Heterogeneous Data Integration Techniques and E-Health Paradigm: A Review

[1] VitalisNdume, [2] Yaw Nkansah-Gyekye, [3] JesukKo

[1, 2] School of Computational, Communication Science and Engineering,
Nelson Mandela-African Institute of Science and Technology, Arusha, Tanzania

[3] Department of Healthcare Management, Gwangju University
Gwangju, Korea

Abstract - Even though data integration has been there for long, integrating e-health data remains an open research challenge. Many studies have discussed data integration approaches in isolation, causing a disjoint of information flows and narrowing the understanding of alternative solutions in the cause of choosing an appropriate approach for integrating e-Health data. This problem can be solved by undertaking a comprehensive review on different approaches and techniques in data integration as well as paradigm in e-Health frameworks. The main contribution of this paper is a narrative description of most widely used integration architectures as well as comparative analysis of each of them. The review reveals that despite the business needs of integration approach, each has limiting factors based on the data size, unstructured or structured nature of data set as well as business agility of the organization. Moreover, it is found that some technologies were not designed for data integration but for document management or information integration. It is concluded that data integration approaches and the field in which it applies mostly, is widely spread that always results in integration problem. Also, it is noted that designing a framework for integrating e-Health data using web service approach in a distributed network characterized with short lived connection is still a challenge.

Keywords - *Heterogeneous data, distributed architecture, integration techniques, e-Health paradigm*

1. Introduction

Even though data integration has been there for long, integrating e-Health data remain an open research challenge. Many studies have discussed data integration approaches in isolation, causing a disjoint of information flows and narrowing the understanding of alternatives solutions in the cause of choosing an appropriate approach for integrating e-Health data. integration has been there since long, almost more than three decades[1]. The first integration application was MULTIBASE developed in 1980. Federated architecture was also among the oldest approaches in data integration[2]. Much literature in data integration framework has potential limitation in the selective literature review regarding integration frameworks, approaches and techniques in relation to e-Health. Data integration has been applied in many fields; however, there is limited application in the health sector. The topic of heterogeneous database integration is difficult but still important in bioinformatics[3].There are many reasons for the limitation including, inherent complexity of the domain, semantically definition in health data, conceptualization or definition of terms, unstructured nature of health data and infrastructure that support integration architecture. Even though data integration has been there for long, integration in health system remains an open research problem. It is noted in [4] that despite more research that has been done in integrating e-Health, no prime solution has been achieved. Research in DNA and its protein sequencing result in new data sources that complicate integration calling for more research that provides a solution that yields coherent, shared information for each individual need. Data standards and interoperability in e-Health system count for the complexity for health data integration.

The challenge to heterogeneous distributed data is interoperability of data. The interoperability of data at the system level is defined in [5]as the ability of two or more information systems to exchange both computer interoperable data and human interoperable information and knowledge. It is argued in [6] that the interoperability does not require two systems be identical in design or implementation rather that they can exchange information and use the information they exchange and that the information being exchanged is conceptually equivalent. Lack of semantic interoperability at the layer of information is a problem in achieving a true interoperable e-Health system [5]. Practically, it means that the precise meaning of exchanged information from one system must be understood by any other system or application; as if the information was created within the first system[7]. In that lines the data standards in e-health is necessary to be defined. Data standards are described as the absence of common frame of references or common business terms definition. Lack of agreed format for exchange makes it difficult for parties to understand each other, and this is extremely true when different businesses need to share information. Therefore, data standard is basically an agreement between parties on the definition of common business terms, the way those terms are named and represented in data and as set of rules that may describe how data is stored, exchanged, formatted or presented [8]. Nevertheless, standards describe the policies and procedures for defining rules and reaching agreement about standardized data elements.

The goal of this review paper is twofold: first is to describe various approaches and techniques in data integration gearing to e-Health frameworks. Second, is to describe theoretical and practical approaches in data integration with focus on e-Health paradigm. The paper adds the following to existing knowledge: first clear description of the approaches of data integration in the knowledge of computer science and constraints of each approach in different environment of application, and second, in the context of e-Health paradigm an existing gap in the cause of integrating e-Health in a distributed network with a short lived connection is identified.

Organization: This paper is organized as follows. Section 2 explains the approach used to select literature. Section 3 provides the knowledge on two major concepts of in data integration, which are ontological and semantic techniques. Section 4 discusses theoretical approach in data integration layers. Section 5 discusses approaches in data integration and provides descriptions of the major operations in the integration layers and the wrappers used to connect the layers. It also gives comparative analysis of the integration architecture. Section 6 explains several techniques in data integration. Section 7 provides paradigm in e-Health data integration frameworks, and finally conclusion is provided in Section 8.

2. Methodology

A comprehensive review of peer reviewed and grey literature was undertaken to identify the previous and current state-of-the-art regarding the alternative choices of data integration. Information was searched from libraries, platform of online journals, and conference proceedings, as well as web sites from grey sources. The inclusion criteria for literature were those papers with technical and theoretical approaches of data integration. The literature reviewed was from the year 1998 to 2013. The reason for including papers aged above five years old was to understand the historical challenges on data integration in the last three decades. A total of 367 papers were collected and only 85 met the criteria and were selected for review.

3. Conceptualization and Meaning of Concepts in Data Integration

Data integration is the study that is concerned with combining data residing at different sources, and providing users with a unified view of the data[9-11].It is focused on a controlled sharing of data and business process between any connected application and data sources[12]. Data integration has two major concepts: ontology which means integration by conceptualization and semantic which means integration by meaning. These concepts are common in biological research and practice where multiple groups from diverse disciplines collaborate or store experimental data in a local, autonomous database with different schema. It is argued by [3] that the complex problem in life science research has given rise to multidisciplinary collaboration, and hence, to the need for biological health database integration. It is depicted in the literature [13] that biological data exhibit a wide variety of technical, syntactical and semantic heterogeneity. In most cases, biological data integration is by conceptualization.

