Real Time Web-based Data Monitoring and Manipulation System to Improve Translational Research Quality

Matthew N. Anyanwu manyanwu@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Venkateswara Ra Nagisetty nnagise@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Emin Kuscu ekuscu@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Teeradache Viangteeravat tviangte@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Abstract

The use of the internet technology and web browser capabilities of the internet has provided researchers/scientists with many advantages, which includes but not limited to ease of access, platform independence of computer systems, relatively low cost of web access etc. Hence online collaboration like social networks and information/data exchange among individuals and organizations can now be done seamlessly. In practice, many investigators rely heavily on different data modalities for studying and analyzing their research/study and also for producing quality reports. The lack of coherency and inconsistencies in data sets can dramatically reduce the quality of research data. Thus to prevent loss of data quality and value and provide the needed functionality of data, we have proposed a novel approach as an ad-hoc component for data monitoring and manipulation called RTWebDMM (Real-Time Web-based Data Monitoring and Manipulation) system to improve the quality of translational research data. The RTWebDMM is proposed as an auditor, monitor, and explorer for improving the way in which investigators access and interact with the data sets in real-time using a web browser. The performance of the proposed approach was evaluated with different data sets from various studies. It is demonstrated that the approach yields very promising results for data quality improvement while leveraging on a web-enabled environment.

Keywords: Bioinformatics, Health Management, Clinical Trial, Basic Research, Data Manipulation, Data Monitoring, Data Cleaning, Data Comparison

1. INTRODUCTION

A data with good quality is needed in-order to produce high quality results from scientific researches and discoveries. This has generated a considerable amount of interest in software/algorithms that can facilitate data quality control. The value of data highly depends on its quality. To enhance the quality of data to be analyzed many data management systems tend to facilitate data quality control before using it in data mart, data mining, or other analytical processes. Data quality control includes all the processes involved in producing and validating good quality data. The processes include but not limited to data-preprocessing/cleaning, data processing, data aggregation and data quality assurance [11]. Data preprocessing involves noise and dimension reduction in the data. In Data processing stage, data is analyzed, aggregated, and incorrect data items eliminated. Also data is examined in this stage for reasonable output [11]. Data aggregation is a measure of the statistical test and analysis of the processed data. Also different statistical tests used in analyzing the data out-put are validated.

Data quality assurance involves the use of quality assurance techniques/methods to validate the data output. Data quality control and validation is used to ensure that good and authoritative data is produced for its purpose [12]. Fan et al., proposed the Semandaq (Semantic Data Quality) [9]. Semandaq is a data quality system that uses conditional functional dependencies for improving the quality of relational data. Galhardas et al., proposed a declarative language along with 1-5 transformation operations to enhance improvements of data monitoring and cleaning process [3]. Harris et at., proposed the Research Electronic Data Capture (REDCap) [4]. The REDCap project uses PHP + JavaScript programming language and MySQL database engine driven methodology and workflow process. The REDCap proposed Data Cleaner and Data Comparison tools to assist in data monitor and cleaning process. However the cleaning actions have to be explicit by the investigators or users. Viangteeravat et al., introduced the Scientific Laboratory Information Management–Patient-care Research Information Management (Slim-Prim) system [7, 8].

The Slim-Prim [7,8] proposed Data Monitor tool to assist the user/researcher with real-time visualization and tracking of the historical data set through Asynchronous JavaScript and XML (AJAX) capability interface. However, it gives users the ability to refine and manipulate their data set to support monitoring. Also the cleaning of poor quality data is still needed under the principal best known as “Your Data, Your Decision”. Raman et al., introduced an interactive ad-hoc technique by providing a spreadsheet-like interface to facilitate the specific transformation operation that can automatically trigger a bad quality of data in the background [6]. Mury et al., proposed the Informatics for Integrating Biology and the Bedside (i2b2) that queries data by dragging and placing the data items into the query environment. This approach is used in retrieving data item from the repository which contains clinical data records [13]. In practice, however, it is estimated that data cleaning is the most labor intensive and a complex process compared to other processes in data quality control [10]. In order to minimize the data cleaning efforts, data quality control process should be part of the data processing stage as it involves data monitoring and manipulation. This stage produces good errors and detects inconsistencies in data items.

