Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2017 February 11; 10137: . doi:10.1117/12.2254689.

Federating Heterogeneous Datasets to Enhance Data Sharing and Experiment Reproducibility

Juan C. Prieto[a], **Beatriz Paniagua**[a], **Marilia S. Yatabe**[b], **Antonio C.O Ruellas**[b], **Liana Fattori**[b], **Luciana Muniz**[b], **Martin Styner**[a], and **Lucia Cevidanes**[b]

[a]NIRAL, UNC, Chapel Hill, North Carolina, United States

[b]DCBIA, UMICH, Ann Arbor, Michigan, United States

Abstract

Recent studies have demonstrated the difficulties to replicate scientific findings and/or experiments published in past.[1] The effects seen in the replicated experiments were smaller than previously reported. Some of the explanations for these findings include the complexity of the experimental design and the pressure on researches to report positive findings. The International Committee of Medical Journal Editors (ICMJE) suggests that every study considered for publication must submit a plan to share the de-identified patient data no later than 6 months after publication. There is a growing demand to enhance the management of clinical data, facilitate data sharing across institutions and also to keep track of the data from previous experiments. The ultimate goal is to assure the reproducibility of experiments in the future. This paper describes Shiny-tooth, a web based application created to improve clinical data acquisition during the clinical trial; data federation of such data as well as morphological data derived from medical images; Currently, this application is being used to store clinical data from an osteoarthritis (OA) study. This work is submitted to the SPIE Biomedical Applications in Molecular, Structural, and Functional Imaging conference.

Keywords

Data federation; reproducibility; clinical data; sharing; de-identified; web visualization

1. INTRODUCTION

The primary motivation of this work is to improve the state of clinical research data organization in order to facilitate data sharing across institutions and collaborators and ultimately to facilitate the reproducibility of clinical trials. A number of issues have been identified when working and managing clinical data recorded during clinical trials funded by the National Institute of Health (NIH) or private institutions. Frequently, clinical data is recorded and stored in spreadsheets by the clinicians. Very often, before the data analysis begins, a significant amount of time is needed to parse, format and detect outliers in the data in order to produce a dataset suited for analysis.

Send correspondence to J.C.P., jprieto@med.unc.edu.

After the analysis is completed and the results have been published, in the majority of cases, the clinical data is not public and/or shared beyond the data holder and collaborators. This situation has drawn the attention of the scientific community due to the fact that many scientific findings cannot be easily replicated by other groups.

As time progresses, the reproducibility of a study decreases for a number of reasons: losing track of the location of the data, the scientist is mutated to another institution and/or the the data needs to be reprocessed. As shown by The Open Science Foundation, in an effort to replicate 100 experiments reported in psychological science during 2008,[1] about one-third to one-half of the original findings were also observed in the replication studies. Similar patterns can be observed in other areas such as cancer research and engineering. After the experiments were replicated, the effects seen were smaller than previously reported. Some of the explanations for these findings include the complexity of the experimental design and the pressure on researches to report positive findings.

Recent guidelines proposed by The International Committee of Medical Journal Editors (ICMJE) suggests that a study considered for publication should contain a plan to share the de-identified patient data (IPD) and also supplementary material no later than 6 months after publication.[2] The main goal is to improve the transparency and robustness of the publications.

Today, there is a growing demand to enhance the Clinical Data Management (CDM)[3] in the scientific community. CDM describes the processes to maintain high-quality and reliable data from clinical trials. Some of the procedures in CDM include data annotation, database design, data validation etc. Recent advancements in web technologies could provide the necessary means to generate publicly available tools and/or services to facilitate data sharing across institutions, keep track of data from previous experiments and more importantly to assure the reproducibility of experiments in the future.

This paper describes Shiny-tooth, a web based application created to facilitate data acquisition during the clinical trial; data federation of such data as well as morphological data derived from medical images; interactive visualization of the clinical and morphological data; and task submission to remote computing grids using the data stored in the system.

