moocRP: An Open-source Analytics Platform

Article · March 2015

DOI: 10.1145/2724660.2724683

Zachary A Pardos
University of California, Berkeley

moocRP: An Open-source Analytics Platform

Zachary A. Pardos
School of Information / School of Education
UC Berkeley
zp@berkeley.edu

Kevin Kao
EECS
UC Berkeley
kkao@berkeley.edu

ABSTRACT

In this paper, we address issues of transparency, modularity, and privacy with the introduction of an open source, web-based data repository and analysis tool tailored to the Massive Open Online Course community. The tool integrates data request/authorization and distribution workflows as well as a simple analytics module upload format to enable reuse and replication of analytics results among instructors and researchers. We survey the evolving landscape of competing data models, all of which can be accommodated in the platform. Data model descriptions are provided to analytics authors who choose, much like with smartphone app stores, to write for any number of data models depending on their needs and the proliferation of the particular data model. Two case study examples of analytics and interactive visualizations are described in the paper. The result is a simple but effective approach to learning analytics immediately applicable to X consortium institutions and beyond.

Author Keywords

Open learning analytics; modularization; MOOC; dashboards; visualizations; reproducible research; edX

ACM Classification Keywords

K.3.1 Computer Uses in Education: Distance learning

INTRODUCTION

Matters of practicality and principle face the communities of educational data mining and learning analytics with both matters coming into particular relief on the topic of Massive Open Online Course data collection, management, distribution, and analytics. We introduce a new tool called moocRP, an open-source[1] analytics platform for the greater MOOC community that situates it self at the intersection of these issues.

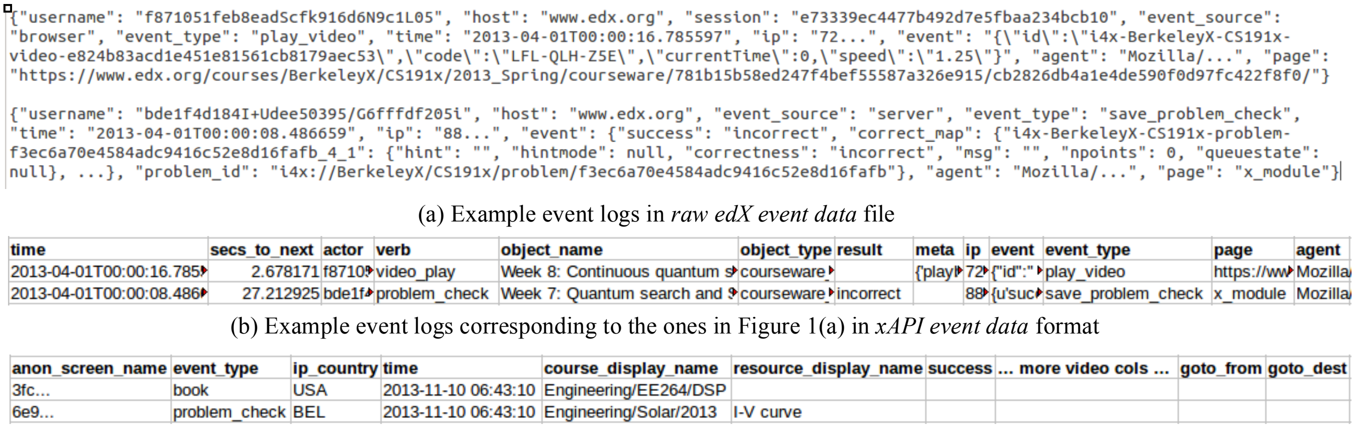
We will start with the issues of practicality. The non-profit higher education platform, edX, now has 36 partner universities offering or set to offer free courses online. The majority of these universities are not well prepared to handle the data they receive from edX, nor what to do with it once received. Coursera and other MOOC providers put stakeholders in a similar state of disorientation with the nuanced distinction that Coursera and others deliver data direct to the instructor of record while edX delivers data at the feet of the institution that must then co-ordinate distribution to the appropriate parties. The learning analytics community can provide much needed guidance on the cutting edge of what is actionable in the data and how it can be leveraged by instructor and student [16] towards the betterment of the learning experience. Our tool can help expand the impact of learning analytics by enabling its use among a quickly expanding new cohort of online learners and instructors. The tool also provides answers to the practical questions of how to prepare the data, grant users access to it securely, and scrutinize the data for instructionally actionable information.

Matters of principle relevant in education data analysis and distribution include a) transparency – what data is being collected, how is it being represented, and what exactly are the various technical components of moocRP doing in the background to manipulate these data behind the analytic dashboard. To address this, our tool is open source, as is the technical design of the pipeline that serves analytics to its users. New analytics and visualizations that are uploaded must also have the source viewable to all users to function on the system. In order for a data model to be imported into the system, it must describe the data elements it exposes and this description is available both to the authors and consumers of the analytics. The next principle is b) modularity – in order to foster an inclusive community of analytic contributors, analytics must be easily incorporated into the tool. We specify a simple but feature rich modular format that lowers the technical and time requirements to contribute to analytics in higher-ed. Lastly, c) privacy is of particular concern in the current climate of big data, surveillance, and ethics surrounding the beneficent use of sensitive education information [5, 17].

L@S 2015, March 14–18, 2015, Vancouver, BC, Canada.

