Data Fetcher

An Interactive Qualifying Project submitted to the Faculty of WORCESTER POLYTECHNIC INSTITUTE in partial fulfilment of the requirements for the degree of Bachelor of Science

by

Andrew Shanaj

20 March 2021

Submitted to:

Dr. Neil Heffernan

Acknowledgments

Throughout my last two terms, working with the ASSISTments' crew has been an absolute pleasure. I want to reach out and give a personable thank you to everyone on the team. I want to specifically thank Ethan Prihar. Without you, this IQP would've not been possible, and your guidance, encouragement, and expectations were appreciated. I also want to personally thank Ryan Emberling. You often went above and beyond with all the education, office hours, and resources you were able to provide. Your organization, professionalism, and teaching were by far exceptional. On top of that, the personal development skills you were able to offer me are appreciated and something I will hold on to. On top of that, I want to thank Korinn, Taylor, and everyone that was part of the water-cooler slack conversations during these times of distance were appreciated, supporting our morale. Lastly, I want to also thank Professor Neil Heffernan for allowing me to join his class and providing me with an opportunity to work on an Interactive Qualifying Project related to my interests especially after international travel was canceled. Thank you all!

Abstract

As online learning has become increasingly more common, a significant amount of interest in educational research has grown at WPI and throughout other universities. With the rise of COVID-19, educational platforms have become more prominent than ever racking up a massive amount of data points within their ecosystems. One university-based platform for online learning, ASSISTments, is unique for its built-in capacity for research. One prominent problem within the ecosystem is related to its data distribution technique. In this IQP report, I am going to outline the issues of data distribution inherent to ASSSITments' research features, and its high reliance on human intervention by the ASSISTment's research team. On top of these issues, I will also discuss the web application that I built, which integrates the Center for Open Science (COS) Open Science Framework (OSF), supporting the research workflow and transparency of academic research. I will also be discussing future work and integration with the E-Trials research testbed.

Contents

Table of Figures

1 Introduction

1.1 The ASSISTments Platform

ASSISTments is a free interactive tutoring system that aims to help teachers provide the best possible educational experience to students (Heffernan 2014). In normal usage, ASSISTments' online platform allows teachers to assign homework problems to their students. Teachers can provide a wide variety of homework assignments, varying from traditional, multiple-choice problem sets, open response problem sets, and mastery-based problem sets which are also known as "Skill Builders". The platform easily allows teachers to build homework assignments based on their teaching curriculum but also provides the option to source proven problem sets. The platform also provides immediate feedback to students and encourages students to try again, pushing educational growth. On top of immediate feedback, teachers receive instant data to help inform and improve instruction, as well as pinpoint individuals struggling and needing a little extra motivation.

The ASSISTments framework also allows researchers to combine several assignment components to produce a randomized controlled trial (RCT). Researchers can utilize the large population of classrooms to find potential subjects for their experiments (ASSISTments, 2019c)

In 2019, Dr. Heffernan and his wife Christina founded the ASSISTments Foundation to scale up and expand the ecosystem. This has prompted a rewrite of the platform to "version 2.0", to rid the technical debt built up from the graduate-led development of the project. Due to the ongoing effort of development of "version 2.0", research is currently taking place in "version 1.0"

1.2 E-TRIALS, RCTs and ALI

In addition to ASSISTments' utility as a classroom tool, the software is also used to conduct randomized controlled trials (RCTs) to test the efficacy of educational techniques. This feature of the platform is known as EdTech Research Infrastructure to Advance Learning Science or E-TRIALS for short. In short, RCTs can be constructed by randomly assigning students into control or experimental conditions and measuring their performance.

The beauty of E-TRIALS is in its capability to connect researchers to a large sample of the population without the inefficiencies faced when conducting experiments in the education fields. Inefficiencies such as finding schools to conduct these studies as well as building platforms to conduct these studies are all mitigated. This means that researchers can focus more on the content of their study rather than conducting the study.

Preprocessing and delivery is handled by the Assessment of Learning Infrastructure (ALI). This service queries the ASSISTments database for data related to an experiment and delivers that information to researchers after a human intervention has reviewed the data requested.

1.3 COS and OSF

The Center for Open Science is a nonprofit with the mission of increasing the openness, integrity, and reproducibility of research. With the vision that within the scholarly community, the process, content, and outcomes of research are openly accessible by default. With its core mission, COS maintains the Open Science Framework (OSF) to help researchers conduct research more rigorously and manage and share their work more openly.

The Open Science Framework is a paradigm shift for those wanting to effectively share their research process. It supports research in a multitude of ways. It supports the discovery of papers, data, and materials to inspire research projects as well as build on the work that others have started. It allows full collaboration of projects by enabling collaborators, giving them access to protocols and other research materials, while built-in version control tracks the evolution of the study. It allows for the storage of data, code, and other materials within the platform. It also allows for users to share their papers in OSF preprints, allowing others to easily cite your work.

2 Data Retrieval Today

2.1 ALI Usage

As briefly described in section 1.2, ALI is the tool used for preprocessing and delivery of studies and/or data about problem sets. Researchers begin by filling out a Google Form (Figure 1) with their email address and a problem set ID. If this is a user's first-time requesting data, the user is then prompted to fill out a second page aimed to sign off on contingencies. From there ALI inputs the requested fields in a google spreadsheet. While on the spreadsheet, an ASSISTments staff needs to validate the filled-out fields and decide how they should proceed. If the form seems good to go, they go ahead and run the query on the ASSISTment's database which automatically generates and emails out a form of the preprocessed data as well as a CSV file with the raw anonymized data. These CSVs could contain tens of thousands of rows and are far too large for more researchers to process by hand. Researchers must employ a variety of analysis methods. An example would be programmatic with a Jupiter notebook or R script.