In this article, we have proposed the real-time web-based data monitoring and manipulation system (RTWebDMM) as an ad-hoc component that can easily be integrated and interfaced with an existing data management system while having the advantages of being an online web-based system. RTWebDMM provides a graphical user interface (GUI) platform in form of a spreadsheet with build-in macros analytical tools that perform both data monitoring and manipulation in-order to produce good quality data for research purpose. It has also shown to be an indispensable tool in data cleaning efforts, thereby relieving the researcher the burden of data cleaning.

2. RTWebDMM SYSTEM ARCHITECTURE AND FUNCTIONALITY

The Real-Time Web-based Data Monitoring and Manipulation system (RTWebDMM) architecture is depicted in Figure 1. As shown Figure 1, the RTWebDMM is composed of four main components. The first component, Ad-hoc API, uses PHP + Asynchronous JavaScript and XML(AJAX) and JavaScript programming language to build an ad-hoc API (Application Programming Interface) to communicate with the existing Clinical Data Management System (CDMS) and uploads data set of interest into RTWebDMM. The second component comprises the Data monitoring and Data Manipulation sub-components. The Data monitoring sub-component consists of built-in Macro tools, data analysis tools, Data graph (which produces the graphical user interface) and Data tracking tool while the Data manipulation sub-component consists of Data Export which may be in Comma-Separated Value (CSV) or Tab Separated Value (TSV) format, Data Aggregation and Comment Exchange modules. Data Query module initiates the built-in Macro module to work on analytical process using Data Analysis. The Data Analysis is composed of many built-in complex analytical functions such as Mean, Median, Standard deviation, Age calculation, Probability Density Function, BMI (Body Mass Index), Standard error, and Outlier detection.

The RTWebDMM is based on a Simple Spreadsheet [6, 14]. A Simple Spreadsheet provides a basic excel-like framework that supports charts, formulas, and simple custom macros. It also includes sorting capabilities, which has the features of expanding all or specified columns, complex macro built-in analytic formulas, and other features necessary for basic science and clinical research. We have also extended it to include the organization of what we call “Data Query”. Data Query is the real-time SQL (Structured Query Language) that allows a user to easily access and monitors the quality of the data set. The user can interactively manipulate, track the state/status of data, if it has been modified since it was last collected by using the Data Tracking module. The Data Aggregation gives users the ability to manipulate the raw data set to a suitable format before statistician(s) can provide further analysis on the data. The Data Conversion becomes the third component used in creating custom user report, which is static mode (i.e., data snapshot). The fourth component, Custom Report, is an online report module. The report module is used to know the state of the system at any given time.

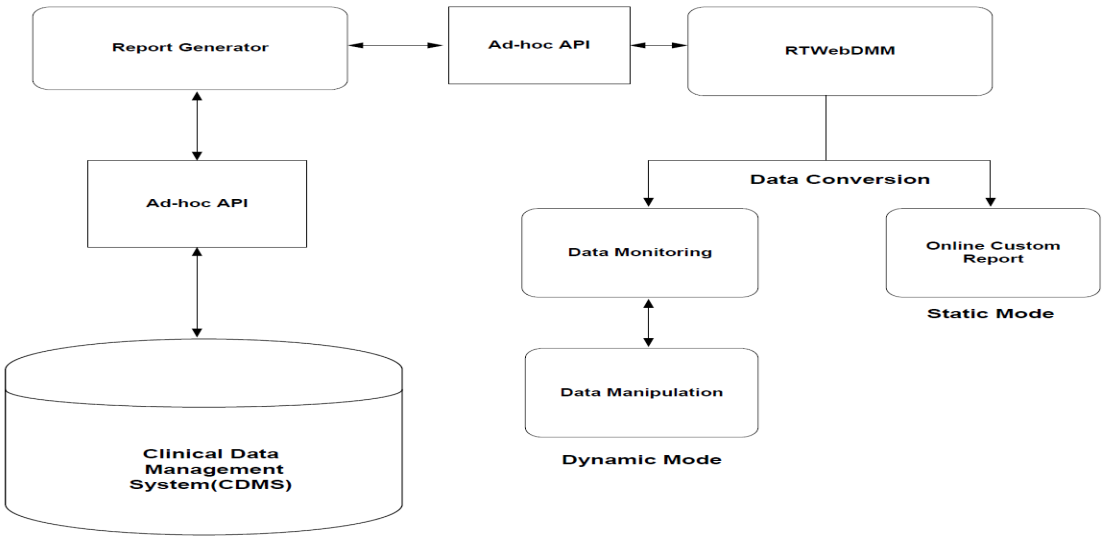

FIGURE 1: RTWebDMM system architecture and functionality

3. RTWebDMM APPLICATION PROGRAMMING INTERFACE (API)

For the purpose of our study, there is an API between our system and CDMS [7, 8] which enables the uploading of data into the RTWebDMM project using the API in Data Query menu as depicted in Figure 2. The RTWebDMM uses PHP application running at the server site to communicate and upload raw data set from Slim-Prim system. The JavaScript (JS) programming language is used to display the query results at the client-side as shown in Figure 3.