ACM 978-1-4503-3411-2/15/03…$15.00.
http://dx.doi.org/10.1145/2724660.2724683

[1] https://github.com/CAHLR/moocRP



```
{"username": "f871051feb8eadScfk916d6N9c1L05", "host": "www.edx.org", "session": "e73339ec4477b492d7e5fbaa234bcb10", "event_source":
"browser", "event_type": "play_video", "time": "2013-04-01T00:00:16.785597", "ip": "72...", "event": "{\"id\":\"i4x-BerkeleyX-CS191x-
video-e824b83acd1e451e81561cb8179aec53\",\"code\":\"LFL-QLH-Z5E\",\"currentTime\":0,\"speed\":\"1.25\"}", "agent": "Mozilla/...", "page":
"https://www.edx.org/courses/BerkeleyX/CS191x/2013_Spring/courseware/781b15b58ed247f4bef55587a326e915/cb2826db4a1e4de590f0d97fc422f8f0/"}

{"username": "bde1f4d184I+Udee50395/G6fffdf205i", "host": "www.edx.org", "event_source": "server", "event_type": "save_problem_check",
"time": "2013-04-01T00:00:08.486659", "ip": "88...", "event": {"success": "incorrect", "correct_map": {"i4x-BerkeleyX-CS191x-problem-
f3ec6a70e4584adc9416c52e8d16fafb_4_1": {"hint": "", "hintmode": null, "correctness": "incorrect", "msg": "", "npoints": 0, "queuestate":
null}, ...}, "problem_id": "i4x://BerkeleyX/CS191x/problem/f3ec6a70e4584adc9416c52e8d16fafb"}, "agent": "Mozilla/...", "page": "x_module"}
```

(a) Example event logs in *raw edX event data* file

time	secs_to_next	actor	verb	object_name	object_type	result	meta	ip	event	event_type	page	agent
2013-04-01T00:00:16.785▸	2.678171	f8710▸	video_play	Week 8: Continuous quantum s▸	courseware▸		{'play▸	72▸	{"id":"▸	play_video	https://ww▸	Mozilla
2013-04-01T00:00:08.486▸	27.212925	bde1f▸	problem_check	Week 7: Quantum search and ▸	courseware▸	incorrect		88▸	{u'suc▸	save_problem_check	x_module	Mozilla

(b) Example event logs corresponding to the ones in Figure 1(a) in *xAPI event data* format

anon_screen_name	event_type	ip_country	time	course_display_name	resource_display_name	success	... more video cols ...	goto_from	goto_dest
3fc...	book	USA	2013-11-10 06:43:10	Engineering/EE264/DSP					
6e9...	problem_check	BEL	2013-11-10 06:43:10	Engineering/Solar/2013	I-V curve				

(c) Example event logs in VPOL format

Figure 1. Examples of event logs represented in different data models

With moocRP, the analytic can be brought to the data instead of the other way around. Institutions needing to keep data close to the chest can control its distribution but still open the window to education researchers by allowing analytic module collaborations and granting researchers access to a level of acceptable transformed aggregate data or sharing the analytics results returned by the module. Similarly, individuals possessing data do not need to upload it to a centralized location to have analyses run but can instead spin up their own local instance of moocRP and import their desired analytic modules to run locally. These issues of transparency, openness, modularity, and privacy were outlined in the open learning analytics vision document [1] written by leaders of the Society for Learning Analytics Research (SOLAR).

In last year's Learning at Scale, authors suggested how modern visualization techniques could aid instructors in understanding how student activity corresponded to course success [25]. Other work categorized visualizations into groups consisting of; quantitative information (assignment grades, demographics, age) 2) qualitative information (discussions, course surveys), and 3) "real-time" visualization while the course is running. The moocRP platform allows for the visualization of all three concepts through its analytic module system. In the case studies shown in section 2.1 we discuss two proofs of concept analytics that provide interactive visualizations that are the confluence of the categories described above. These analytics, uploaded to moocRP, serve as examples of what types of modules an instructor could have access to through the usage of moocRP. With constantly growing amounts of MOOC data, visualizations and analytic modules shared among instructors and researchers can be an effective way to explore and analyze a course.

Related Work on Data Models

The pedagogical developments in online education are still changing form and function and thus the model used to represent the data coming from these new pedagogical interfaces is ever changing. In this section, we provide an overview of a few of the most visible data models currently in play in the MOOC ecosystem. All of these models are compatible with moocRP, which takes a data model agnostic approach in its design.

EdX tracking log and database model

The most primitive form of data that is obtained from edX by partner schools is the tracking and database logs [18] hosted by Amazon S3 storage service. The raw edX data consists of a set of compressed and encrypted files that contain *event data* of all courses offered by an institution. Every course is served by multiples servers. As a result, to obtain complete *event data* for one course, data must be aggregated from multiple servers. While this is getting a bit into the weeds of data wrangling, it's a reality that can be overwhelming and keep an institution and its learners from reaping the benefits of analytics. The moocRP platform has built into it data cleaning and processing scripts that manage these processes from data ingestion, decryption, through to viewing the analytics on the web.