For data storage, we use couchdb, a NoSQL database that uses Javascript Object Notation (JSON) format to store documents and offers a document-based query and indexing mechanism Additionally, couchdb allows attaching binary data to the documents, i.e., medical images, tessellations, etc.; it is implemented following Representational State Transfer (REST) architecture to insert, modify, retrieve and delete records; and offers a synchronization feature between two couchdb instances. Additional frameworks have been used to develop the application, the following section describes the application's architecture.

Currently, this application is being used to store clinical data from an osteoarthritis (OA) study. OA is the most prevalent arthritis worldwide, is associated with significant pain and disability and affects 13.9% of adults at any given time. OA is very complex and the pathogenesis of TMJ OA remains unclear to this day. OA affects the temporomandibular joint (TMJ), among other joints. TMJ OA represents 42.6% of the diagnosed disorders of the TMJ, and results in $4 billion annual health care costs in the US[4.5] In the future we expect to create repositories to study rare diseases such as OA and facilitate anonymized data sharing across institutions.

2. METHODS

Shiny-tooth is focused on the re-usability of components and creating a robust and scalable application.

The back-end of the application is built using Node.js[*], Hapi.js[†], Couchdb[‡] and user authentication is done with Json Web Tokens (JWT)[§].

The front-end of the application is build using Angular.js[¶], Data Driven Documents (D3.js)[||] and Three.js[**] for 3D visualization.

Additionally, a plug-in for 3DSlicer[††] is implemented to commit and retrieve data directly from the system.

The following sections gives additional detail about the tools used to build this system.

2.1 Back-end framework

Node is a Javascript engine that facilitates building scalable network applications. Using Hapi as the server framework, we are able to build services and focus on writing reusable application logic instead of pure infrastructure. Hapi is fully REST and orchestrates communication between all components in the system. The tool used for storage is Couchdb, a NoSQL type database, i.e., it does not store data and relationships in tables. Instead, each database is a collection of independent JSON documents. JSON is a flexible format and facilitates encoding data without enforcing a predefined rigid structure.

For our application, storing data in such format presents an advantage as we don't know beforehand the structure of incoming data. Once the data is stored in the system, the relationships between documents are discovered using the map reduce algorithm and generating views indexing the data in the system.

User authentication is handled with JWT. A plug-in is developed allowing storage and retrieval of user information, as well as JWT encryption exchanged with the user upon login.

Finally, clusterpost[6] is integrated in the system. This plug-in allows submitting tasks to remote computing grids using the data stored in the system.

[*] https://nodejs.org
[†] http://hapijs.com/
[‡] http://couchdb.apache.org
[§] https://jwt.io
[¶] https://angularjs.org
[||] https://d3js.org/
[**] https://threejs.org/
[††] https://www.slicer.org

2.2 Front-end framework

The front end of the application is based on Angular.js, this framework facilitates the development of reusable HTML components. The clinical data visualization is done using D3.js and 3D visualization of morphological data is accomplished with Threejs. The application is hosted using Amazon web services or the Elastic Computing Cloud (EC2). Shiny-tooth has been applied to store clinical OA data.

2.3 3D Slicer plug-in

3D Slicer is an open source software for 3D image processing and visualization. It is known as the “Swiss knife” for medical images as it provides several tools to manipulate them. Furthermore, it allows software developers to contribute plug-ins and Slicer users to install them via an extension manager.

We have developed a Slicer extension that enables clinicians to download and upload data directly to shiny-tooth server. Additionally, we have implemented the interface to use clusterpost and submit high intensive computational tasks to remote computing grids.

3. MATERIALS

The current study consists of 218 TMJ joints, (153 TMJ OA, 65 Controls) obtained from CBCT images.

Label maps of the TMJs were generated and point distribution models (PDM) with 1002 correspondent points were constructed using SPHARM-PDM.[7]

Previously acquired clinical and biological data is added to the repository. Incoming data is being recorded directly in the system via study specific questionnaires.

4. RESULTS

4.1 Back-end server architecture

Figure 1 shows the architecture of the application and the plug-ins deployed in the system. This architecture allows adding plug-ins and enable new services. A full documentation of all the different REST services is available at https://ec2-52-42-49-63.us-west-2.compute.amazonaws.com:8180/docs.