Figure 1 Data Request Form

2.2 ALI Issue

The main problem that my IQP set out to tackle was the human intervention from the use of the Data Request Form. Having a staff member consistently validating and creating a query for the data is inefficient. As the ASSISTments foundation is currently growing at a rapid pace, the research testbed is also to follow, but the inefficiency of validation is not one built for scaling. The motif of an ASSISTments team member to step in and help researchers is not feasible or scalable for the long run.

3 Planning Process

3.1 Goal

The goal of my project was a simple but integral part of ASSISTments' E-Trials' expansion. The goal of sharing data openly and ethically is one that the ASSISTments' team wanted to move towards. Alleviating the need for human intervention was also an important part of building a product that is scalable for the future. This set the goal of the project to be simple. Create a web application that can share the requested data with researchers in a simple and trackable manner.

3.2 Considerations

Based on what I learned from my experience of working with the ASSISTments' database and discussing the downsides of ALI and data retrieval, I knew that I had to create an application that can easily be managed and alleviate efforts from staff.

3.2.1 Data Storage

A prominent issue was the way of distributing the data. ALIs' current system was that of emailing the data to researchers directly. While this might seem like an okay idea, there were some prominent problems with it especially if we wanted to alleviate human intervention. One of the problems is version control of the data. As the data changes, the requestor would need to continue to submit more forms for the requested project ID. Another problem that falls under version control is managing the collaborator and access to the original data.

3.2.2 Research Integrity

Another important consideration while tackling the outgoing goal was keeping integrity. This means knowing that the researcher is going to use the provided data in good faith and is credible. Having a drawn-out plan of how the researcher would conduct the experiment and use the data is important to ASSISTments as the data related is that of students.

3.2.2 Publication

Research publication and citation are also of importance. As a nonprofit foundation, ASSISTment's is do in part of the funding from national agencies. Research conducted and published with ASSISTment's data is an easy way to display the contribution and impact of the platform. This in turn can provide additional funding for the development of the platform and additionally contribute to the learning science community.

3.3 OSF A Tool Worth Considering

With the following considerations considered, a prominent tool used throughout the research community is the Open Science Framework from the Center for Open Science. OSF can provide a solution to a lot of the problems considered. OSF requires an account that researchers

need to create. This account falls under the Terms and Conditions that closely revolve around ASSISTment's. OSF also provides a dependable repository for data distribution and storage. Within the research integrity front, Pre-registrations and registrations are public tokens used to create the content of a project and how it is to be conducted. These tokens are persistent and reside within the framework, acting as a barrier for bad science practice. OSF also alleviates the need for citation when researchers are ready for publication, make it easy for them to cite.

3.4 The OSF API and Architecture

On top of filling out the box on the considerations we set out, OSF has an API that we can easily integrate through RESTful calls. The OSF Architecture revolves around components that can be thought of as nodes. All these components have an ID that OSF provides for them. I will briefly describe those of interest to this project.

3.4.1 Project and Component Nodes

The project and component nodes are what OSF calls 'Nodes' in their API reference. The idea of these nodes is the project and where it resides. There are a few different categories of these such as data, software, project, etc... We can stack nodes with children, as well as add contributors that can only view nodes below the node they get added to.

3.4.2 Users

The User is simply a user in the OSF database. We can search and validate users by RESTfully retrieving them with a GET command. An important part of users is that we can make them contributors to our Project and Component Nodes.

3.4.3 Registrations

Registrations are frozen, time-stamped versions of a project that cannot be edited or deleted. They are made public immediately or embargoed for up to 4 years. When creating a registration, researchers propose a hypothesis and how they are going to go about conducting their project.

3.5 Technology Used

The technology used for the web application falls in line with the technology used by the ASSISTment's ecosystem for "version 2.0". I used JAVA Spring to build a RESTFul API that connects our request with the OSF API, as well as the ASSISTments' database. For the frontend, I used Vue.JS to build a reactive application that includes a validation form and a description of the data. On top of Vue.JS, I used Vuetify to easily add a material design theme.