Figure 1(a) shows an example of event logs in an event data file. One *event data* file contains all event data for all courses within one-day period in JSON format. The *database data* of a course consists of a set of SQL and mongo database files which include information about students, demographics, and survey answers provided by students, as well as grades and certification statuses.

xAPI data model

HarvardX has adopted a data format, developed by Jim Waldo [19], to package, analyze, and manipulate edX data [19]. The tool converts *edX tracking log data* into more tangible csv files based on the xAPI standard [20] (formerly Tin Can) created by Advance Distributed Learning (ADL). We call the output csv file, the *xAPI event data*. Figure 1(b) shows the same event logs as in Figure 1(a) in *xAPI event data* format.

Stanford VPOL data model

The Stanford Vice Provost Office for Online Learning (VPOL) provides MOOC data for researchers and instructors in a multi-file format [21]. It supports data derived from NovoEd, Coursera, and edX platforms. For edX data, the platform uses scripts developed by Andreas Paepcke [22] to transform *raw edX event data* and *raw edX database data* into relational tables. For example, *EventXtract* table contains event data of all courses. Figure 1(c) shows an example of an *EventXtract* table. Observe that the two logs are from different courses. Other tables such as *AcitivityGrade, Demographics,* and *Forum post* are generated from *edX database data*.

MOOCdb

The MOOCdb relational database schema [3, 23] organizes the data with respect to four different interaction modes: observing, submitting, collaborating, and feedback. Figure 2 shows the schema of the observing mode tables.

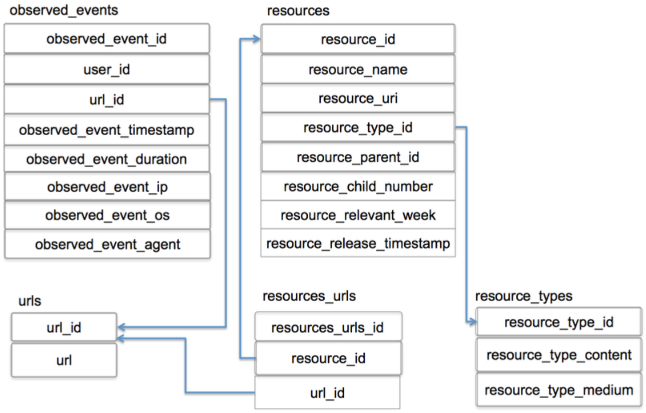


Figure 2. MOOCdb's observing mode tables

LearnLab DataShop model

In the space of data repositories, the Pittsburgh Science of Learning Center's DataShop [2] has hosted an abundance of both public and private datasets. While it has not been a MOOC repository, it is widely used in the Intelligent Tutoring Systems community, and we would be remiss to not acknowledge its impact and use in the learning community. DataShop hosts datasets as a service to the community, largely from the Cognitive Tutor [9], which was born out of the same lab, as well as from other tutoring systems such as the Andes Physics Tutor [4], the Open Learning Initiative [10] and a web-based mathematics platform, ASSISTments [7], among others. These tutoring systems are based on mastery learning [12] and the effective practice of providing immediate feedback [11] in achieving that mastery. As such, these tutoring systems are problem-solving centric, with most of the tutors breaking problems down in to a collection of steps to be answered with hints available as the primary pedagogical device. The singular data model underlying the DataShop was therefore designed to be problem centric, with each row in the data format representing an answer to a step. In MOOCs, where problem solving takes a back seat to lecture videos in student time on task [14], a wider variety of behavior and interaction is desirable to be captured. The DataShop data model was not designed to incorporate this variety of information nor social network information from the many social contexts in which learning takes place [13]. This is not to say that the DataShop model falls short of serving its purpose, or that a globally compatible data model should be sought, but rather that a single data model is not sufficient to accommodate the numerous philosophies of learning represented by various platforms of learning.

ANALYTIC MODULE CASE STUDY

In this section, we will describe some proof of concepts implementations of analytics modules[2] into the moocRP system.

Integration of analytics modules into moocRP is relatively straightforward and requires minor modifications in which the analytics source code is structured. The basic structure of an imported analytics package follows this package structure:

- `main.html`
- `/css`
- `/js`

The main display of the analytics module is placed in `main.html`, where the HTML will be rendered in a container on the analytics page of moocRP. Any script dependencies and styling dependencies must be declared in this file as well – this allows moocRP to have a single location to rewrite dependencies in a more compatible fashion. CSS files and JavaScript files are placed in the `/css` and `/js` folders, respectively.

Another minor change that needs to be made is how the data is read into the analytics module. The analytics module should access the dataset using "`<%= dataset %>`", which represents a dictionary that holds the various files associated with a data model.

To access each data model's files in an HTML file, we can use JavaScript and write the following code:

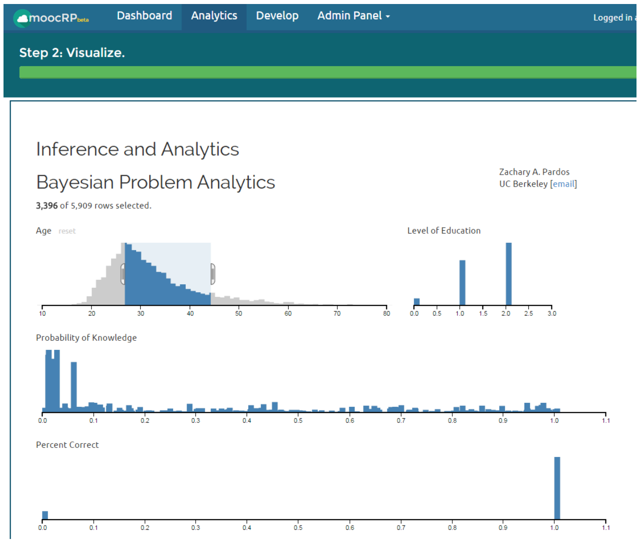
```
var dataFiles = {}
<% for (var dataFile in dataset) { %>
  <% if (dataset.hasOwnProperty(dataFile))
{ %>
    dataFiles["<%= dataFile %>"] = <%-
JSON.stringify(dataset[dataFile]) %>;
  <% } %>
<% } %>
```