In Figure 2, RTWebDMM uses Structure Query Language (SQL) to communicate with the CDMS system. The result of the raw data set is then translated into a required format in which its structure is compliant for rendering to the client using JS Data Grid Writer and Render Class. The JS Data Grid Writer and Render Class is the PHP class written in OOP (Object-Oriented Programming) fashion. It is used to create the compliant JS format for RTWebDMM to render and display its result in dynamic data grid style. Once RTWebDMM successfully renders the displays of its results, also the user is allowed to manipulate and monitor the data set using Data Graph, Data Tracking, and Data Aggregation modules thus detecting outlier or poor quality of data. This process reduces the extensive labor in data manipulation and assists in data cleaning.

The RTWebDMM uses Data Graph and Tracking components to monitor the value of the data set. The user uses either simple built-in formulas (e.g., Mean, Median, or Standard deviation) or complex built-in formula (e.g., Body Mass Index(BMI)) in detecting the quality of data. To enhance collaboration, RTWebDMM provides comment functionality for data comment exchange as shown in Figure 3.

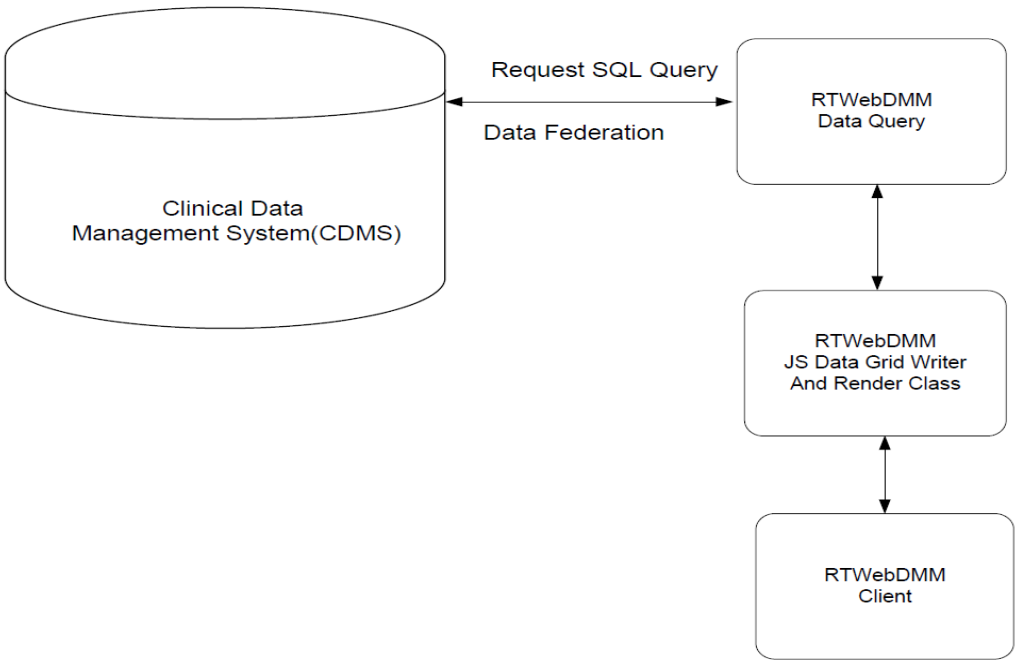

FIGURE 2: RTWebDMM system API (Application Programming Interface)