4 Design Process

4.1 OSF Project Design

Figure 2 Describes the OSF Process. At the top level, we have our ASSISTment's main project node. There we have ASSSITments staff members as contributors who can overlook any children's projects underneath. Under our parent project, we have the Researcher Requested Project. At the level, the researcher that requested the project is a contributor to it. They are not able to view the top-level project or any other nodes that are not below them. The Researcher Requested project requires a registration, that we can access. Once registration is provided, we can provide the requested data to that project with read-only capability. Top-level contributors trickle down as contributors towards projects. From there, researchers are also able to create their components, such as software ones or even documents. Once they want to publish their study, they are easily able to cite the data and ASSISTments'
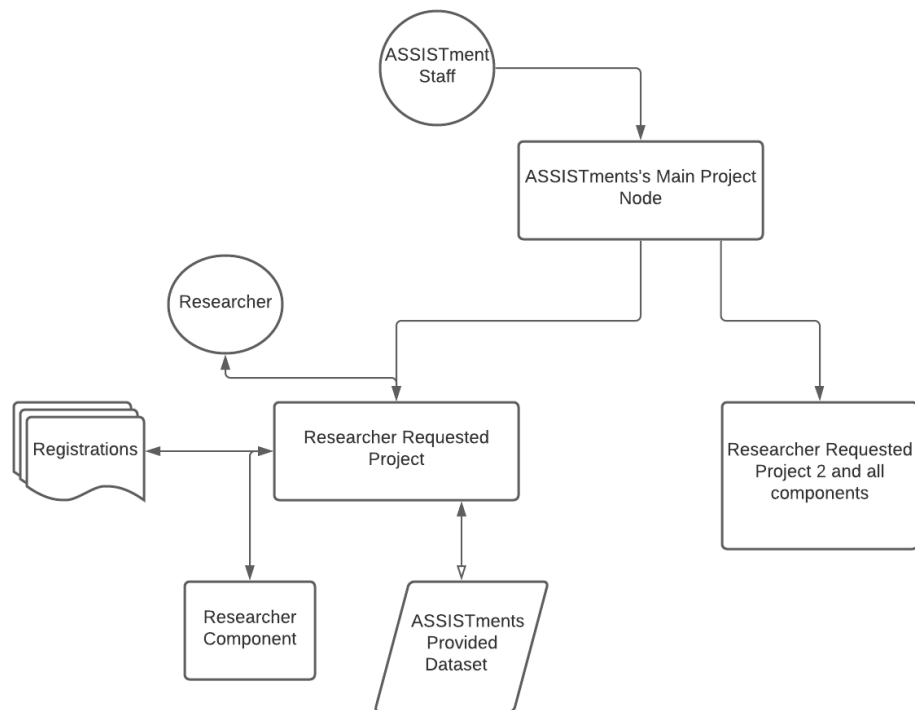

Figure 2 OSF Process

4.2 Data Fetcher Process

The process for the data fetcher API is as shown in Figure 3. A user fills out a form that is verified for correct credentials with our API. Once they have filled out the form with the required fields, our API creates the requested project on the OSF server. From there, the researcher needs to go back to OSF, and create a registration for the project detailing how they are going to conduct it. Once they have created the project. The researcher returns to our website and submits a second form with the project ID that we created for them. Once we confirm the registration, our API can go ahead and conduct a query on the database, and then submit the CSV files to the OSF repository.
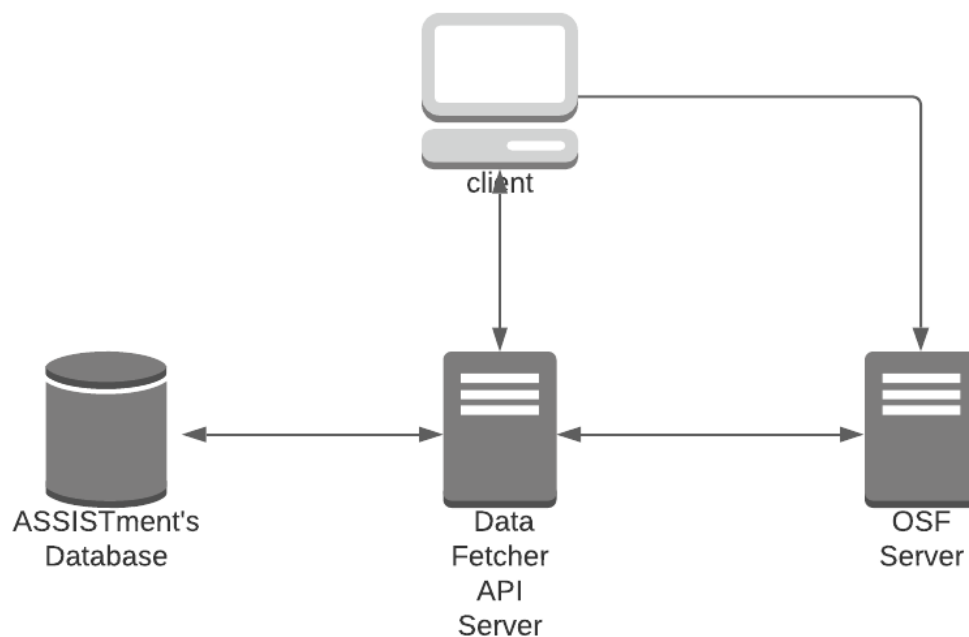
Figure 3 Data Fetcher Process

4.3 Data Fetcher Back End

Figure 4 shows the UML Design of the Data Fetcher API. When designing the API, I choose to settle for a RESTFul architecture. I used the model-view-controller software design pattern to implement the RESTful architecture. The model directly manages the data, logic, and rules of our application. The Controller accepts restful inputs and converts them to commands for the model view. The commands related to the OSF API or databases.