Then, to access any of the data model's file contents, simply select the file using the name of the data model file as the key to the dictionary. Raw content will be stored in the dictionaries for parsing and manipulation by the analytics modules.

[2] http://www.github.com/CAHLR/moocRP_visualizations

Bayesian Network Knowledge Analysis

In this first case, the Bayesian Knowledge Tracing algorithm [4] is employed to assess student current and prior knowledge for each given problem in the course (see Figure 3). The visualization is presented along with age and level of education information so that an instructor may inspect for which demographic students lacked the requisite prior knowledge. Additional attributes can be added to this visualization including country, survey information, and outcomes for previous parts of the course.



Step 2: Visualize.

Inference and Analytics

Bayesian Problem Analytics

Zachary A. Pardos
UC Berkeley [email]

3,396 of 5,909 rows selected.

Figure 3. This analytic module is based on BNT.

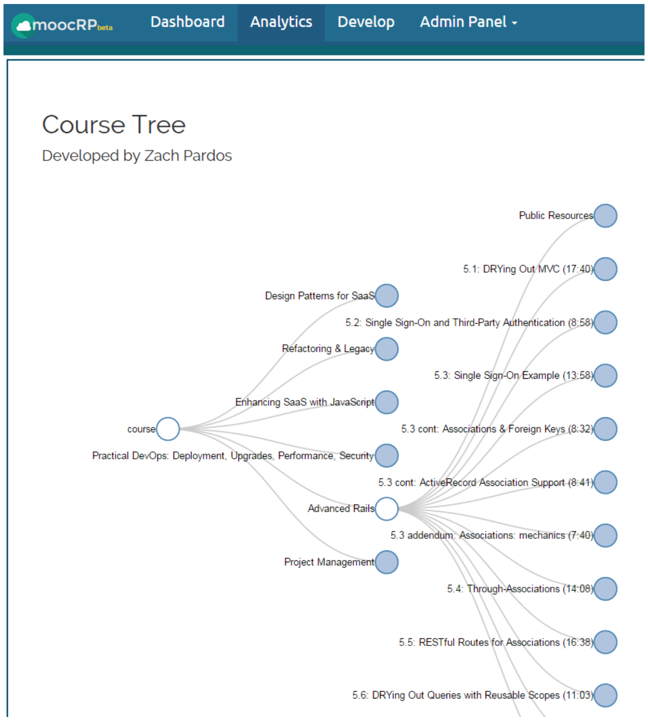
To integrate this module into moocRP, we simply had to move all JavaScript and CSS dependencies that were not in the main HTML file into their respective folders as defined earlier. In the original HTML file, `d3.csv(“filepath”, function)` was used to read the CSV contents of a file to perform the D3 visualizations. With moocRP, instead of directly reading the file from disk using a function like `d3.csv`, we can receive the CSV file contents from the data model by including the code previously mentioned with the addition of the snippet `var data = dataFiles[“output.csv”]`. Then we can use the function `d3.csv.parse` on the `data` variable, which contains the raw contents of the file. With the modifications complete, we can then zip up the `main.html` file along with the `css/` and `js/` folders and upload the archive to moocRP to be shared and applied on other datasets.

A simple example of how the data can be used with D3:

```
var raw_data = dataFiles["bnt.csv"]
                   .replace(/\\n/g, "\n");
var flights = d3.csv.parse(raw_data);
flights.forEach(function(d, i) {
    // do something to data rows
}
var flight = crossfilter(flights), ...
var chart = d3.selectAll(".chart")
    .data(charts)
    .each(function(chart) { ... });
```

Course Structure Visualizer

In the second case study, we present a module for visually inspecting the flow and course structure of a course (see Figure 4). The tree is created from parsing of the `course_structure.json` file and visualized using an expandable tree layout in D3. While this visualization shows only an interactive course tree structure, it can be used as the basis for navigating to analytics about different elements of the course, such as selecting the problem to receive the aforementioned Bayesian problem analytics for.


moocRPbeta
Dashboard
Analytics
Develop
Admin Panel ▾

Course Tree

Developed by Zach Pardos

course
Design Patterns for SaaS
Refactoring & Legacy
Enhancing SaaS with JavaScript
Practical DevOps: Deployment, Upgrades, Performance, Security
Advanced Rails
Project Management
Public Resources
5.1: DRYing Out MVC (17:40)
5.2: Single Sign-On and Third-Party Authentication (8:58)
5.3: Single Sign-On Example (13:58)
5.3 cont: Associations & Foreign Keys (8:32)
5.3 cont: ActiveRecord Association Support (8:41)
5.3 addendum: Associations: mechanics (7:40)
5.4: Through-Associations (14:08)
5.5: RESTful Routes for Associations (16:38)
5.6: DRYing Out Queries with Reusable Scopes (11:03)

Figure 4. An analytic module used to visualize the course components of CS169 Software Engineering from BerkeleyX.