3.1. Ontological Approach

The term ontology has long been used in many ways and domains. As referenced in the literature [11], in computer science world ontology was introduced by Gruba (1992) as an explicate specification of a conceptualization. A conceptualization refers to the abstraction model of how people think about a real thing in the world. Ontology gives the name and description of the entities of specific domain using predicates that present relationship between these entities. It provides a vocabulary to present and communicate knowledge about the domain and set of relationship containing the term of vocabulary at a conceptual level. Integrating data in the health fields, especially in computational biology, requires techniques of transforming data syntactically and semantically to bridge the gaps across data sources. In addressing the problem of semantic heterogeneity, ontological approach is mostly used in data integration[14].

Nevertheless, ontology in biological science is hard to solve; however, ontology provides a rich predefined vocabularies that serves as a stable conceptual interface to the database and it is independent of the database schema[11]. Some researchers attempt to describe ontology in the form of glossaries rather than making a deeper analysis of the concept[11, 15-18].These attempts mainly contain terms and their primary definitions, and therefore lack expert conceptualization. This argument is supported by[19] that armored issues concerned with ontology description requires expert to provide knowledge and mechanisms for validation.

Ontology, from the respective field of research, needs to be harmonized to reduce the conceptual and terminologies confusion and to allow a common and shared understanding in order to improve communication, data sharing, interoperability and reusability[20]. However, it is argued by[15]that even the most expressive ontology specification language is not sufficient for information integration in the semantic web. This is because in a real world setting, different ontology is built by different organizations for difference purposes. Hence, one should expect the same information to be represented in different forms and with different level of abstraction in various ontological approaches. Nevertheless, it is suggested in the literature[11] that despite the challenges in ontology descriptions, the concept might be used for data integration tasks because of its potential to describe the semantic of information sources and to solve heterogeneity problem.

3.2. Semantic Method

Semantic heterogeneity is a general term referring to disagreement about meaning, interpretation or intended use of the same or related data[21]. It is argued by[22] that semantic heterogeneity is one of the key challenges in integrating and sharing data across disparate sources, data exchange, and migration, data warehousing, model management, semantic web and peer to peer databases. The reason for studying semantic information system is to ensure that the interpretation of information is well understood by the user in collaboration [23]. Different data mode provides different structure primitive [24]; this is because when databases are designed and developed independently, it is common to have information in one database inconsistent with information in another. Therefore, structural semantic heterogeneity is experienced when two different data models are connected for the purpose of data sharing. The structure heterogeneity means that different information system models store their data in different structure[21].

Despite its pervasiveness and importance, semantic integration remains an extremely difficult problem[25]. Sources of semantic interoperability arise due to different formalization of terms, conceptualization differences and difference in perceived context. This problem is elaborated by [26]as follows: first, the semantic of the elements can be inferred from only a few information sources, typically the creator of the data, documentation and associated schema and data. Second, schema and data clues are also often incomplete. Third, it is difficult to design a system that is matched globally. Finally, schema matching is often subjective, depending on the application.

Matching techniques of semantic integration can be categorized into two groups: Rule and learning based solution as discussed in the literature [26, 27]. Rule based solution employs handcraft rules that exploit schema information such as element names, data types, structure numbers of sub elements, and integrity constraints. On the other hand, the key idea in learning based solution is that a matching tool must be able to learn from the past matches, to predict successfully matches for subsequent, unseen matching scenarios. It is suggested by [26] that the complementary nature of rule and learning-based techniques suggest an effective matching solution that should employ both each on the type of information it can effectively exploit. The semantic and ontology data integration are considered as search engines and are in third generation techniques in data integration.

4. Theoretical Approach in Data Integration

In practice, concrete integration solution is realized based on six integration approaches, including: manual integration, common interface, integration by application, integration by middleware, uniform data access and common data storage [1].Integration of heterogeneous database system is computational model and software implementation that provides a single uniform query interface to data that are stored and managed in multiple heterogeneous data sources [12, 28, 29]. Integration of multiple information systems aims at combining selected systems so that they form a unified new whole and give users the illusion of interacting with one single information system ([1, 9, 30]. The key issue behind data integration is transparency, which means abstraction from secondary feature of distributed resource is combined to give a single view. It is observed in [31]that there are many forms of transparency, in particular location, access, concurrency, implementation, scaling, fragmentation, replication, indexing and failure transparency. All aim at supporting flexibility and maintainability of software products. Thus, theoretically [9, 32, 33] a *data integration system* I is defined in terms of a triple *{* G, S, M *}*
where,

- G is the *global schema*, expressed in a language L_G over an alphabet A_G. The alphabet comprises of a symbol for each element of G (i.e., relation if G is relational-oriented, class if G is object-oriented, etc.).
- S is the *source schema*, expressed in a language L_S over an alphabet A_S. The alphabet A_S includes a symbol for each element of the sources.
- M is the *mapping* between G and S , constituted by a set of *assertions* of the forms

$$qS \rightarrow qG$$

$$q_G \rightarrow q_S$$

Where

q_S and q_G are two queries of the same respectively over the source schema S , and over the global schema G . Queries q_G are expressed in a query language $L_{M,S}$ over the alphabet A_S , and queries G are expressed in a query language $L_{M,G}$ over the alphabet A_G. Intuitively, an assertion $q_S \rightarrow q_G$ specifies that the concept represented by the query q_S over the sources corresponds to the concept in the global schema represented by the query q_G (similarly for an assertion of type $q_G \rightarrow q_S$).

Queries I are imposed in terms of the global schema G , and are expressed in a query language L_Q over the alphabet A_G. A query is intended to provide the specification of which data to extract from the virtual database represented by the integration system.

5. Frameworks for Integrating Heterogeneous Data

Different architectures have been suggested by many authors. However, almost all have the basic concept which relates to RDBMS architecture or OODBMS[14]. Theoretically, integration architecture has three layers: presentation layer, mapping layer and data source layer. Three basic operations on these layers are *Extract, Transform and load*(ETL) as discussed in the literatures[34-36]. The *extract ()* function is responsible to request data from the source system. It deals with semantic heterogeneity and therefore it is a challenging aspect of the ETL. An intrinsic part of extract involves the parsing of extracted data, resulting in a check if the data meet an expected pattern or structure. If not the data may be rejected entirely. The *Transform ()* is a series of rules or function responsible for gathering data into a format understood by the user. It deals with syntactic heterogeneity. And the last operation is *Load().*This function loads the data into the end target, usually the data warehouse. Depending on the requirement the process may overwrite, update or add data to the destination.