4.3.1 Controller

Our controller has 5 methods that get called from the client end. Three GET commands, two of them validating that the user or project is in our persistent database. These two commands take in a Path Variable in the URL call. As of right now they only return a Boolean, but for future iterations, if an admin panel is created, we can easily retrieve all the data related two the project or user. The third GET command gets our problem set IDs. Because there are hundreds of thousands of problem sets, and I don't want to load all of them directly in the beginning, the way that I tackle loading them is by querying the database once a client stops typing the ID. I only look for the IDs that start with the typed ID. From there we have 2 POST commands. When a researcher first fills out his form, the createOSFProject gets called with all the form data in the request body. From there we can create the project on the OSF ecosystem. The second POST command confirms the registration, starts the query on the ASSISTment's database, and submits it to the OSF data repository.
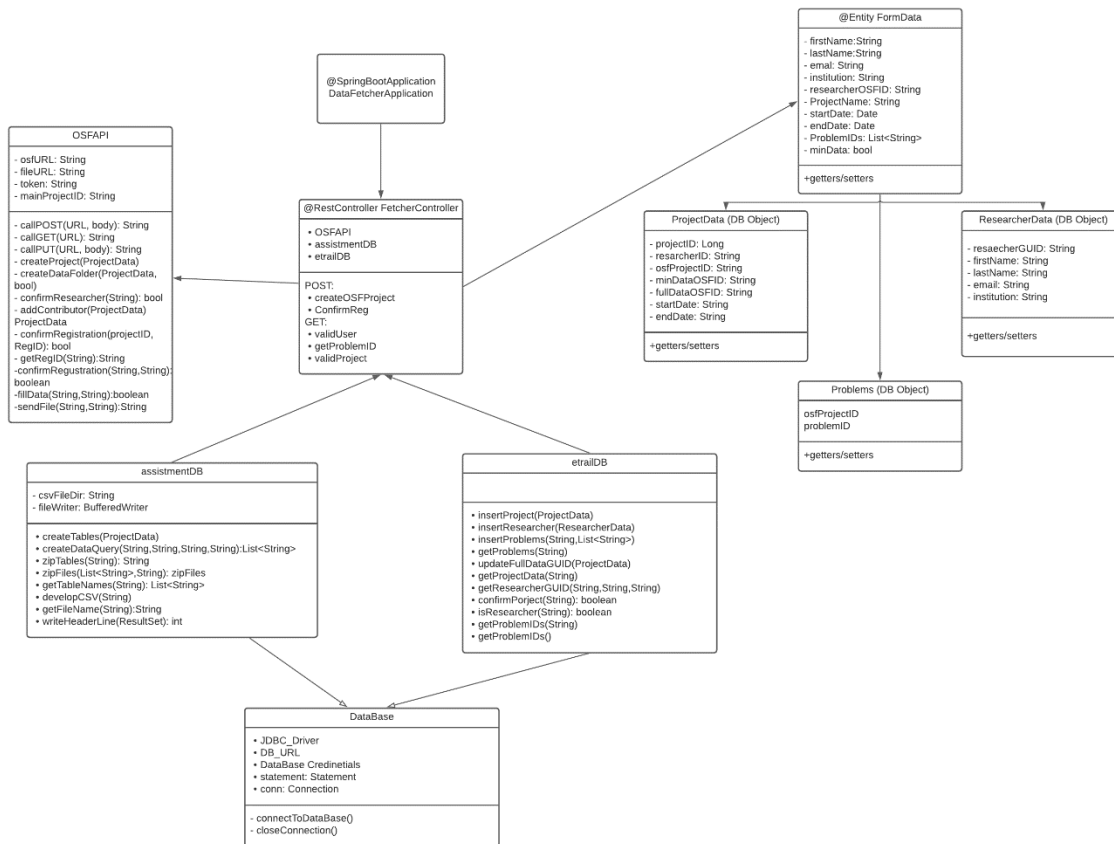
Figure 4 Data Fetcher UML

4.3.2 OSF API Object

The OSF Object the object that handles any communication with the OSF Servers. Based on the OSF architecture that I described above, this object's main purpose is to adhere to the OSF's RESTFul protocol. The object has three main purposes.

The first is two POST data to the API. Within the architecture, we post 3 items. Firstly, we create a project by posting a child node to our parent node. Secondly, we create a data project node and attach it as a child by posting it to the researcher's project. Lastly, add the researcher as a contributor to those projects by posting the researcher's data.

The second is PUT. The correlated PUT method for the API is towards putting a zipped file on the OSF data repository. After turning the requested query into a zip file of CSVs, we push the bytes towards the repository for the researcher to get.

The third purpose is to call GET requests. The get requests are used to confirm research-filled data for confirmation before proceeding.

4.3.3 E-Trials Database Object

The E-Trials Database object inherence from a parent database object that abstracts some filler code. After that, the E-Trials database is used to persistently store data on our database. Figure 5 described the data structure. All the items starting with OSF correlate to the IDs that OSF gives for their nodes. The Entities in figure 4 represent these data tables as Java objects to maintain RESTful protocol.
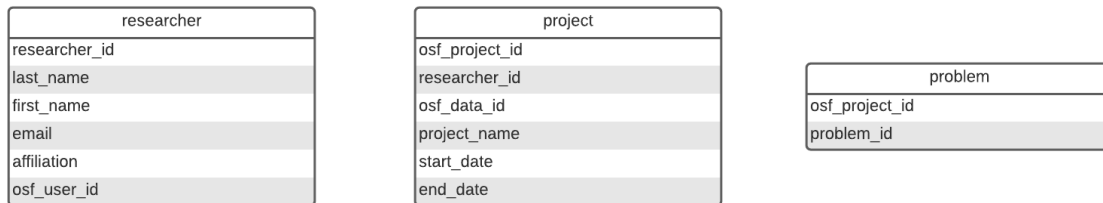
researcher
researcher_id
last_name
first_name
email
affiliation
osf_user_id

project
osf_project_id
researcher_id
osf_data_id
project_name
start_date
end_date

problem
osf_project_id
problem_id

Figure 5 Persistent Data

4.3.4 ASSISTments Database Object

The ASSISTments Database Object also inherits from the parent database object. This database object has two main purposes. Querying and creating all the tables with the requested data and creating those tables as CSVs and zipping them up. For reference, figure 6 shows the ASSISTments' data structure that we query.