4.2 Front-end views

Figure 2 shows different views to input data to the system. Figure 2a shows a collection of Biological Markers. A collection in this system is defined as a grouping structure for clinical data and morphological data. The clinical data is stored as a JSON document, each entry is indexed separately and may be retrieved using the document’s content with the help of the map reduce algorithm. A collection may be utilized to facilitate the retrieval of meaningful datasets in a study. Figure 2b displays the clinical data in a table format. Another feature of the system is the possibility to import previously acquired data in spreadsheets. However, the content of the spreadsheet must be converted to a comma separated value (CSV) format with the first row being the table header. The header or column names will be used to index the data in the future. If the data is imported from a spreadsheet, it is necessary to specify the column that contains the patient anonymized id. This id will be used to index the JSON document and perform joints between clinical and morphological data. Figure 2c shows an extract from a form used to acquire data in an ongoing OA study. Figure 2d shows the contributions performed by the clinicians to the system. Each time a clinician enters a survey in the system, his contribution is recorded. We seek to record usability statistics and develop project management capabilities in the future.

Figure 3a shows a view where the user can select different clinical variables, in this case, the variables correspond to genetic data acquired for the OA study. Figure 3b shows patient selection feature. Figure 3c shows a plot using the selected variables in the clinical data and patient section.

The plot produced in figure 3c is an interactive plot where the user can select a line and/or the legend on top in order to retrieve morphological data from the database. Figure 4 shows a drop down menu that is populated with the patient's data and the visualization of the structure in the interactive 3D viewer.

The following section describes a desktop plug-in for the 3D Slicer application. This plug-in may be used to populate the system with new data.

4.3 3D Slicer plug-in

This extension contains multiple panels that allow the user to manage the data stored in the remote server. Figure 5a shows the view to login to the system or retrieve the JWT. Once the token is acquired, the user may download or upload data to the system. Figure 5b describes the file structure used when downloading data to the desktop application. Figure 5c shows a view were the user may download data using ids in the system. Figure 5d shows the view to upload new data to the system.

5. CONCLUSIONS

The current state of the application allows gathering clinical data and morphological data in a structured manner. The tools and plug-ins described shown in Figure 1 have been published and are available in the node package manager repository.

The current implementation of shiny-tooth has been deployed in the EC2 container and is available here [‡‡]. The access to the data is restricted and will only be allowed after authorization from the project managers.

We have developed a plug-in to facilitate interaction with the system. The extension is developed for one of the most popular software for medical image processing and three-dimensional visualization. 3DSlicer is used by many research groups worldwide, is open-source, and provides a wide range of processing tools to physicians, researchers, and the general public.

[‡‡] https://ec2-52-42-49-63.us-west-2.compute.amazonaws.com:8180

With the tools presented here, we seek to provide new possibilities to record previous studies, facilitate data-sharing, and improve experiment reproducibility.

References

1. Collaboration OS, et al. Estimating the reproducibility of psychological science. Science. 2015; 349(6251):aac4716. [PubMed: 26315443]
2. DBT, JB, CB, et al. Sharing clinical trial data: A proposal from the international committee of medical journal editors. JAMA. 2016; 315(5):467–468. [PubMed: 26792562]
3. Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS. Data management in clinical research: An overview. Indian Journal of Pharmacology. 2012; 44(2):168–172. [PubMed: 22529469]
4. Cevidanes L, Hajati AK, Paniagua B, Lim P, Walker D, Palconet G, Nackley A, Styner M, Ludlow J, Zhu H, Phillips C. Quantification of condylar resorption in temporomandibular joint osteoarthritis. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 2010; 110(1): 110–117.
5. Paniagua B, Cevidanes L, Walker D, Zhu H, Guo R, Styner M. Clinical application of spharm-pdm to quantify temporomandibular joint osteoarthritis. Computerized Medical Imaging and Graphics. 2011; 35(5):345–352. [PubMed: 21185694]
6. Danaele Puechmaille, MS., Prieto, JC. SPIE Medical Imaging. International Society for Optics and Photonics; 2017. Civility: Cloud based interactive visualization of tractography brain connectome.
7. Styner M, Oguz I, Xu S, Brechbuehler C, Pantazis D, Levitt J, Shenton M, Gerig G. Frame-work for the statistical shape analysis of brain structures using spharm-pdm. Jul.2006

