Extending the Medici Data Management Service for Medical Research involving Distributed Teams

Luigi Marini[1], Ashwini Vaidya[1], Nick Tenczar[1], Norma Kenyon[3],
Amelia Bartholomew[2], Dan Salomon[4], Kenton McHenry[1]
[1] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
[2] University of Illinois at Chicago
[3] University of Miami
[4] Scripps Institute
{lmarini, mchenry}@illinois.edu

Abstract—**Managing and analyzing data in medical research is often an ad-hoc process using a wide variety of tools if any at all. To help organize and analyze data produced from medical research studies we propose a lightweight series of services that attempt to complement existing workflows of research teams. In particular, we have been developing and applying a prototype to help improve islet and renal allograft survival and function in nonhuman primates using mesenchymal stem cells.**

Keywords—data management; bioinformatics; medical informatics; data mining; cyberinfrastructure;

I. Introduction

Different research labs have access to different data management and analysis tools and individual researchers are more proficient with certain tools than others. General-purpose applications such as the Excel spreadsheet application are popular for data management, but users can easily reach their limits due to size constraints and limited processing and organizational capacities [2]. Providing brand new data management and analysis tools can be expensive and instrumenting existing tools can be complicated and prone to error or excessive complexity. Instead we propose complementing existing tools with services to improve the individual workflows by adding value without removing familiarity.

We present a use case for which we are trying to improve islet and renal allograft survival and function in nonhuman primates using mesenchymal stem cells by providing data management and analysis tools. Most of the data is provided as Excel workbooks organized in three categories and formats: (a) monitoring information tracking a variety of variables such as blood glucose, insulin, weight, C-peptide, rapamycin, FK506 (two immunosuppressant drugs) of the subjects used in the study tracked on a daily/weekly resolution; (b) haplotype information for a subset of subjects; (c) complete blood count (CBC) variables obtained at irregular intervals.

Instead of building web interfaces specific to the data and formats seen so far, we propose a more extensible framework that allows the researchers to keep using Excel to track the original data, but that is capable of organizing the original workbooks and extracting the embedded data. The system then makes this data available in open interfaces using modern web standards to make it easy to access and analyze the data in different analysis environments, such as the R environment for statistical analysis, the Pandas Python library and the scikit-learn data mining toolkit. We provide data mining examples to model time of decrease in function of the transplant and time of graft loss as well as calculating haplotype matching for identification of donor and recipient.

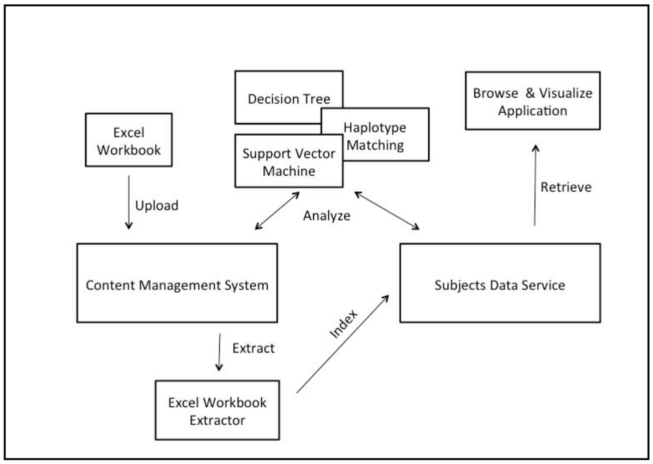


Fig. 1. Architecture and data flow.

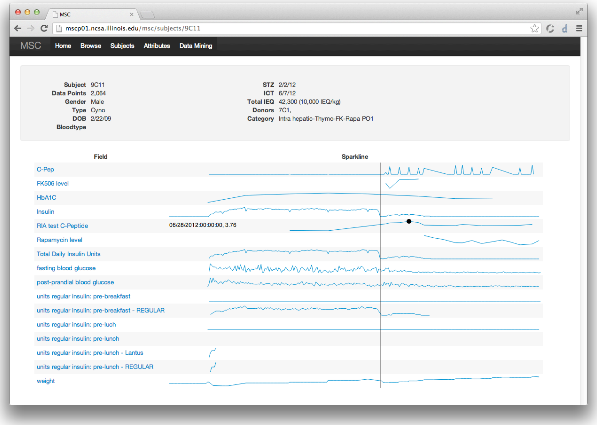
II. Proposed System

The proposed framework consists of a content management system to store the original Excel workbooks containing the raw data and any derived dataset; a service for storing and querying large collections of biomedical timeseries data; a collection of extractors to extract information from the workbooks and ingest it into the biomedical timeseries service; a web application to visualize and launch analytics on the data; and a series of machine learning algorithms to analyze the data.

Individual researchers can upload excel workbooks containing data collected in the lab to the Medici web-based content management system [1]. They can use this service to share access to the data with distributed teams and more easily discuss and organize the files using built-in social annotation features. Relationships between files can be established in Medici to store provenance information [3]. This can be used to keep track of versions of a particular file or to establish relationship between the original file and derived files created by executing analytical tools on the data.