District_Details

district_id
location
opportunity_zone
locale_description

teacher_details

teacher_id
district_id
account_creation_date
assignment_reports_viewed_fraction
student_reports_videwed_fraction
graded_open_response_fraction
open_response_comment_fraction

problem_logs

log_id
student_id
assignment_id
problem_id
start_time
time_on_task
answer_before_tutoring
fraction_of_hints_used
attempt_count
answer_given
problem_completed
correct

assignment_details

assignment_id
class_id
release_date
due_date
assignment_type
assigned_student_count
started_student_count
completed_or_mastered_student_count
problem_count
mean_correct
mean_time_on_task

class_details

class_id
teacher_id
class_creation_date
student_count
problem_sets_assigned
skill_builders_assigned

student_log

log_id
student_id
assignment_id
problem_id
action
timestamp

probelm_details

problem_id
content_source
skills
problem_type
tutoring_types
sudent_answer_count
mean_correctness
mean_time_on_tak

student_details

student_id
class_id
account_creation_date
started_problem_sets_count
completed_problem_sets_count
started_skill_builders_count
mastered_skill_builders_count
answered_problems_count
mean_problem_correctness
mean_problem_time_on_task

teacher_logs

teacher_id
assignment_id
student_id
problem_id
action
timestamp

assignment_logs

log_id
student_id
assignment_id
start_time
mean_correct
time_on_task
assignment_completed

Figure 6 ASSISTments Data Query

5 Web Application

5.1 Inspiration

The inspiration of the front-end theme revolves around the assistments.org design aspect. It also borrows inspiration from the original google form.

5.2 Data Fetcher Front End

Their web application contains front-end two pages.

5.2.1 Home Page

The first page is the home page (Figure 7). On the home page, we are greeted with a navigation bar on the top. Right under we have a form with two options, either confirming
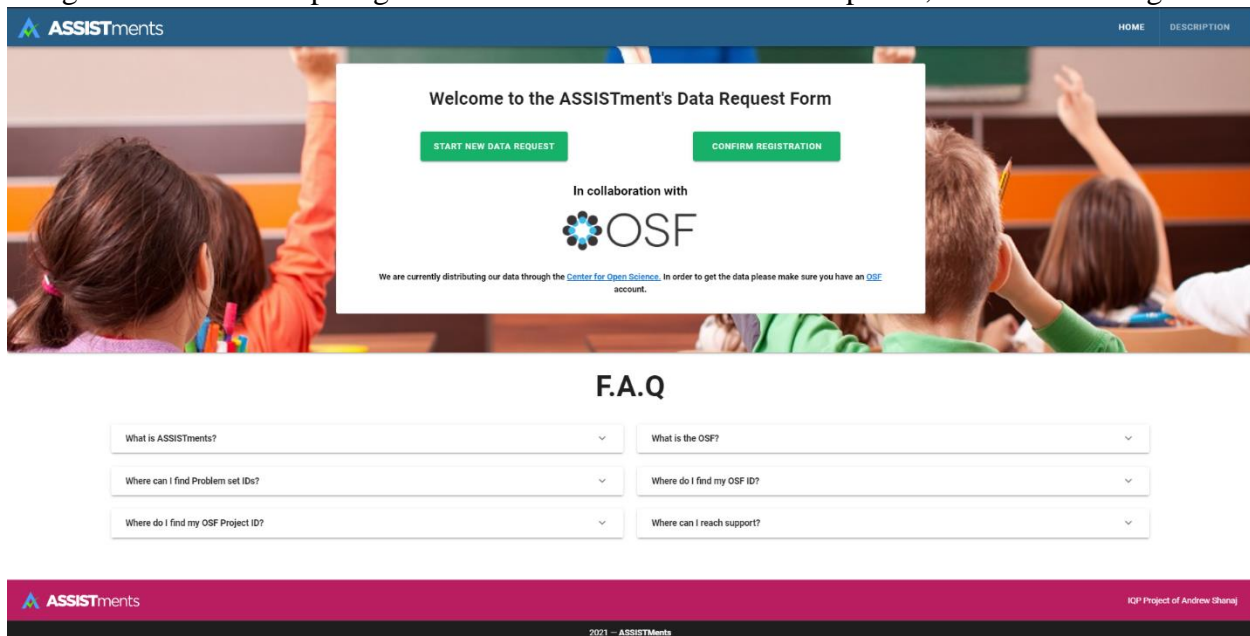

Figure 7 Home page

registration or starting a new data request. Bellow that we have a frequently asked section that is currently a filler section. Lastly, we have a bottom bar adding a warm purple accent to our page.

5.2.2 Description Page

The Description Page (Figure 8) is used to describe the data that I referenced in Figure 6. It is a scrollable page the describes all the tables and how they are connected. For future work, the Description page can get a lot added to it. I think a few nice additions would include a downlink link to a small percentage of data so that researchers can play around with the data. As well as a jupyter notebook view of how researchers can manipulate and use the data. I also think a section of past coded projects with the data would be nice to model ways the data has been used before.

Figure 8 Description Page

5.3 Start a New Data Request

Figure 9 shows the first part of our data form which correlates to the researcher persistent database. The form automatically validates itself, requiring it to be filled, limiting the characters, and requiring the correct email format.