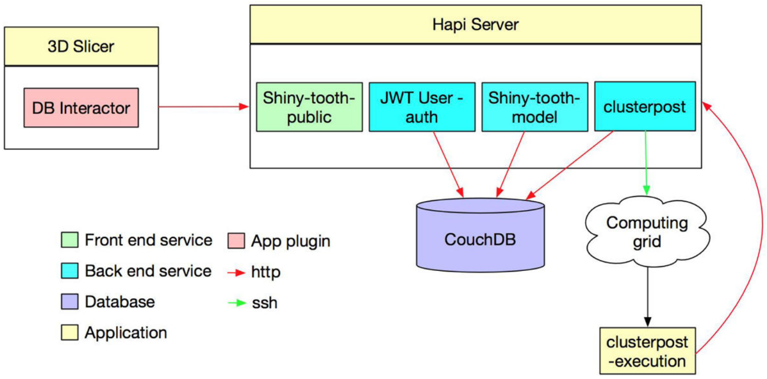
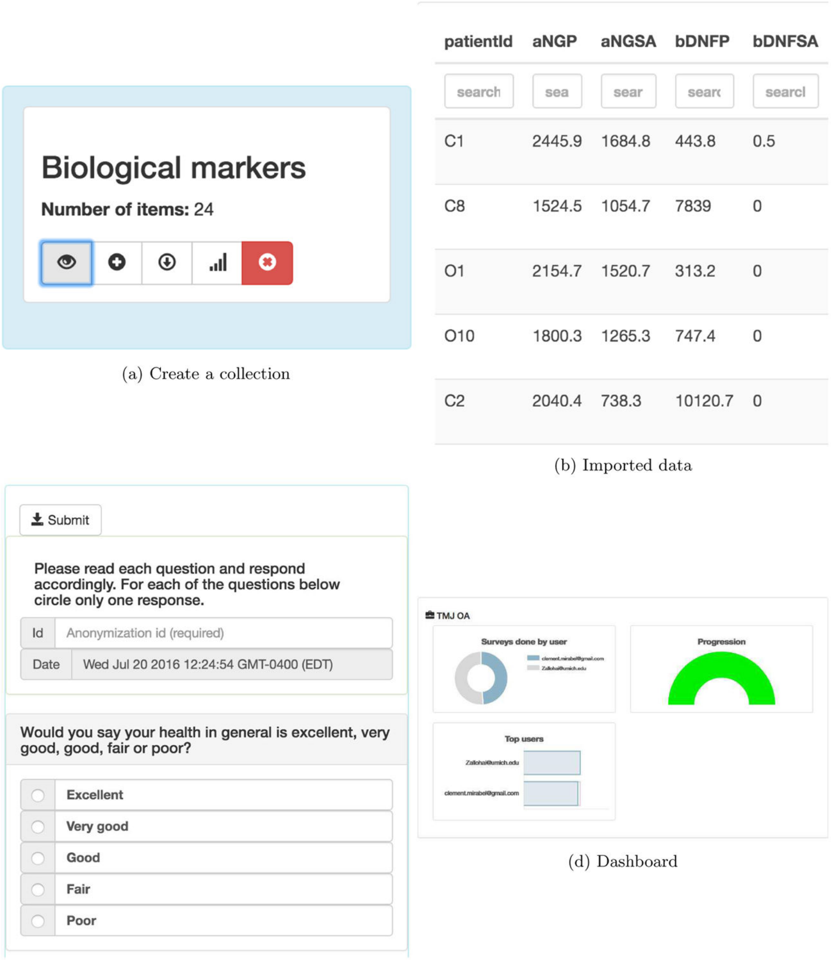


Figure 1.

Server architecture based on plugins for a distributed application. The framework allows easy integration of a variety of plug-ins. The authentication system is based on JSON Web Tokens. The clusterpost plug-in allows submitting heavy computational tasks to remote computing grids.



