter, and A. T. Sanil, "Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers," *Journal of Statistical Science*, vol. 50, pp. 163–177, 2005.
[20] B. E. Haarbrandt, M. Tute, and S. Marschollek, "Automated population of an i2b2 clinical data warehouse from an open EHR-based data repository," *Journal of Biomedical Information*, vol. 59, pp. 277–281, 2016.
[21] S. Hadi and R. B., "Emerging artificial intelligence methods in structural engineering," *Journal of Science Direct- Engineering Structures*, vol. 12, pp. 170–189, 2018.
[22] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Burlington: Morgan Kaufmann Publishers, 2006, pp. 1–42.
[23] "Health Information Exchange for Continuity of Maternal and Neonatal Care Supporting," *Journal of Applied Clinical Information*, vol. 8, pp. 1082–1094, 2017.
[24] K. Hines, "Facebook reporting tools for in-depth analysis of fan pages," *Postplanner*, https://www.postplanner.com/6-facebook-reporting-tools-in-depth-analysis.
[25] I. M., "IFRS application and the comparability of financial statements," *Journal of Account Finance*, vol. 7–8, 2017.
[26] "Information Science," "New Aspects on Using Artificial Intelligence to Shape the Future of Entrepreneurs," http://dxdoi.org/10.18576/isl/090106.
[27] W. H. Inmon and R. Hackathorn, *Using the Data Warehouse*, New York: Wiley, 1994.
[28] ISO/IEC 11179, *Metadata Registries - Part 3 (Edition 3)*, http://metadata-stds.org/11179/index.html#A3.