Indeed, accessing data from external data source mediator may contain one or several wrappers which process data from different kinds of external data sources. An example of wrapper is Open Database Connectivity (ODBC), Java Database Connectivity (JDBC), CAD system, ADDO.NET (set of classes that exposes data access service to the net programming). Other interfaces such as DOM, SAX, STAX, and JAXB interfaces are popular when dealing with XML mapping. The function of wrapper in the physical layer is to provide an interface to data sources. Wrapper logically converts the underlying data objects to common information model [37]. However, things get even more difficult when dealing with different data formats such as Electronic Data Interface (EDI) or Flat files [38, 39]. As described in the literature [1], the data integration approach falls into six categories: manual integration, common interface, integration by application, integration by middleware, uniform data access or common data storage. These categories can be implemented by any of the integration architectures as discussed below:

5.1. Federation System Architecture

The Federation Database Management System (FDBMS) is the first generation strategy of integrating data primarily focusing on issues of interoperability and schema integration. The federated architecture provides a solution for integrated data coming from heterogeneous database via a computer network[2]. The emphasis is on federation approach which is on system and data management, as opposed to information or knowledge management[40]. All federated schema are defined and controlled by the federal DBA. Each schema is virtual in the sense that there does not exist a physical database that corresponds to it; rather, a specification is provided that describes how the federal schema constructs are materialized from data maintained by individual components[24, 41, 42].In a federated system the query is submitted through a client to the federated server. The federated system cooperates with wrappers to develop execution plan. In the execution plan a query is decomposed into fragments for individual sources. The optimizer chooses the best plan on the basis of minimum estimated resource consumptions and the wrappers execute the fragments assigned to them. The result stream of data is set to the federal server where they are combined and results from wrappers perform any additional processing[43].However, the strategy provides limited scalability.

5.2. Mediated Systems

The mediator approach, originally proposed in 1992, has been used for integrating heterogeneous data in several projects. It is second generation data integration after federated framework[40]. Most mediator systems integrate data through a central mediator server accessed on or several sources through a number of wrappers interface that translate data to a common data model. However, the original goal of mediator is a distributed software module that transparently encodes domain-specific knowledge about data and share abstraction of that data with higher layers of mediators or applications [39]. It is noted in the literature[44] that mediation architecture also provides an effective promising module that promotes the integration of hospital information system that are autonomous, heterogeneous, and semantically interoperable and independent platform. The mediation manages a unified query interface over a set of distributed heterogeneous and autonomous data sources and plan for execution of the fragment which is assigned to them [43]. The TSIMMIS project (1994) specifies the architecture for mediation and claims that mediator embeds the knowledge that is necessary for processing specific types of information. The mediator may also process answers before forwarding them to the user. The label of the given data value may also be provided in the mediation process. The challenge to such mediation is due to heterogeneity nature which is caused by syntactic, schematic or semantic.

5.3. Data Warehouse

Data warehouse is defined by[45] as a consolidated integrated view of cooperate data drawn from disparate operational data sources and a range of end to end user access tools capable of supporting simple to highly complex queries. The goal of data warehouse is to support decision making. The framework of data warehouse composes of operational data, staging data, low summarized data, historical and highly summarized data [46].Data warehouse framework is a collection of decision support technology, aiming at enabling the knowledge workers to make better and faster decision[47-49]. It is noted in[49] that data warehouse aims to make an organization information easily accessible; consistent, that is data must be adaptive and resilience to change. The information generated from a warehouse should be capable of improving decision making and also data from a warehouse must be accepted by community. However, the warehouse suffered from data update, duplication of function in warehouse instead of operation database, higher maintenance cost, cleaning and manipulation of data stream. A research by [50] reveals that in order tofacilitate complex analysis and visualization, the data in data warehouse is typically modeled in multi-dimensions requiring the capability of viewing the data from a variety of perspective. The dimension of data can be subject oriented, integrated, time variant and detailed, summarized or highly summarized. The operational system might have overlapping and some times contradicting definition such as data type. It is argued by [51] that Entity Relation Model is not suited for multidimensional conceptual modeling in the data warehouse because they cannot represent adequately the semantic data in data warehouse. It is further claimed that the level in data warehouse can roll up to any number of levels thus forming multiple hierarchies on a single dimension. This can occur if different criteria of classification which are possible to dimension member of decision makers.

5.4. Work Flow Management Systems

Work flow can be seen as collection of tasks that are processed in distributed resources in a well-defined order to accomplish a specific task[52]. Workflow management techniques have been developed over 20 years especially in business management and office automation and production management. Work flow management system (WFMS) allows implementing business processes where each single step is executed by a different application or user. Generally, WFMS supports modeling, execution, and maintenance of business processes that are comprised of interactions between applications and human user approach[53].WFMS is a piece of software that provides an infrastructure to set up, execute and monitor scientific workflow[54]. Even though described as an approach in data integration but workflow management focuses on process rather than a document. It is a term used to describe the task of procedural steps, organization or people involved or requiring input and output information task needed for each step in a business process.

5.5. Peer to Peer (P2P)

In a P2P data integration, each peer has its own local data store (LDS) managed by the local management system[55].The essence of P2P is that nodes in the network directly exploit the resources present at other nodes of the network without intervention of any centralized server[56]. The P2P system has a query interface for accepting query and return answers during interaction within other peers. An important component of data integration in P2P system is distributed query manager (DQM). The DQM is responsible for planning execution of received query using P2P's own LDS and propagating queries to its partners. The partial results are merged and returned to the enquiring user. The metadata necessary to understand the query and to plan its execution are managed by metadata manager. Information about partners as well as rules defining integration strategy and conciliation action is managed by semantic integration manager. According to [57] in a P2P data integration system (P2PDIS), each peer is an autonomous information system providing parts of the overall information available from a distributed environment, and acts both as client and as server. Information integration in P2P does not rely on a single global view, as in federated data integration instead its achieved by the establishing mapping between peers, and by exploiting such mapping to collect and merge data from the various peers when answering user query.