Biological markers

Number of items: 24

(a) Create a collection

patientId	aNGP	aNGSA	bDNFP	bDNFSA
search	sea	sear	searc	searcl
C1	2445.9	1684.8	443.8	0.5
C8	1524.5	1054.7	7839	0
O1	2154.7	1520.7	313.2	0
O10	1800.3	1265.3	747.4	0
C2	2040.4	738.3	10120.7	0

(b) Imported data

Please read each question and respond accordingly. For each of the questions below circle only one response.

Id	Anonymization id (required)
Date	Wed Jul 20 2016 12:24:54 GMT-0400 (EDT)

Would you say your health in general is excellent, very good, good, fair or poor?

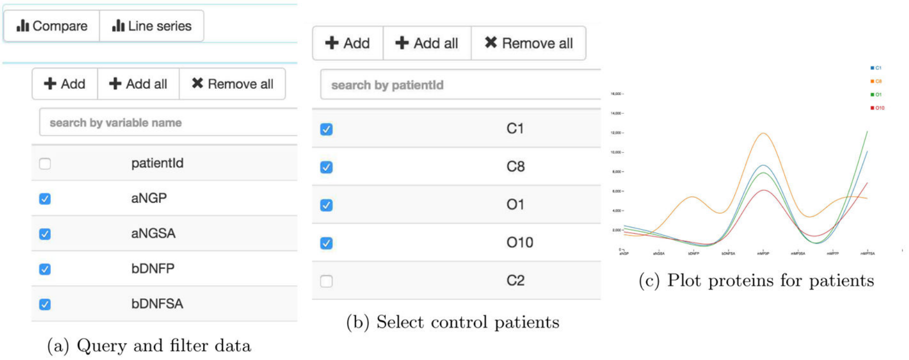
- ○ Excellent
- ○ Very good
- ○ Good
- ○ Fair
- ○ Poor

(c) Questionnaire

(d) Dashboard

Figure 2.

a) Create a group or collection for clinical data. b) Forms or questionnaire to record data. c) Imported data from spreadsheet d) Dashboard with contributions from clinicians