Figure 9 Research Form Fields

We also validate the OSF public profile ID by directly calling our API. If the OSF ID is not valid, we inform that the client as shown in Figure 10. We also throw an error when the client tries to continue without filling out all the data.
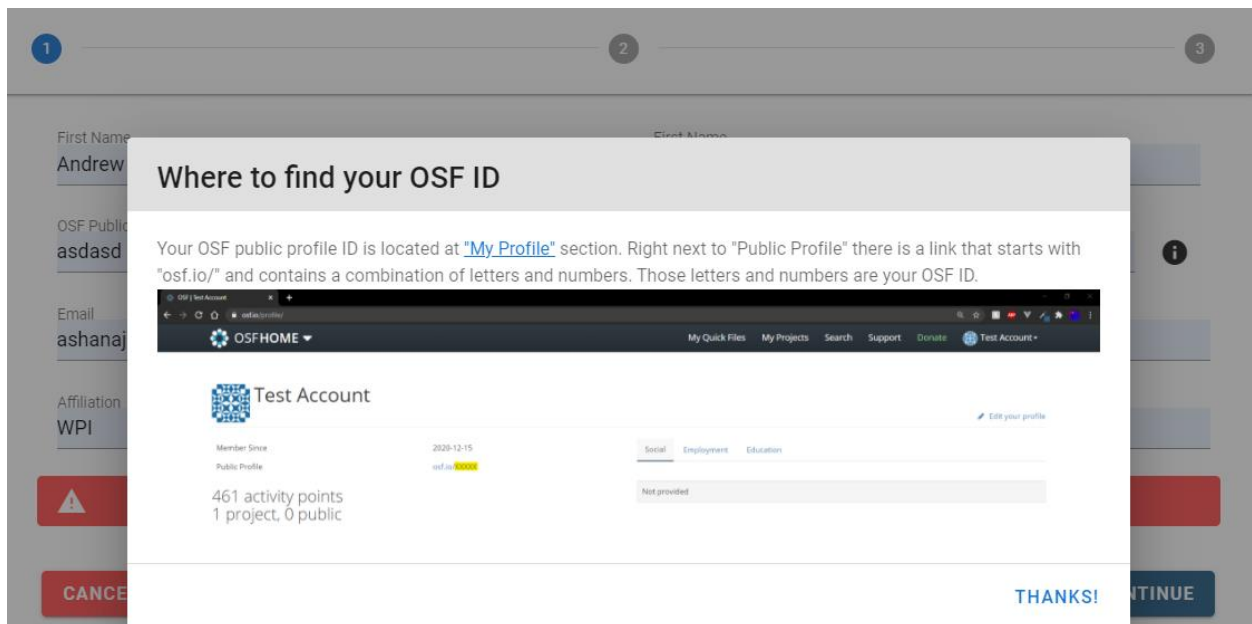

Figure 10 Invalid OSF ID

If the client doesn't know what an OSF Public Profile ID is the information icon to the right will pop up where to find the ID as displayed in Figure 11.

Figure 11 Where to Find your OSF ID

Once the researcher provides the correct ID, he can move on to the second part of the form which is shown in Figure 12. On the second page of the form, the researcher can provide a Project Name, the Project Set IDs requested, start date, and end date. The second page of the form also calls our API and does validation. Project Ids that are part of the ASSISTments database are only available for processing by validating the problem set IDs. Another verification needed is that the Start date is before the end date.
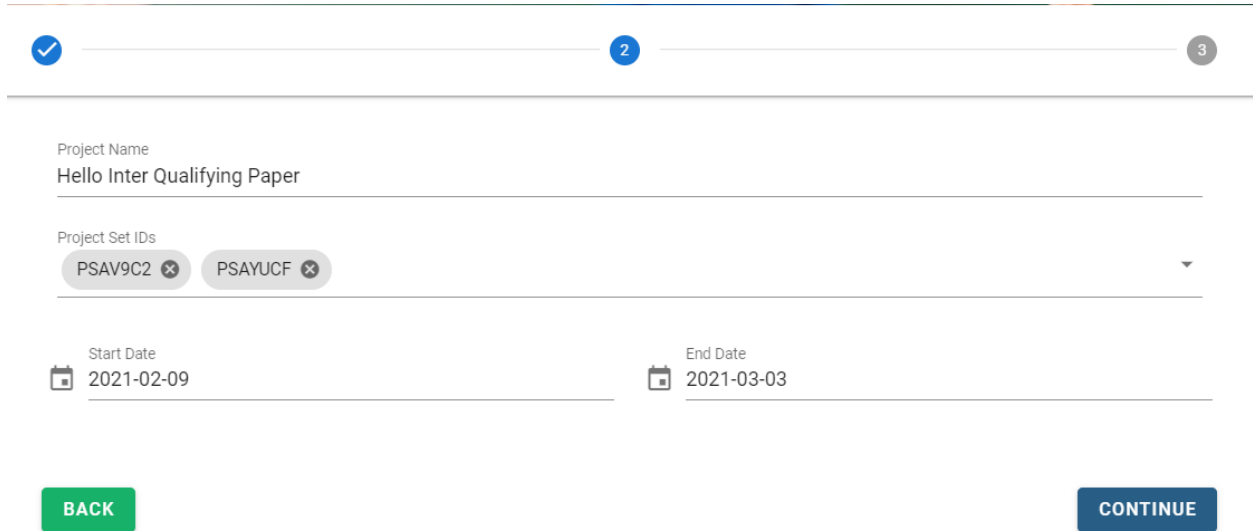
Figure 13 Second Page of Form

Once we continue from there, the form calls the API again and POSTs all the data we just received which then creates the account.