Once a file is uploaded to Medici, it is submitted to a cloud of special purpose extractors configured by the administrators of that particular instance. Based on the file type, matching extractors will attempt to extract relevant information from that file. This allows for great flexibility and the ability of adding new extractors as needed.

For each data type we developed a separate extractor specific to the format of the data. Each extractor submits data points to a Subjects Data Service where they are stored in a common format as streams of datapoints indexed by time and subject. Since it is not known in advance what variables might be tracked for a specific subject or study, we adopted JSON as the interchange format of an HTTP based RESTful API and developed the service as a Scalatra web application on top of MongoDB, a document oriented database where each document is a JSON object. This enabled us to create a flexible and scalable service that does not require one to know in advance what variables are being tracked. It also provided a common layer against which other tools can be written. As most data analysis environments provide ways of making HTTP requests and manipulating JSON objects it also makes it easier for researchers to retrieve data in the tools with which they are most comfortable. We also provide a simple way of downloading the data as comma-separated values (CSV) files (which can be opened by Excel).



Subject 9C11
Data Points 2,064
Gender Male
Type Cyno
DOB 2/22/09
Bloodtype

STZ 2/2/12
ICT 6/7/12
Total IEQ 42,300 (10,000 IEQ/kg)
Donors 7C1,
Category Intra hepatic-Thymo-FK-Rapa PO1

Field | Sparkline

C-Pep
FK506 level
HbA1C
Insulin
RIA test C-Peptide 06/28/2012:00:00:00, 3.76
Rapamycin level
Total Daily Insulin Units
fasting blood glucose
post-prandial blood glucose
units regular insulin: pre-breakfast
units regular insulin: pre-breakfast - REGULAR
units regular insulin: pre-luch
units regular insulin: pre-lunch
units regular insulin: pre-lunch - Lantus
units regular insulin: pre-lunch - REGULAR
weight

Fig. 2. Interactive spark line visualization of all known variables tracked for a specific subject.

A visualization and browsing service built against the Subjects Data Service is provided for the researcher to get an overview of the available data (Fig. 2). This service provides views across all data, not just individual spreadsheets, and makes it possible to compare specific variables across all subjects or all variables for one specific subject.

A series of statistical and data mining analyses were developed using the Python language using the Pandas data analysis library, numpy, scikit-learn and the R statistical environment. Pandas was used to develop several preprocessing methods to clean up the data, such as filling missing data, summarization, and pivoting. To help match recipients and donors based on haplotype, we developed several statistical methods based on feedback from the medical researchers. We modeled insulin requirements using the rpart library in R and trained decision trees (Fig. 3) and support vector machines to model decline in function using the scikit-learn library.

Using the Subjects Data Service to retrieve data about all subjects has made writing new algorithms in different languages and using different libraries a lot easier than if we had to manually aggregate all the data from all the spreadsheets manually as needed by the specific data analysis. Also not requiring the use of a specific data analysis framework leaves much more flexibility to the researcher.

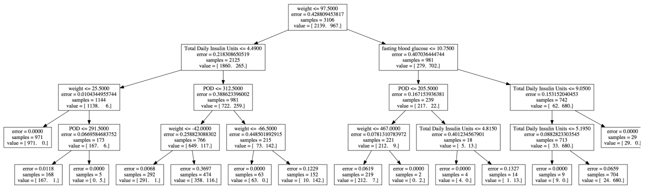


Fig. 3. A decision tree automatically constructed from the data to model decline in transplant function as predicted by other attributes.

III. Conclusions

Complementing the use of data management tools, such as Excel by medical researchers, with web based services running in the cloud and special purpose data analysis tools running on the desktop has minimized disruption to medical researchers, provided new grounds for advanced functionality, and allowed the data scientists the ability to use different analytical frameworks as they see fit. While the web interfaces developed so far have been received positively by the medical researchers, the complexity of some of the data analytics tools has made it difficult to make them available directly to the medical researchers. We are currently trying to provide better interfaces to the data analysis tools.

Acknowledgments

This work at NCSA/UIUC has been supported by the National Institutes of Health under grant 1P01AI089556-01A1.

References

[1] Marini, L., R. Kooper, J. Futrelle, J. Plutchak, A. Craig, T. McLaren, and J. D. Myers, "Medici: A Scalable Multimedia Environment for Research", Microsoft Research eScience Workshop, Berkeley, CA, 10/11/2010.

[2] Anderson, Nicholas R., et al. "Issues in biomedical research data management and analysis: needs and barriers." *Journal of the American Medical Informatics Association* 14.4 (2007): 478-488.

[3] Luc Moreau, Paul Groth, Simon Miles, Javier Vazquez-Salceda, John Ibbotson, Sheng Jiang, Steve Munroe, Omer Rana, Andreas Schreiber, Victor Tan, and Laszlo Varga. 2008. The provenance of electronic data. *Commun. ACM* 51, 4 (April 2008), 52-58.