Integration of this module was similar to the first case study. This particular analytic module used the `course_structure.json` from the edX database model. moocRP can handle JSON files in an even more efficient manner, reading and passing the contents of the file as a JSON object to the HTML page instead of raw string contents. For this module then, we took out the original `d3.json(“filename”, function)` and simply used

```
var data =
         dataFiles[“course_structure.json”]
```

to access the data, without any parsing necessary.

FUNCTIONALITY OF MOOCRP

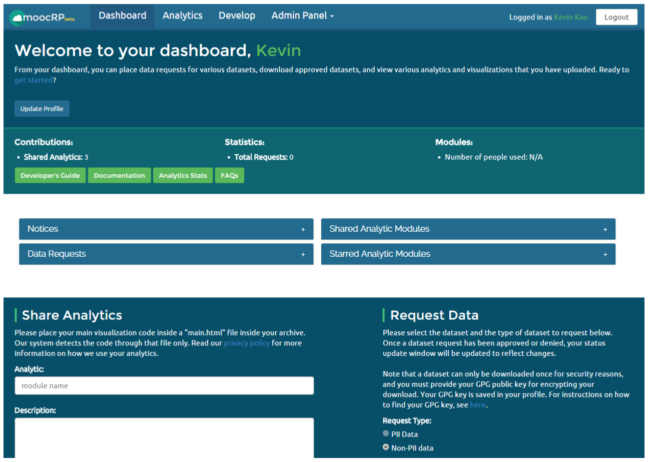
In this section we discuss the various practical functionality of the tool, such as the data request/authorization workflow, security details, technical implementation details, and multiple data model compatibility.

Instructor-Oriented Interface

A prominent design principle kept in mind during the design of this application is the creation of an instructor-oriented dashboard.

Instructors have much to gain from inspecting various aspects of their courses through the analytics modules created by researchers and other instructors (and their GSIs), so the primary use case was that of instructors running MOOCs extracting actionable analytics.

To achieve this, we concentrated on three major areas: ease of access through an aggregated and simplistic interface, facilitation of data distribution, and security.



Welcome to your dashboard, Kevin

From your dashboard, you can place data requests for various datasets, download approved datasets, and view various analytics and visualizations that you have uploaded. Ready to get started?

Contributions:

- **Shared Analytics:** 3

Statistics:

- **Total Requests:** 0

Modules:

- Number of people used: N/A

Share Analytics

Please place your main visualization code inside a "main.html" file inside your archive. Our system detects the code through that file only. Read our privacy policy for more information on how we use your analytics.

Analytic:

Description:

Request Data

Please select the dataset and the type of dataset to request below. Once a dataset request has been approved or denied, your status update window will be updated to reflect changes.

Note that a dataset can only be downloaded once for security reasons, and you must provide your GPG public key for encrypting your download. Your GPG key is saved in your profile. For instructions on how to find your GPG key, see here.

Request Type:

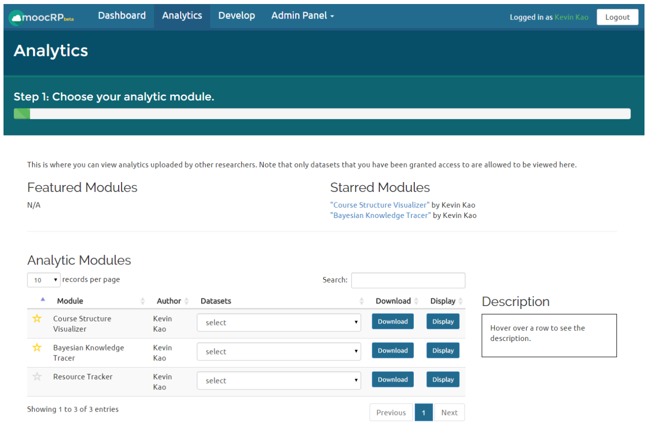
● PII Data
○ Non-PII data

Figure 5. The main interface for moocRP, providing easy access to data requests, analytic modules, and basic stats.

Aggregated interface

moocRP provides a simple, aggregated interface to all of its basic functions. On its main page, moocRP has a series of four closed but expandable panels. The four panels provide easy access to 1) data requests, 2) data downloads, 3) shared analytics modules, and 4) starred analytics modules (see Figure 5). Below, there are two forms for requesting data and uploading analytic modules. The system for data requests and downloads are discussed further in the next section, so here, we focus on the interface for analytics. To facilitate an instructor's access to interesting modules, moocRP allows users to "star" analytic modules and visualizations, essentially creating a bookmark for quick access to the module.

The analytics page provides a paginated list of approved visualizations and analytic modules uploaded by various users of the system (see Figure 6). Each row representing a module also has a dropdown list of datasets for the module to be applied on. The datasets listed are 1) restricted to the datasets that have already been granted to the browsing user, and 2) limited to the compatible data models as specified by the uploading analytic author. For example, modules with a particular Coursera data model most likely is incompatible with an analytic module developed for a particular edX data model. A search feature is implemented in to allow instructors to quickly scout out relevant modules to their problems.