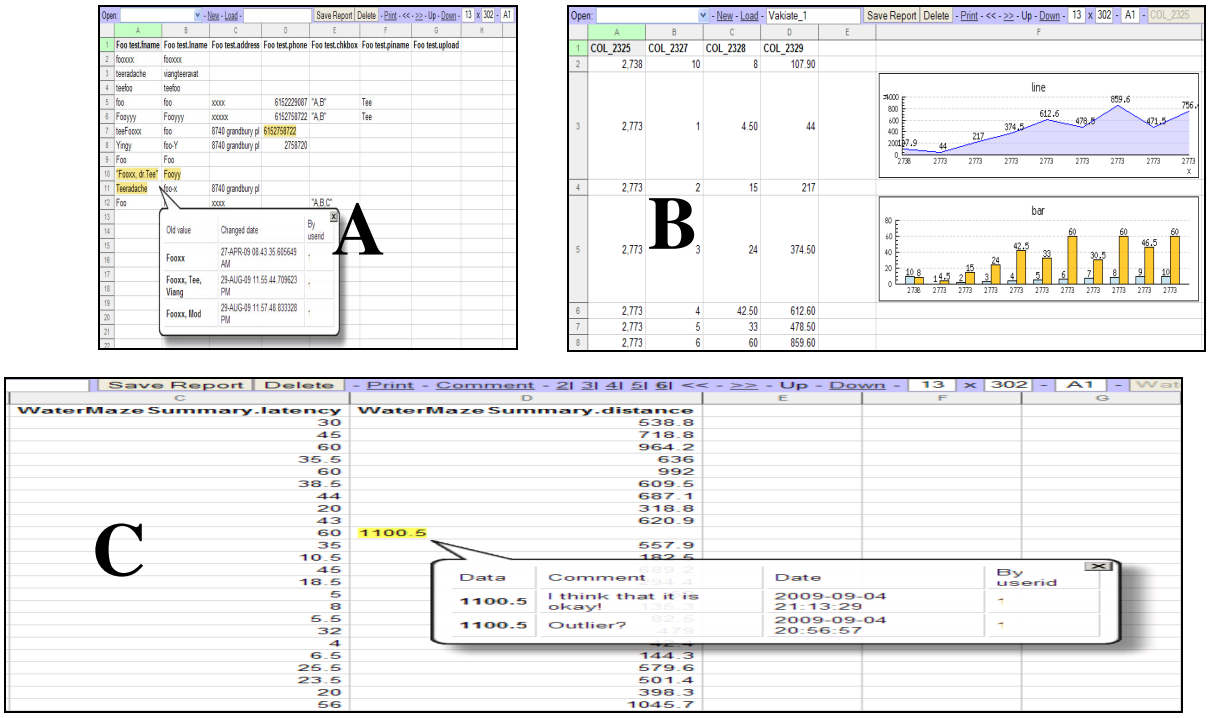

FIGURE 3: RTWebDMM User Interface (web-based) for Historical Data Tracking(A) Data Graph / Analysis(B) and Collaboration supported in comment exchange functionality(C)

4. DISCUSSION AND FUTURE CHALLENGES

The proposed Real-Time Web-based Data Monitoring and Manipulation (RTWebDMM) system has shown significant improvement in reducing extensive labor for data cleaning process. Studies available in literature have shown that it is difficult and challenging to detect poor quality of data conditions in both basic science and clinical research studies [12, 15]. The RTWebDMM attempts to assist in providing a new alternative method in which we are able to monitor uncertainties in relational data sets using built-in independency functions of the data sets. Compared with other data management tools mentioned (See Section 1 and 2) our proposed tool (RTWebDMM) is a real-time web-based data monitoring and manipulation tool and it ensures that good quality data is generated for research purpose. It also has graph Data Graph and Tracking components which monitors the value and quality of data in real-time. RTWebDMM also has API for ease of integration with legacy systems or other data management systems. In practice, basic science and clinical research data deal with relational data dependencies. For instance, the specific zip code within the same county must result in the same name of a city. The ad-hoc functional dependencies that the user can define will have to be established in our future work. The improvements in user-friendly interface (UI) and data cleaning are also part of our future implementation of our system. We also intend to make this tool play a key role in decision support management by creating a repository of commonly used data sets in clinical research. An association between the data items will be created using Apriori Principle which states that “if an item-set is frequent, then all of its subsets must also be frequent” [16]. There will be an alert to signify the presence or absence of a data set item used in the query analysis that has an associate in the repository.