5.6. Web Service Technology for Data Integration

Web service is a Graphical User Interface less web application. The term web service describes a standardized way of integrating web based application using the extensive markup language extensible marks up language (XML), Simple object access protocol (SOAP), Web service description language (WSDL) and Universal Description, Discovery and Integration (UDDI), and open standards over the internet protocol backbone that is HTTP[58, 59]. The XML is used to tag the data, SOAP is used to transfer the data, WSDL is used for describing the service available and UDDI is used for listening what services are available[60].There are basically three major roles with the web service architecture. These are service provider, service requester, and service register. Even though there are more advantages of using web service there are some challenges. The HTTP and HTTPS are stateless, i.e., the interaction between the server and client is typically brief and when there is no data being exchanged the server and client have no knowledge of each other [61]. In spite of the difficulties, web service and other technologies for Service Oriented Architecture (SOA) promise a future in which businesses are able to discover each other, exchange electronic documents and format, and conduct transactions with or without prior agreements [6].

It is observed that Web service and standards technologies such as XML and SOAP are essential for the development and deployment of interoperable heterogeneous health systems [58, 62]. Web service is loosely coupled and cross-platform with ability to integrate distributed application [63-65]. Use of open standards technology ensures that risk is minimized when developing new technology, prevents vender lock-in, enables re-use of solution, and eliminate costs in custom development and integration. In addition, web service architecture addresses the requirements of loosely coupled standards-based, and protocol independent distributed computing, mapping enterprise information system appropriately to the overall business process flows [38, 66]. However, it is claimed in [67] that it takes more than just standards to achieve interoperability at system level. It is also argued by [68] that other important aspects that contribute to smoothness of data integration includes intelligence data filtering at the data collection point, backward compatibility, leadership, political, organization legal aspect, psychological and commercial or business issues as well as emerging technologies. Despite many technical standard challenges, web services have become a key technology for bioinformatics, since life science databases are globally decentralized and the exponential increase in the amount of available data demands for efficient systems without the need to transfer entire databases for every step of analysis is important [69, 70].

With tools that add XML support to a database such as Oracles XML DB, some of the modeling effort is simplified, but the indexing strategy still requires lengthy planning to achieve usable, scalable system. However, achieving modeling of multiple formats using Entity Attribute Value (EAV) is also claimed to be unsuitable for direct analytical processing [71]. Nevertheless, web service model is based on open standards and it is expected to allow biological data to be fully exploited using relation model [72]. Standardization is a base for integrating heterogeneous data set. Standard is a free implementable specification developed by consensus among the important stakeholders in some domain working in a framework [6]. In order for a company to participate in a relation of business they need to follow loosely coupled standards such as a well-definedprotocol, web services or use of XML for data sharing.

5.7. Linked Servers Architecture for Heterogeneous Data Integration

A linked sever is a recent approach in data integration mostly applied in Microsoft applications. It is a technique of integrating heterogeneous data by using object linking and embedded database (OLEDB) technology. The OLEDB is a Microsoft application program interface (API) for retrieving data from a wide variety of sources. The significant advantage of OLEDB is that it can link non database files allowing users to run queries on a remote distributed server. Two approaches are used to create linked server: one is ad hoc which creates linked server for temporary execution of query and another is permanent approach which allows transaction SQL statement to run over and over [73]. In SQL 2008 R2 the technology of linked server has grown to link open source databases like MYSQL and DB2. This approach is categorized as uniform data access. The drawback of linked server is that it experiences slow query processing and may cause memory leakage. However, in the environment where data is updated occasionally; that is, it does not require frequently query processing then it is a good choice. Another drawback is that the linked server works mostly with OLEDB provider. Currently SQL server provides multi-hop linked server.

5.8. Grid on Data Integration

Grid is a natural evolution of the web. It is an organized connection of nodes over the network which contributes various resources. Grid enables the virtualization of distributed, heterogeneous resources using web service[74]. Unlike the web which is client–server oriented, the grid is demand oriented; user sends request to the grid which allocates them to the most appropriate resource to handle them. The main solution for grid data management is the context of computational .Grid is file based processing. A basic solution is that grid combines global directory service to locate the file and secures file transfer protocol.

Table 1: Comparative analysis of the integration architecture

SN	Framework/Architecture	Advantage	Disadvantage
i	Data warehouse	Offers fast query processing, clean data due to storage level, Supports both database and liked file, provides consistency data, Common data model regardless of sources, improved query processing	Complex schema, out of date data, duplication of function in operational model, High maintenance cost, Cleaning and manipulation of data stream
ii	Federation database	Offers current data, flexible architecture, no copying of data, supports database approach, explicitly autonomous DB	Slow query processing, complex schema due to different designers, dirty data, limited scalability, depends much on centralized mediator / mapping
ii	Mediated schema	current data, flexible architecture, schemas tailor to user, support data base approach, makes use of extensive metadata, It can combine RDBMS and OODBMS	Slow query processing, dirty data, mapping of local to global schema is needed, not best for persistently changing environment, does not support much dynamic changes, Less security due to uses of wrappers
iv	Peer to peer	Current data flexible architecture, Schemas tailored to user, Best for linked files, cost sharing reduction, improved scalability/reliability, resource ,aggregation and operability, increased autonomy , dynamic, anonymity/security and ad–hoc communication and collaboration , may not require centralized mapping/ mediator, improved quarry routing	Slow query processing Unclean data is possible when Peer grows, experimental architecture, does not take care of semantic data exchange, Data placement/location is hard to predict It is locally designated; centralized management rather depends on the voluntary contribution of the resources by users. inconsistency of the sources due to automaticity
v	Linked Server	Supports variant database source, easy to configure, easy to put two DBMS on single SQL statement. Supports Remote query, Implementable in low network environment - VPN network Flexible architecture Less maintenance cost Supports remote queries	Hard hoc query is slow, Joining different DBMS cause slow execution, Memory leakage due to data size, type mismatching
vi	Web Service	Gets real data in a single view, Data sources can be displaced but unification is possible at any up time	Down time of services depends from on network reliability. Short lived connection hence requires programmer involved to define size of messages packets, SOAP is built on top of HTTP which actually was also build on TCP IP and the same weakness of TCP is experienced
vii	Grid data integration	Uses Middleware approach, New technology in information integration	Was meant for scientific community (Physician and astronomy), The grid focuses on file directly allocation. The initial design purpose was for physical science and astronomy