Analytics

Step 1: Choose your analytic module.

This is where you can view analytics uploaded by other researchers. Note that only datasets that you have been granted access to are allowed to be viewed here.

Featured Modules

N/A

Starred Modules

"Course Structure Visualizer" by Kevin Kao
"Bayesian Knowledge Tracer" by Kevin Kao

Analytic Modules

10 records per page

Search:

	Module	Author	Datasets	Download	Display
☆	Course Structure Visualizer	Kevin Kao	select	Download	Display
☆	Bayesian Knowledge Tracer	Kevin Kao	select	Download	Display
☆	Resource Tracker	Kevin Kao	select	Download	Display

Showing 1 to 3 of 3 entries

Description

Hover over a row to see the description.

Figure 6. A two column header provides access to popular analytics modules as well as the visualizations that the user has starred. Below it is the list of available analytics modules, each with drop-down lists of compatible datasets.

A set of links is also provided for quick access to various significant moocRP developer guides, such as the documentation and data model guides for developing analytic modules.

Data distribution

A basic but essential feature of moocRP is its facilitation of data distribution. The processes to request and receive data at numerous institutions often involve a large number of email handoffs culminating in an exchange of a thumb drive with encrypted data in tow. This is just one example of a significantly inefficient process that could become more streamlined and secure. To solve this problem, moocRP provides a shorter data distribution pipeline. Upon registration, users provide a PGP public keys (step-by-step instructions on how to do so are provided on the same page). The user can then login and view the "Request Data" form, where they simply select their choice of dataset and submit the form to initiate the data request. The data scrubbing scripts packaged with moocRP can produce PII (personally identifiable information) and non-PII versions of a dataset with separate authorization for each.

After a data request is sent, any moocRP administrator can go to the administration panel and see a searchable table of data requests, which displays information about each data request (i.e. user's full name, type of dataset, dataset name, date requested). The administrator can either grant or deny the request, which would then reflect on the researcher's dashboard. The users can then easily download their approved datasets with a one-time download provided by the server.

Security

With a web application that works with sensitive datasets containing personal information, security is a major concern. This concern is reflected in the development of moocRP where security comes first at every level.

The first set of security measures comes directly from the user authentication system built into moocRP. The current release of moocRP uses the CAS protocol, otherwise known as the Central Authentication Service. Using CAS has two benefits in one: a secure way to manage users using an institution's existing CAS setup and the ability to use one set of credentials to login to multiple services. Focusing on the security aspect of CAS, we can inspect the CAS authentication process: a user logs in to moocRP, which redirects to the CAS portal of the institution, where the user enters login credentials. The moocRP back-end server waits for a response from the CAS portal that has a "ticket" – moocRP then sends this ticket back to the CAS portal for validation (to prevent access through an expired session or fabricated tickets). If validation is successful, then the user is allowed access to moocRP. CAS provides us *authentication* of the user as well as the user's identity. Future modularization of the authentication module will allow for substitutions of alternative authentication protocols into moocRP, based on each institution's needs.

The next level of security arrives directly when dealing with the datasets moocRP distributes. Upon registration, the user inputs his/her PGP public key, allowing moocRP to utilize the power of encryption to secure datasets. moocRP automates dataset encryption when a dataset is approved for download for a particular user, encrypting the dataset with 2048-bit PGP encryption using the user's public key, allowing only the original owner of the private key to decrypt the dataset.

This leads to another security measure: the entire moocRP application runs over the HTTPS protocol, with SSL built in. This secures the connection between the client and the moocRP server, preventing spoofers and eavesdroppers from gaining unauthorized access to communications. The data distribution is finally secured with CSRF protection and a one-time download link, preventing a malicious attacker from attempting to resubmit a data request form or re-download an approved dataset.

Some additional minor security features are found in the administration panel and logging. The administration panel provides features to manage users, including removal and editing of users, if needed. Uploaded analytics modules can also be removed if found to include malicious code. Logging keeps track of user actions within moocRP, which can be parsed to check for unusual activities.

The last major security measure is inherent to moocRP itself: moocRP in its present state is a relatively closed system, meaning that each moocRP instance will be used and managed by individual institutions, so only members or affiliates of the institution will be accessing moocRP. A *required level of trust* is necessary to re-inforce all of the security precautions that moocRP takes: users are only granted access to the system by each institution's administrators. Likewise, data requests and analytic modules must all go through an approval process by the administrators. PII data will not be viewable via analytics; however, many institutions have security policies around non-PII data. We believe that encryption of the data on the server and visualization of the data over SSL is a high standard of security. Data passed to analytic modules in the browser are not stored on disk and the level of risk associated with these data is not commensurate with any concern over information being resident in browser memory. While it is true that a capable user could intercept and store data passed to the analytic module, it is also the case that such an individual would also only be able to view analytics if already granted full privileges to download the data. At some point the trust bestowed to the individual must be relied upon to cover the balance of scenarios.

Technical Design

Technology stack

The choice of technologies was selected with care. During the planning phase of moocRP development, there were two obvious choices of technologies: Ruby on Rails or Node.js with a backend MVC framework. Note that Node.js should not be directly compared to Rails, since Node.js is a web server while Rails is a framework.