Next Steps

We are currently creating your project. Once the progress bar loads your project structure will be made. All you have to do from here on is create a project Registration and come back with the project ID. From there on we will distrubute all of the requested data.
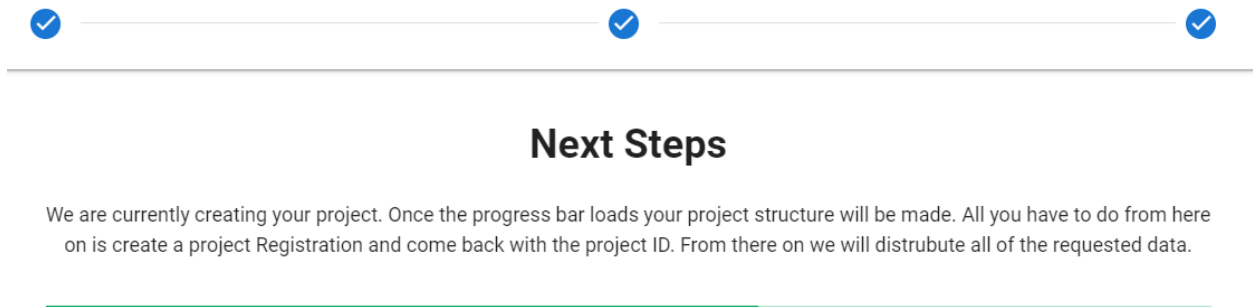
Figure 12 Final Page

5.4 Confirm Registration

The confirm a registration form is a simple form that only requires the Project ID that we created for the researcher. Once we have the Project ID, we can search through the OSF API and get information about the registrations attached to our current project.
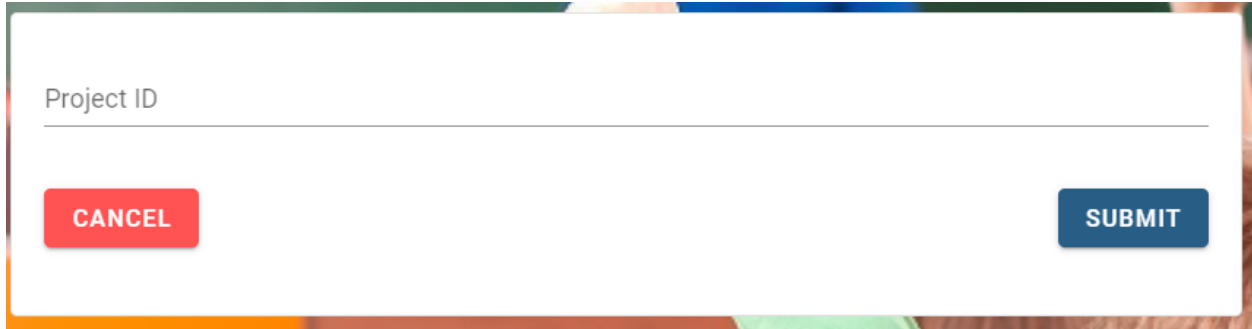
Figure 14 Registration Confirmation

6 OSF View

From the ASSISTments' account view, we can view our main parent project and all the projects that we created for different researchers. The main parent is the ASSISTments IQP Data project. The parent project houses two children's projects. One of them is the one that we just created. There is only one contributor to this project because the OSF ID I used the form is for the IQPShanaj account. The "P vs N Mistake Msgs" project one that I created a little while ago demonstrating the web application. That project also has the ASSISTMent's Complete Project Data repository which was created after confirming the registration. We also notice that the projects have Heffernan as a contributor.

Name ∧ ∨	Contributors	Modified ∧ ∨
− 📦 ASSISTments IQP Data	IQPShanaj	2 minutes ago
📦 Hello Inter Qualifying Paper	IQPShanaj	14 minutes ago
− 📦 P vs N Mistake Msgs	IQPShanaj, Heffernan	a month ago
🗄 ASSISTMent's Complete Project Data	IQPShanaj, Heffernan	a month ago

Figure 15 OSF project components.

Figure 16 shows the researcher's view that includes where the zipped file with the requested data.

Zip File:

Download the .zip file to view the contents.

File Name	Modified	Size
adets_x8jhk_Export.csv	2021-02-23 20:00:32	1.5KB
alogs_x8jhk_Export.csv	2021-02-23 19:54:30	16.6KB
cdets_x8jhk_Export.csv	2021-02-23 19:54:30	861B
pdets_x8jhk_Export.csv	2021-02-23 20:00:54	9.0KB
plogs_x8jhk_Export.csv	2021-02-23 20:01:04	63.5KB
sdets_x8jhk_Export.csv	2021-02-23 19:54:30	16.7KB
slogs_x8jhk_Export.csv	2021-02-23 19:54:30	333.1KB

Figure 16 Researcher View

7 Future Work

7.1 Database Query

The database query used for this project was written by Ethan Prihar to query a large chunk of the data for our published poster from a graduate class that I had the honor of being apart of. I edited the query to specify it towards the goal of selecting certain problems at a certain time frame. One of the big problems with this query is that it takes quite a while to run. The ASSISTments database has over 500 million data points relating to user actions. The query queries the actions and aggregates data and created the previously showed data tables. This has resulted in long query time, on top of the problems correlated to running on the development database, that crashed quite often.