6. Techniques in Data Integration

There are many technologies that are being used in facilitating data integration including flat file, data dictionary, Programming language classes, database schemas, UML Models, Spreadsheet templates and XML schemas and many others[6].

6.1. Flat File Integration

The easiest way to transfer data between the systems is to use flat files. The SQL server integration services (SSIS) uses flat files that are fixed width, delimited and ragged right format types for integrating data[75].

Fixed width: Fixed width format is mostly applied in mainframe or legacy system. Fixed width files use different width for column, but the chosen width per column stay fixed for all the rows, regardless of the contents of those columns. The fixed width format has disadvantage of having much more blank space between the columns. As most of the data in a column with variable data tends to be smaller than the width provided, it experiences a lot of wasted space. As a result, these types of files are more likely to be large in size than the other format.

Delimited: the most common format used by most of the system to exchange data with foreign system, delimited files separate the columns using a delimiter such as a comma or tab and typically use a character combination (for example, a combination of carriage return plus linefeed characters {CR}{LF} to delimit rows/records. These techniques are commonly known as CSV. Generally, importing data using this format is quite easy, unless the delimited used also appears in the data. For example, if users are allowed to enter data in a field, some users may use a comma while entering notes in the specific field, but this comma will be treated as column while entering notes in the specified fields, and will distort the whole row format. This free format data entry conflicts with the delimiter and imports data in the wrong column. If not careful in using delimiter it may degrade data quality. Because of the potential conflict, one needs to pay particular attention to the quality of data he is dealing with while choosing a delimiter. This approach is preferred to small size.

Ragged right: in some cases if one have fixed width file and one of the column (the rightmost one) is a non-uniform column, if saving some space is needed then one can add a delimiter such as {CR}{LF} at the end of the row and makes it a ragged right file. Ragged right files are similar to fixed width files, the last columns are of variable size. This makes the files easier to work with when displaying in notepad or imported into an application. Also, some vendors use these types of format when they want the flexibility to change the number of column in the files. In such situation, they keep the entire regular column (The column that always exists) in the first part of the file and the columns that may or may not exist combined as a single string of data in the row. Depending upon the columns upon the columns that have been included the length of the last column will vary. The application generally uses substring logic to separate out the column from the variable length combined column.

6.2. XML in Data Integration

The eXtensible Markup Language (XML) is a markup language that defines a set of rules for encoding document in a format that is both human readable and machine readable. The design goal of the xml emphasizes simplicity, generality and usability over the internet. It is a textual data format with strong support via Unicode for the language of the world. Despite of focusing on documents it is widely used for representation of arbitrary data structure in web service. The reason behind is about interoperability of data that are exported in xml format. It is argued in [6] that to exchange document or business application requires a precise and unambiguous language for describing information model. The XML has the following advantage that makes it suitable for data integration. It enables robust application to be deployed efficiently and at a reasonable cost, that is, XML contents can be taken from document, database and enterprise application, combined and treated as single source, and delivered to multiple users, devices or applications. Conversely, a number of issues may arise when working with complex the XML schemas.

For example, Oracle XML Developer Kits (XDK) parse the XML document outside the Oracle database and store the XML data on rows in one of more tables in the database. In this sense, the Oracle database is unaware that is managing xml content but object schema. Also, oracle store the XML document as character large object–CLOB, Binary Large Object BloBle, Binary file–Bfile or VARCHAR column, treating these as flat files. An error ORA-0172 maximum number of column in a table may also be encountered when registering an XML schema or creating table based on a global elements defined by the XMLschema. Nevertheless, using XML to encode implementation model yields an overall reusability and programmability unmatched by other representations. Furthermore, XML facilitates document encoding exchange architecture of the internet and the XML schema can be used to generate any of the other format if requires.

6.3. Spreadsheet

A spreadsheet is an interactive computer application program for an organization and analysis of data in a tabular form. The program operates data representation as a cell of arrays, organized in rows and columns. Each cell of the array is a model–view –controller element that can contain either numeric or text data or the resulting formulae that automatically calculates and displays values based on the content of the cells. Changes can be made in any stored value and observe the effect in calculated values. Each row in the spreadsheet can be mapped with a fairly different application allowing the integration of the systems. These instances of a spreadsheet can be associated or mapped with different records of a database or xml, wherein each database records has record field corresponding respectively to data entry cells of the form[76].These make spreadsheet valuable in data integration. However, the drawback of spreadsheet is that their simplicity often results in data tables that do not follow the best practice of database design such as attention to keys and normalization late alone the richer features enabled by knowledge bases[77].

6.4. Data Table/Schema Matching

The data table or relation schema is based on the concept of the data set which actually is individual tables that make a single database. Data table therefore is more advanced than spreadsheet. It contains data rows and data columns in which row are uniquely identified in the schema and it is rich of data types [78]. Using ID in the records of the tables allow schema to be normalized and be related and hence increased update, deletion as well as addition operation in T-SQL. The format of table's schema can be output to XML, flat file, CSV or spreadsheet. Schema matching has been used in data integration[79]because data table responds positively to data serialization between client and server in distributed network. The objective of serialization is to find non-serial schedule allowing transaction to execute concurrently without inter leave[80]. Data serialization format used in schema matching is more preferred in a higher structure data or relatively unstructured contents than XML.