Rails has been on the rise, since before 2010. Node.js, on the other hand, has quickly been gaining popularity in the last couple of years. A few notable differences and considerations when using Node.js with an MVC framework versus using Rails:

- Rails is built on Ruby. Node.js is built on JavaScript, which (often times) has comparable or faster benchmark times than Ruby due to the V8 Engine[3].
- Rails is an opinionated framework, forcing a developer to adhere to its culture and stigmas. Node.js is the opposite, more allowing for a step-by-step build of various components.
- Rails has a steep learning curve with numerous Ruby intricacies to consider. Learning JavaScript is rather quick with experience in any major language like Java or C.

With these observations in mind, we can weigh some advantages and disadvantages of each. For the purposes of moocRP, we heavily favored speed and customization, so we chose Node.js. To make up for what Rails provides out of the box, we used a popular Node.js MVC framework: Sails.js. This allowed us to quickly develop a web application with basic functionality in a Rails manner while still maintaining the ability to customize the application as we desired. Since our work began, moocRP has developed greater support for web sockets, better support of databases, and will soon extend integration of AngularJS, which will allow for a smoother presentation and transfer of data to the client.

Analytics pipeline

The analytics pipeline for sharing and visualizing modules created by users is centric to moocRP. There are a few notable components of the analytics pipeline to discuss: the

[3] http://benchmarksgame.alioth.debian.org/u32/compare.php ?lang=v8&lang2=yarv

file watch, the syncing of datasets to clients, and the scaffolding of uploaded analytic modules.

Essentially, moocRP simply keeps a file watch on specific folders in the base directory of moocRP for datasets. These folders are associated with the datasets of each data model in moocRP. The folders are created when an administrator creates a new data model in the moocRP admin panel.For example, a sample structure of the data directories used for analytics modules. When a new folder is detected inside these directories, moocRP will add new data models and datasets to be available for requesting by users. When a user's request is granted, the user will also be allowed to select the dataset for relevant analytic modules.

The other component of the analytics pipeline involves the uploaded modules. As mentioned in the case studies, the analytic modules is a package consisting of one HTML file, CSS files, and JS files. moocRP extracts these files into a temp folder and moves them into the correct directories to be served. The CSS and JS files are moved into the assets folder, while the HTML file is moved to a view folder used by the framework Sails.js to serve pages with additional features, e.g. template variables between the client and server. Once placed in the correct folders, moocRP will display them on the analytics page.

Support for Multiple Data Models

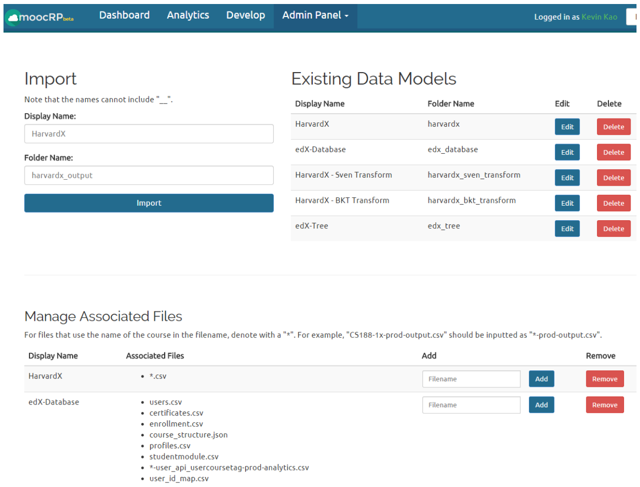
The moocRP tool allows an administrator to add a new data model easily. The process of adding a new data model consists of only a few steps. The administrator can navigate to the "Data Models" tab in the administration panel. The administrator then fills out a form that asks for only the user-friendly name of the new data model and a system-friendly name for the data model (see Figure 7). The system-friendly name is a name that the moocRP can use to automatically create the appropriate directories on the server. Then, the administrator only needs to configure the tools necessary for transforming the data into the new data model formats in a way such that the tools output the datasets in the directories that moocRP created. Therefore, once a new data model is added in the administration panel, moocRP will be able to recognize the new data model and be able to provide support for it for analytics modules and for data distribution.

edX data scrubbing

On the administrator's side, moocRP includes an edX data scrubbing tool out of the box that helps the administrator

- download edX data from the Amazon servers
- decrypt the downloaded data
- organize event data logs across multiple servers into one event data log file for each course, and
- transform the aggregated event data log files into *xAPI event data* format.

Our edX data scrubbing tool is built on top of the HarvardX Tools. We modified several parts of the original scripts to make the tool compatible with the current format of raw edX data, and to speed up the data organizing and transformation time.



Import

Note that the names cannot include "__".

Display Name: HarvardX

Folder Name: harvardx_output

Existing Data Models

Display Name	Folder Name	Edit	Delete
HarvardX	harvardx	Edit	Delete
edX-Database	edx_database	Edit	Delete
HarvardX - Sven Transform	harvardx_sven_transform	Edit	Delete
HarvardX - BKT Transform	harvardx_bkt_transform	Edit	Delete
edX-Tree	edx_tree	Edit	Delete

Manage Associated Files

For files that use the name of the course in the filename, denote with a "*". For example, "CS188-1x-prod-output.csv" should be inputted as "*-prod-output.csv".