7.2 E-Trial Integration

The main purpose of this project is to demonstrate the ability to use OSF to verify and share our dataset in an open and scalable manner. As the project and proof of concept are getting passed down my experience of working on this web app has brought up a few new considerations that would be nice to get tackled as the new E-Trial back end is getting developed.

7.2.1 Registration Submission

I believe that building the process of the application fully in-house would very much improve the scalability of E-Trials. Instead of having to fill out the registration on OSF, researchers fill it out on our web app. OSF registrations can vary from strict to very open-ended. By creating a strict in-house registration process and modifying the open-ended OSF registration we would be able to submit this registration openly to the science community as well as maintain the data and keep a strictness to who we pass the data to.

7.2.2 Registration Verification

The current version of the application relies on the integrity of the user that signed up and read the Terms and Conditions of OSF. I believe that the usage and distribution of student data that can be anonymized need to be a lot more considered and controlled. Continuing from the last point of the in-house registration submission, I think that a natural language processing pipeline should be implemented on sections of the registration submission. With an NLP pipeline, we would be able to detect registrations that might raise an eye without the need for an ASSISTments staff checking up on submitted registrations.

7.2.3 Admin Panel

For scalability purposes, I believe that an admin panel containing, all the persistent data collected would be a good tool to give us control of our data and provide version control if needed. A graphical user interface would easily allow for staff to be able to manage E-Trials, without the worry of having to code, query data, or scour OSF. It would bring even more in-house and control to the ASSISTments E-Trial's testbed.

7.3 Project Sharing

To progress learning sciences. By using the right tools to enable collaboration and project version control, I think the base of supporting collaboration and openness is off to a great start. I think an important feature that can push discoveries would be the ability to easily share experiments between researchers. Setting up a page and ecosystem of allowing researchers to build off the work of others and easily find projects related to theirs that use the ASSISTments' data would be able to alleviate a lot of time wasted for researchers and allow them to design better and more solidified experiments.

8 Conclusion and Lesson Learned

At the onset of this IQP, the goal was to implement a method of providing researchers with data in an ethical manner. I was met with a problem to solve with little pre-determined direction. However, after significant effort alongside my advisors, I was able to land on a solution that could pave the future for E-Trials and contribute to the learning science community. This process has taught me a lot about software development and project management. It has taught me how to bring an idea to life by creating a web application and the usefulness of RESTful principles. It has also taught me that development is truly a group effort. Throughout the conversations with my advisors and meetings partook the importance of collaboration led my IQP project towards a successful ending.

Throughout my two terms, the process of creating the application wasn't always smooth. Tackling technical challenges and persisting through problems was an important lesson learned. Another important lesson learned is the importance of project planning. Throughout my time, I found myself going back and making changes, especially near the end. This was due to my newly acquired knowledge of the MVC design pattern and RESTful principles. Some minor changes could've been avoided by creating strong UML and use-case diagrams at the start and going over them with my advisors. Something that I will be sure to slow down and work on for my next major project.

As I am ending my IQP, I hope my contribution to the ASSISTments team and more importantly E-Trials has been useful. As the new architecture is getting developed, I hope some of my future work get considered, and that E-Trials can develop into a self-sustained application that pushes the bounds of the learning community and sets a standard for RCTs within all online learning platforms.

As for Dr. Ostrow and Mr. Emberling, it was truly a pleasure to contribute to your project, and look forward to seeing the completion of E-Trials.

References

"ALI's Raw Data." *Etrialstestbed.Org*, https://www.etrialstestbed.org/research/research-process/4-analyze-data/ali-s-raw-data. Accessed 24 Mar. 2021.

ASSISTments. "ASSISTments TestBed Resource Guide." *ASSISTments TestBed Resource Guide*, 2015, http://www.assistmentstestbed.org/.

"ASSISTments | Free Education Tool for Teachers & Students." *ASSISTments | Free Education Tool for Teachers & Students*, https://new.assistments.org/. Accessed 24 Mar. 2021.

"EdTech & Learning Science Research & Efficacy | ASSISTments." *ASSISTments | Free Education Tool for Teachers & Students*, https://new.assistments.org/research. Accessed 24 Mar. 2021.

"Introduction — Vue.Js." *Vue.Js*, https://vuejs.org/v2/guide/. Accessed 24 Mar. 2021.

"OSF." *OSF*, https://osf.io/. Accessed 24 Mar. 2021.

says, Ravan. "What Is REST." *REST API Tutorial*, https://restfulapi.net/. Accessed 24 Mar. 2021.

Science, Center for Open. "Center for Open Science." *Center for Open Science*, https://www.cos.io/. Accessed 24 Mar. 2021.

---. "Our Mission." *Center for Open Science*, https://www.cos.io/about/mission. Accessed 24 Mar. 2021.

---. "The Open Science Framework." *Center for Open Science*, https://www.cos.io/products/osf. Accessed 24 Mar. 2021.

"Spring | Microservices." *Spring | Home*, https://spring.io/microservices. Accessed 24 Mar. 2021.

"Why You Should Be Using Vuetify — Vuetify." *Vuetify*, https://vuetifyjs.com/en/introduction/why-vuetify/. Accessed 24 Mar. 2021.

Heffernan, N.T., Heffernan, C.L. The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *Int J Artif Intell Educ* **24,** 470–497 (2014). https://doi.org/10.1007/s40593-014-0024-x