7. Paradigm in e-Health Data Integration Framework

Integrating and sharing health research data is very crucial because they hasten the analysis of the prevalence, incidence and risk factor of diseases which are crucial to understanding and treating them[81].The application of information and communication technology on the whole range of e-health activities can simplify the access to health care services, boost quality and effectiveness of health management. E-Health tools allow the construction of patients-centric healthcare services. It enables providers to support patients to access health related information to prevent their possible diseases and monitor their health status. Many integration framework and standards has been developed. The three famous ones are explained in brief as follows:

Common Object Request Broker (CORBA) is a middleware standard architecture that provides some integrative functions[82]. CORBA exists in various versions with different functionality. CORBA realizes a strong object oriented concept providing object oriented features, global object identification, managing of distributed object, persistence of inheritance object lifecycle services and modularization up to an atomic level[83]. CORBA has long stay connection allowing integration of huge data in the network.

*Distributed Healthcare Environment (DHE):*is also a middleware architecture permitting cooperation and data sharing between end-user applications including legacy system [82]. This framework was developed by EU project RICHE and EDITH and is implemented in more than 20 hospitals in east and western European countries. The DHE information is presented by using basic concept of entities relationships (ER) modeling, i.e., entity, relations and attributes. Therefore, each service is connected to one of more entities or relations between entities and the information it deals with is described by means of attributes of the involved entities or relations.

Health Industries Level Seven (HL7) interface standards: was founded in the USA as association of vendors, users and organizations who were interrelated to support and promote communication between information systems within hospital environment. The HL7 is based on health care domain related electronic data interchange (EDI) standards. The HL7 is a communication standard for information interchange in health environment, especially in hospitals. HL7 aims at enabling communication between application provider, different venders, using different platforms, operating system, and application environment. In principle, HL7 enables communication between systems regardless of their architecture[84].HL7 is a protocol for information interchange of health care information defining both message and message exchange format[85].The architectural framework of HL7 indicates that it is a point to point information interchange paradigm 1:1 or 1: N in the case of broadband and communication is controlled by trigger events. HL7 has been dominant in USA, Netherlands, German, Finland, UK and Japan.

Despite their importance, all of these systems only go a part of the way towards solving the problem of semantic and or ontology interoperability. They do not provide lightweight application that is a plug-and-play which can run in an environment with higher latency and low bandwidth. Moreover, the basic requirements of science are that experiments be repeatable in retrospective or prospective studies. Furthermore, organizations now need to reuse and analyze shared data to acquire information and knowledge that can underpin business intelligence [40]. This requirement entails data integration to be in open standards and be integrated efficiently regardless of the network infrastructure. To this end, there is a need to develop a dynamic link library for data collection. Likewise, designing an algorithm that can run in a network infrastructure with a short-lived connection is required.

8. Conclusion

It is concluded that data integration is a primary concern with combining data residing at different sources, and providing users with a unified view of the data. The integration may also be concerned with managing the relationship of data from different independent systems. The degree to which the data are coupled may vary depending on the organization business needs. On one side, a need for tightly integrated solution is required while on the other side loosely integrated data model is preferred depending on the organization agility. Therefore, a practitioner should consider the need for better conceptualization and method for implementing partial integration of their data between and within the organization. This can be achieved by using an architecture which is loosely integrated while enforcing standardization. Additionally, choosing the right approach for data integration requires a practitioner to consider the IT infrastructure, better conceptualization and method for implementing partial integration of their data. Also, understanding and making a right choice of business model of the integration technology such as XML, flat file, comma separate values and data table matching is the basic step in enhancing integration framework. Moreover, e-Health records integration is a key feature for health improvement and planning nationwide.

Acknowledgement

This work is carried as a case study research in improving health data collection and integration in distributed low bandwidth area. The financial support was granted to the scholar by the Ifakara Health Institute (IHI) under the scheme of Post Graduate scholarship. Additional fund was granted to scholar by Tanzania Commission for Science and Technology through the Nelson Mandela African Institution of Science and Technology (NM-AIST). We are grateful to the NM-AIST scientific committee and Institution review board of IHI for granting permission to undertake this research in the health sector.

References

[1] P. Ziegler and K. R. Dittrich, "Data Integration—Problems, Approaches, and Perspectives," *Conceptual Modelling in Information Systems Engineering,* pp. 39-58, 2007.

[2] P. Cudré-Mauroux*, et al.*, "Gridvine: An infrastructure for peer information management," *IEEE Internet Computing,* vol. 11, pp. 36-44, 2007.

[3] V. Y. Bichutskiy*, et al.*, "Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: A p53 Cantcer Research Database," *Cancer Informatics,* vol. 2, p. 277, 2006.

[4] N. Arch-int and A.-i. Somjit, "Semantic information integration for electronic patient records using ontology and web services model," in *Information Science and Applications (ICISA), 2011 International Conference on*, 2011, pp. 1-7.

[5] K. Atalag*, et al.*, "Putting health record interoperability standards to work," *electronic Journal of Health Informatics,* vol. 6, p. e1, 2010.

[6] R. J. Glushko and T. McGrath, *Document engineering Analysing and Designing Document for Business information & Web service*: MIT Press, 2005.

[7] K. A. Stroetmann and V. N. Stroetmann, "Towards an Interoperability Framework for a European e-Health Research Area–Locating the Semantic Interoperability Domain," 2005, pp. 14-15.

[8] D. Loshin, *The praactitionners's Guide to Data Quality Improvement*: Morgan Kaufmann 2011.

[9] M. Lenzerini, "Data integration: A theoretical perspective," in *Symposium on Principle of Database system*, 2002, pp. 233-246.

[10] A. Calì, "Reasoning in data integration systems: why lav and gav are siblings," in *Foundations of Intelligent Systems*, ed: Springer, 2003, pp. 562-571.

[11] A. Buccella*, et al.*, "An Ontology Approach to Data Integration," *International journal of simulation system ,sceince and technology,* vol. 11, Ocotber 203 2010.