Display Name	Associated Files	Add	Remove
HarvardX	*.csv	Filename Add	Remove
edX-Database	users.csv certificates.csv enrollment.csv course_structure.json profiles.csv studentmodule.csv *-user_api_usercoursetag-prod-analytics.csv user_id_map.csv	Filename Add	Remove

Figure 7. The data models management page, where data models can be added or deleted in a few clicks. Associated files of each data model is also managed here.

TOOL SUSTAINABILITY AND FUTURE WORK

Centralized repositories, such as the LearnLab's DataShop, are not fully supported by an institution and rely on government funding to maintain. The moocRP model instead leverages institutional self-interest to maintain its services. Institutions have an interest in facilitating research from its faculty and staff as well as leveraging value from its data in order to provide a more effective and efficient educational experience to its students. No particular platform buy-in or inter-institutional network is required; rather each institution may utilize the tool and benefit from the analytics modules available for it.

Much work remains to be done to improve moocRP. Some important features to be implemented with community support and continued in-house engineering include:

- Automated analytic module security screening
- Alternative authentication protocols and modularization
- Integration of data pre-processing scripts
- Support for scripts of alternative languages, e.g. Matlab, Perl, Python, that could be used for machine learning analytics
- Additional statistics on data model and analytic usage information
- Various server and implementation optimizations

We will also continue collaborations with interested universities and learning systems through webinars and conferences to gain feedback on the tool.

CONCLUSIONS

In this paper we introduced an open-source analytics platform that adopts the three tenants of open learning analytics; transparency, modularity, and privacy. The platform provides value to; MOOC researchers by allowing for replication of results by providing a simple framework to share and develop analytic modules, to instructors by providing an easy to use interface for applying community generated analytics to their course, and to administrators by providing a workflow for managing secured data repository access controls. The moocRP platform is not a panacea to the problem of open learning analytics or reproducible research, but it is a earnest step towards bringing the power of learner data to bear on both the educational and research objectives of an organization.

ACKNOWLEDGEMENTS

The second author was partially supported by a Google Faculty Research Award (#2014_R1_446).

REFERENCES

[1] Siemens, G., Gašević, D., Haythornthwaite, C., Dawson, S., Buckingham Shum, S., Ferguson,R., et al. (2011) Open Learning Analytics: An Integrated and Modularized Platform (Concept Paper): SOLAR.

[2] Koedinger, K. R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010) A data repository for the EDM community: The PSLC DataShop. *Handbook of educational data mining*, *43*.

[3] Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z.A. & O'Reilly, U.M. (20213) MOOCdb: Developing data standards for MOOC data science. In *AIED 2013 Workshops Proceedings* (pp. 17-24)

[4] Pardos, Z. A., Heffernan, N. T. (2010) Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In *Proceedings of the 18th International Conference on User Modeling, Adaptation and Personalization (UMAP)*. Springer. (pp. 255-266)

[5] Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., ... & Chuang, I. (2014) Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, *57*(9), 56-63.

[6] Martin, B., Mitrovic, T., Mathan, S., & Koedinger, K.R. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Model User-Adap Inter* (2011) 21:249–283.

[7] Feng, M., Heffernan, N.T. (2007) Towards Live Informing and Automatic Analyzing of Student Learning: Reporting in ASSISTment System. *Journal of Interactive Learning Research* 18 (2): 207-230.

[8] Gertner, A.S., and VanLehn, K. (2000) Andes: A Coached Problem-Solving Environment for Physics. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 133-142)

[9] Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995) Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences* 4 (2): 167-207.

[10] Lovett, M., Meyer, O., & Thille, C. (2008) The Open Learning Initiative: Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education*, no. 14. Special Issue http://jime.open.ac.uk/2008/14

[11] Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of SIGCHI 2011* (pp. 245-252). ACM.

[12] Bloom, B. S. (1968). Learning for mastery.

[13] Ferguson, R., & Shum, S. B. (2012). Social learning analytics: five approaches. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 23-33). ACM.

[14] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013) Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, *8*, 13-25.

[15] Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., ... & Chuang, I. (2014) Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9), 56-63.

[16] Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2013) Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, 1-16.

[17] The Asilomar Convention for Learning Research in Higher Education (http://asilomar-highered.info/asilomar-convention-20140612.pdf)

[18] edX Research Guide. Data Delivered in Data Packages. http://edx.readthedocs.org/projects/devdata/en/latest/internal_data_formats/package.html, 2014.

[19] Jim Waldo. HarvardX Tools. https://github.com/jimwaldo/HarvardX-Tools

[20] ADL Initiative, Experience API, 2013, https://github.com/adlnet/- xAPI-Spec/blob/master/xAPI.md

[21] Stanford Vice Provost Office for Online Learning. How to Access the VPOL Online Learning Data. http://datastage.stanford.edu

[22] Andreas Paepcke. json_to_relation. https://github.com/paepcke/json_to_relation

[23] MOOCdb. http://moocdb.csail.mit.edu/wiki/index.php?title=MOOCdb

[24] Stephens-Martinez, K., Hearst, M. A., & Fox, A. (2014) Monitoring MOOCs: which information sources do instructors value? In *Proceedings of the first ACM conference on Learning@ Scale* (pp. 79-88)

[25] Xu, Z., Goldwasser, D., Bederson, B. B., & Lin, J. (2014) Visual analytics of MOOCs at Maryland. In *Proceedings of the first ACM conference on Learning@ Scale* (pp. 195-196)