□ patientId
☑ aNGP
☑ aNGSA
☑ bDNFP
☑ bDNFSA

(a) Query and filter data

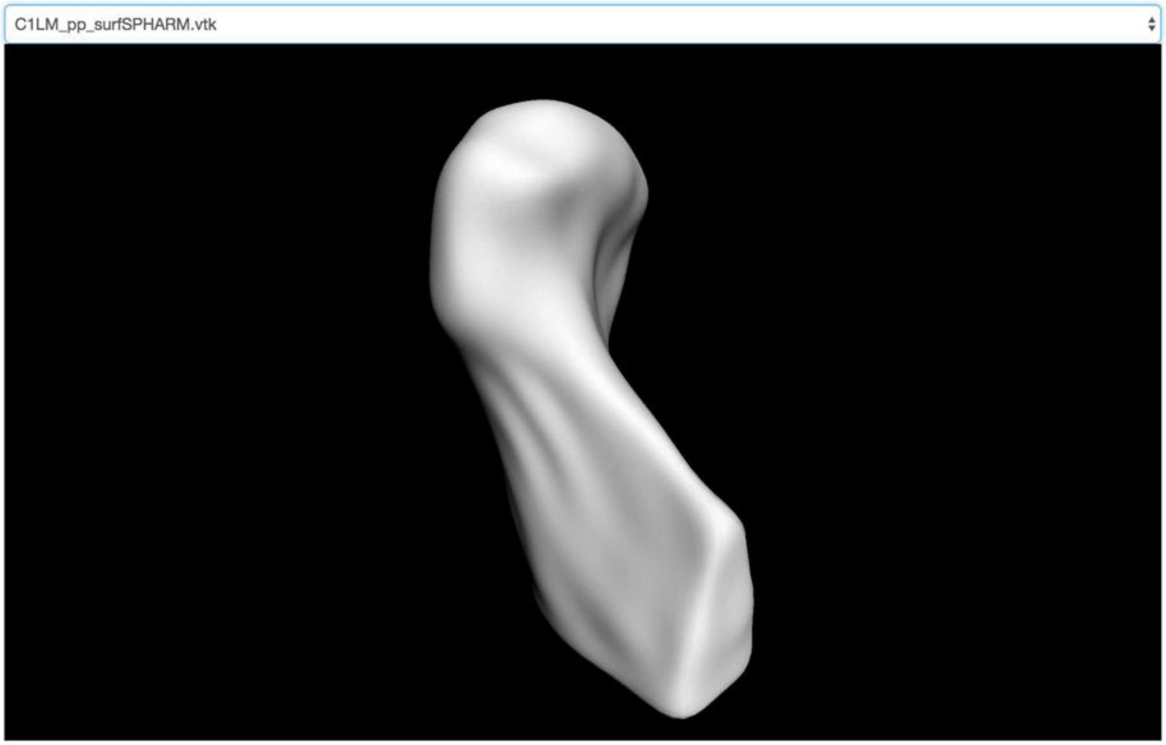
☑ C1
☑ C8
☑ O1
☑ O10
□ C2

(b) Select control patients

(c) Plot proteins for patients

Figure 3.

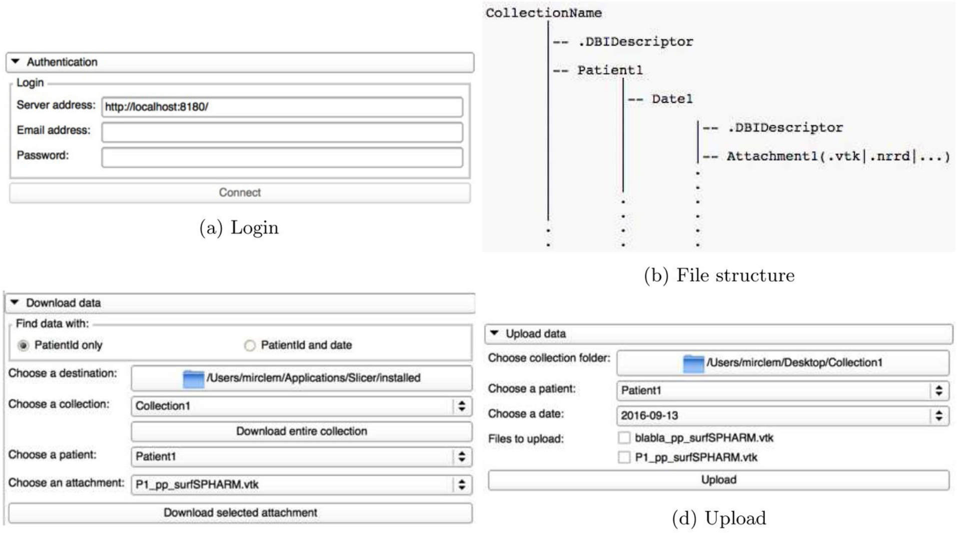
a) Query and select clinical data. b) Query and select patients. c) Plot the selected data



C1LM_pp_surfSPHARM.vtk

Figure 4.

Condyle visualization in 3D. The viewer is embedded in the web.



▼ Authentication

Login

Server address: http://localhost:8180/

Email address:

Password:

(a) Login

```
CollectionName
        |
        |-- .DBIDescriptor
        |
        |-- Patient1
        |         |
        |         |-- Date1
        |         |         |
        |         |         |-- .DBIDescriptor
        |         |         |
        |         |         |-- Attachment1(.vtk|.nrrd|...)
        |         |         .
        |         .         .
        .         .         .
        .         .         .
```

(b) File structure

▼ Download data

Find data with:

● PatientId only ○ PatientId and date

Choose a destination: /Users/mirclem/Applications/Slicer/installed

Choose a collection: Collection1

Choose a patient: Patient1

Choose an attachment: P1_pp_surfSPHARM.vtk

(c) Download

▼ Upload data

Choose collection folder: /Users/mirclem/Desktop/Collection1

Choose a patient: Patient1

Choose a date: 2016-09-13

Files to upload: □ blabla_pp_surfSPHARM.vtk
□ P1_pp_surfSPHARM.vtk

(d) Upload

Figure 5.

a) Login view in 3D Slicer b) download file structure for different time points c) Download data using patient id and date d) Upload generated data in 3D Slicer