[12] L. Han*, et al.*, "Visual model of heterogeneous data sources based on service-ontology," 2010, pp. 2945-2949.

[13] S. J. Cockell, *et al.*, "An integrated dataset for in silico drug discovery," *J Integr Bioinform,* vol. 7, p. 116, 2010.

[14] N. J. Salkind, Ed., *Tests & Measures for people who Hate Tests & Measurement*. London ECITY 1SP: Sage publications Ltd, 2006, p.^pp. Pages.

[15] D. Calvanese, *et al.*, "A framework for ontology integration," 2002, pp. 201-214.

[16] J. A. R. Castillo, *et al.*, "Information extraction and integration from heterogeneous, distributed, autonomous information sources-a federated ontology-driven query-centric approach," in *Information Reuse and Integration, 2003. IRI 2003. IEEE International Conference on*, 2003, pp. 183-191.

[17] I. Cruz and H. Xiao, "Ontology driven data integration in heterogeneous networks," *Complex Systems in Knowledge-based Environments: Theory, Models and Applications,* pp. 75-98, 2009.

[18] I. F. Cruz and H. Xiao, "The role of ontologies in data integration," *Engineering intelligent systems for electrical engineering and communications,* vol. 13, p. 245, 2005.

[19] G. Atemezing and J. Pavón, "An Ontology for African Traditional Medicine," in *International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008)*, 2009, pp. 329-337.

[20] B. Louie, *et al.*, "Data integration and genomic medicine," *Journal of biomedical informatics,* vol. 40, pp. 5-16, 2007.

[21] Z. Xu and Y. Lee, "Semantic heterogeneity of geodata," *INTERNATIONAL ARCHIVES OF PHOTOGRAMMETRY REMOTE SENSING AND SPATIAL INFORMATION SCIENCES,* vol. 34, pp. 216-224, 2002.

[22] A. Doan, *et al.*, "Introduction to the special issue on semantic integration," *ACM SIGMOD Record,* vol. 33, pp. 11-13, 2004.

[23] F. Hakimpour and A. Geppert, "Resolving semantic heterogeneity in schema integration," in *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, 2001, pp. 297-308.

[24] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys (CSUR),* vol. 22, pp. 183-236, 1990.

[25] A. Doan and A. Y. Halevy, "Semantic integration research in the database community: A brief survey," *AI magazine,* vol. 26, p. 83, 2005.

[26] R. Chaudhri, *et al.*, "Open data kit sensors: mobile data collection with wired and wireless sensors," in *Proceedings of the 2nd ACM Symposium on Computing for Development*, 2012, p. 9.

[27] A. Giemza, *et al.*, "A mobile application for collecting numerical and multimedia data during experiments and field trips in inquiry learning," in *International Conference on Computers in Education, Putrajaya, Malaysia*, 2010.

[28] X. Dong, *et al.*, "Data integration with uncertainty," 2007, pp. 687-698.

[29] X. L. Dong, *et al.*, "Data integration with uncertainty," *The VLDB Journal,* vol. 18, pp. 469-500, 2009.

[30] A. Satheesh and R. Patel, "Dynamic Nearest Neighbours Classifier For Integrated Data Using Object Oriented Concept Generalization," vol. 11, pp. 35-40, 2010.

[31] H. Kozankiewicz, *et al.*, "Intelligent data integration middleware based on updateable views," *Intelligent Media Technology for Communicative Intelligence,* pp. 29-39, 2005.

[32] D. L. Cal`ı A, Riccardo R,, "Query rewriting and answering under constraints in data integration systems" 2003.

[33] A. Maedche and S. Staab, "Ontology learning for the semantic web," *Intelligent Systems, IEEE,* vol. 16, pp. 72-79, 2001.

[34] Wikipedia. (2011, 13/03/2014). *Extract Transform ,Load*. Available: http://en.wikipedia.org/wiki/Extract,_transform,_load

[35] K. P. Kornelson, *et al.*, "Method and system for developing extract transform load systems for data warehouses," ed: Google Patents, 2006.

[36] P. Vassiliadis, "A Survey of Extract-Transform-Load Technology," ed, 2011.

[37] S. Chawathe, *et al.*, "The TSIMMIS project: Integration of heterogenous information sources," 1994.

[38] M. Van Cappellen, *et al.*, "Data Aggregation, Heterogeneous Data Sources and Streaming Processing: How Can XQuery Help?," *IEEE Data Eng. Bull,* vol. 31, pp. 57-64, 2008.

[39] T. Risch, *et al.*, "Functional data integration in a distributed mediator system," ed: Springer, 2003.

[40] D. George, "Understanding structural and semantic heterogeneity in the context of database schema integration," *Journal of the Department of Computing, UCLAN,* vol. 4, pp. 29-44, 2005.

[41] D. Heimbigner and D. McLeod, "A federated architecture for information management," *ACM Transactions on Information Systems (TOIS),* vol. 3, pp. 253-278, 1985.

[42] D. McLeod and D. Heimbigner, "A federated architecture for database systems," 1980, pp. 283-289.

[43] Webopedia. (2013, 15 March 2013). *Data Model*. Available: http://www.webopedia.com/TERM/D/data_modeling.html

[44] P. Gupta, Ed., *Business Innovation in the 21st century: A Comprehensive aproach to Institutional Business Innovation*. S.Chand & Company, 2009, p.^pp. Pages.

[45] R. L. Richesson and J. Krischer, "Data standards in clinical research: gaps, overlaps, challenges and future directions," *Journal of the American Medical Informatics Association,* vol. 14, pp. 687-696, 2007.

[46] R. Ramakrishnan and J. Gehrke, *Database management systems*: Osborne/McGraw-Hill, 2000.

[47] R. Atun, *et al.*, "Integration of targeted health interventions into health systems: a conceptual framework for analysis," *Health Policy and Planning,* vol. 25, pp. 104-111, 2010.

[48] M. M. Huynen*, et al.*, "The health impacts of globalisation: a conceptual framework," *Globalization and Health,* vol. 1, p. 14, 2005.

[49] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*: Wiley, 2011.