5. CONSLUSION

RTWebDMM tool has been tested with data from clinical research and basic science studies at UTHSC to evaluate its use in improving the quality of data item sets and in eliminating inconsistencies detected in the data sets. It is a web-based tool that gives users of clinical trial research studies real-time access to monitor, manipulate and refine data set items to obtain good quality data for their studies. Data quality affects the result of clinical research studies as the data items are patient medical records which should contain data of highest possible quality that can be obtained. In-order to obtain credible result from clinical research study, the data must be credible [15]. The tool also relieves the user the burden of data cleaning, thus allowing user to focus on the objective of the research study. In future users of clinical research project will leverage on the built-in intelligence of the tool that will be added to determine the association between clinical data set items. The built-in intelligence tools will include the use of standardized techniques and technologies such as; International Classification of Diseases, Ninth Revision (ICD-9) [17], International Classification of Diseases, Tenth Revision (ICD-10) [18], Current Procedural Terminology, 4th Edition (CPT-4) [19], Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [20], Logical Observation Identifiers Names and Codes (LOINC), Recommended Standard Clinical Drug Nomenclature (RxNorm) [21], and Unified Medical Language System (UMLS) [20].

6. REFERENCES

1. C. C. Shilakes and J. Tylman. Enterprise information portals. Technical report, Merrill Lynch, Inc., New York, NY, Nov. 1998.

2. C. Gao, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: Consistency and accuracy. In VLDB, 2007.

3. H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model and algorithms. In VLDB, 2001.

4. P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzales, J. G. Conde. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics 42 (2009) 377-381.

5. PHP Hypertext Preprocessor. http://www.php.net/. 2009. Ref Type: Electronic Citation.

6. V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In VLDB, 2001.

7. T. Viangteeravat, I. M. Brooks, W.J. Ketcherside, R. Houmayouni, N. Furlotte, S. Vuthipadadon, & C.S. McDonald. Clinical & Translational Science Biomedical Informatics Unit (BMIU): Slim-Prim system bridges the gap between laboratory discovery and practice, 2009.

8. T. Viangteeravat, I. M. Brooks, E.Smith, N. Furlotte, S. Vuthipadadon, R. Reynolds, & C.S. McDonald. "Slim-Prim: A biomedical informatics database to promote translational research". Perspectives in Health Information Management, 2009.

9. W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. TODS, 33(1), 2008.

10. Press release, Garter, Inc.. Quoting Bill Hostmann, Research Director, presented at Gartner Business Intelligence Summit in London, UK., February 3, 2005.

11. Y, Akiyama, and K. S. K. Propheter, “Methods of Data Quality Control: For Uniform Crime Reporting Programs”. Criminal Justice Information Services Division Federal Bureau of Investigation. April 2005.

12. A. S. Loebl, "An organizational and historical perspective of a decade of data validation R&D at the Oak Ridge reservation". Data quality control theory and pragmatics Pages: 1 - 6, 1991.

13. S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. Chueh, C.,Churchill, S., et al. Serving the enterprise and beyondwith informatics for integrating biology and the bedside (i2b2).Journal of the American Medical Information Association, 17,124–130. PMID: 20190053. 2010.

14. D. J Power, "A Brief History of Spreadsheets", DSSResources.COM, World Wide Web,http://dssresources.com/history/sshistory.html, version 3.6, 08/30/2004. Photo. Retrieved: September 24, 2010.

15. J. Rothenberg. “A Discussion of Data Quality for Verification, Validation, and Certification (VV&C) of Data to be used in Modeling”. Rand Project Memorandum PM709-DMSO. 1977. P. Tan, M. Steinbach, and V. Kumar. “Introduction to Data Mining”. Addison Wesley, New York, 2006.

16. Centers for Disease Control and Prevention ."International Classification of Diseases,Ninth Revision (ICD-9)”. http://www.cdc.gov/nchs/icd/icd9.htm. Retrieved: January, 2011.

17. Centers for Disease Control and Prevention. "International Classification of Diseases,Ninth Revision (ICD-9)”. http://www.cdc.gov/nchs/icd/icd10.htm. Retrieved: January, 2011.

18. American Medical Association. "CPT” http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.shtml . Retrieved: January, 2011.

19. United States National Library of Medicine. "SNOMED Clinical Terms” http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. Retrieved: January, 2011.

20. C. McDonald , S.M. Huff, J. Suico , & K.Mercer (eds). Logical Observation Identifiers Names and Codes (LOINCR) users’ guide. Indianapolis: Regenstrief Institute; 2004.