[50] Y.-C. Lu*, et al.*, "A review and a framework of handheld computer adoption in healthcare," *International Journal of Medical Informatics,* vol. 74, p. 409, 2005.

[51] M. Haithcox-Dennis*, et al.*, "Rethinking the Factors of Success: Social Support and Community Coalitions," *American Journal of Health Education,* vol. 44, pp. 110-118, 2013.

[52] J. Yu and R. Buyya, "A taxonomy of workflow management systems for grid computing," *Journal of Grid Computing,* vol. 3, pp. 171-200, 2005.

[53] W. Van Der Aalst and K. M. Van Hee, *Workflow management: models, methods, and systems*: MIT press, 2004.

[54] C. Hagen and G. Alonso, "Exception handling in workflow management systems," *Software Engineering, IEEE Transactions on,* vol. 26, pp. 943-958, 2000.

[55] G. Brzykcy*, et al.*, "Schema Mappings and Agents' Actions in P2P Data Integration System," *J. UCS,* vol. 14, pp. 1048-1060, 2008.

[56] S. Staab and H. Stuckenschmidt, *Semantic web and peer-to-peer*: Springer, 2006.

[57] D. Calvanese*, et al.*, "Inconsistency tolerance in P2P data integration: An epistemic logic approach," *Information Systems,* vol. 33, pp. 360-384, 2008.

[58] Y. Liu and L. Yuefan, "Research on Data Integration of Bioinformatics Database Based on Web Services," in *The 1st International Conference on Networked Digital Technologies (NDT2009)*, 2009.

[59] Webopedia. (2013, Web service.http://www.webopedia.com/TERM/W/Web_Services.html. Available: http://www.webopedia.com/TERM/W/Web_Services.html

[60] C. Walker and D. Walker, "Integration and Data Sharing between WS-Based Workflows," in *Web Services, 2008. ICWS'08. IEEE International Conference on*, 2008, pp. 667-674.

[61] MSDN. (2009, 27 Oct.2013). Microsoft Developer Network. Available: http://social.msdn.microsoft.com/Forums/en-US/435f43a9-ee17-4700-8c9d-d9c3ba57b5ef/advantages-disadvantages-of-webservices?forum=asmxandxml

[62] M. Elammari, "Health Architecture based on SOA and Mobile Agents," in *The 2nd International Conference on Software Engineering and Computer Systems (ICSECS2011), June 27-29, 2011, Kuantan, Malaysia*, 2011.

[63] F. Zhu*, et al.*, "Dynamic data integration using web services," 2004, pp. 262-269.

[64] Y. Zhu*, et al.*, "Research on Web service-oriented data integration in the distributed system," 2011, pp. 568-571.

[65] M. P. Papazoglou*, et al.*, "Service-oriented computing: State of the art and research challenges," *Computer,* vol. 40, pp. 38-45, 2007.

[66] M. P. Papazoglou and W. J. Van Den Heuvel, "Service oriented architectures: approaches, technologies and research issues," *The VLDB journal,* vol. 16, pp. 389-415, 2007.

[67] K. Atalag*, et al.*, "Assessment of Software Maintainability of openEHR Based Health Information Systems–A Case Study In Endoscopy," *Electronic Journal of Health Informatics,* vol. 7, p. e3, 2012.

[68] R. E. Scott, "e-Records in health—Preserving our future," *International journal of medical informatics,* vol. 76, pp. 427-431, 2007.

[69] Nadhan and J.-L. Weldon. (2014, 20 /05/2014). A Strategic Approach to Data Transfer Methods. Available: http://msdn.microsoft.com/en-us/library/aa480064.aspx

[70] T. Katayama*, et al.*, "The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows," *Journal of biomedical semantics,* vol. 1, pp. 1-19, 2010.

[71] C. A. Brandt*, et al.*, "Metadata-driven creation of data marts from an EAV-modeled clinical research database," *International journal of medical informatics,* vol. 65, pp. 225-241, 2002.

[72] L. Stein, "Creating a bioinformatics nation," *Nature,* vol. 417, pp. 119-120, 2002.

[73] C. Bizer*, et al.*, "Linked data-the story so far," *International Journal on Semantic Web and Information Systems (IJSWIS),* vol. 5, pp. 1-22, 2009.

[74] E. Pacitti*, et al.*, "Grid data management: Open problems and new issues," *Journal of Grid Computing,* vol. 5, pp. 273-281, 2007.

[75] A. Nanda, *Hands-on Intergration Services. Microsoft SQL server 2008* 2ed. New York: Mc Graw Hill, 2011.

[76] R. W. Comer*, et al.*, "Automatic Spreadsheet forms," ed: Google Patents, 1998.

[77] L. Han*, et al.*, "RDF123: from Spreadsheets to RDF," in *The Semantic Web-ISWC 2008*, ed: Springer, 2008, pp. 451-466.

[78] S. Holzner, Ed., *Visual Basic .NET Programing Black Book.Comprehensive Problme Solver*. Dreamtech, 2005, p.^pp. Pages.

[79] T. Milo and S. Zohar, "Using schema matching to simplify heterogeneous data translation," in *VLDB*, 1998, pp. 24-27.

[80] P. Helman, *The science of database management*: Richard D. Irwin, Inc., 1994.

[81] D. C. Kaelber*, et al.*, "A research agenda for personal health records (PHRs)," *Journal of the American Medical Informatics Association,* vol. 15, pp. 729-736, 2008.

[82] B. Blobel, *Analysis, Design, and Implementation of Secure and Interoperable Distributed Health Information Systems* vol. 89: IOS Press, 2002.

[83] J. Grimson*, et al.*, "A CORBA-based integration of distributed electronic healthcare records using the synapses approach," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 2, pp. 124-138, 1998.

[84] T. J. Eggebraaten*, et al.*, "A health-care data model based on the HL7 reference information model," *IBM Systems Journal,* vol. 46, pp. 5-18, 2007.

[85] T. Benson, *Principles of health interoperability HL7 and SNOMED*: Springer, 2010.