PETRA: A Crowdsourcing-Based Platform for Rocks Data Collection and Characterization

Raúl A. Mira[1], Patricia L. Suarez[1], Rafael E. Rivadeneira[1] and Angel D. Sappa[1,2]
[1]ESPOL Polytechnic University, Escuela Superior Politćnica del Litoral, ESPOL, CIDIS-FIEC
Campus Gustavo Galindo Km. 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
[2]Computer Vision Center, Edifici O, Campus UAB,
08193 Bellaterra, Barcelona, Spain
{raualmir, plsuarez, rrivaden, asappa}@espol.edu.ec

***Abstract*—This paper presents details of a distributed platform intended for data acquisition, evaluation, storage and visualization, which is fully implemented under the crowdsourcing paradigm. The proposed platform is the result from collaboration between computer science and petrology researchers and it is intended for academic purposes. The platform is designed within a MTV (Model, Template and View) architecture and also designed for a collaborative data store and managing of rocks from multiple readers and writers, taking advantage of ubiquity of web applications, and neutrality of researchers from different communities to validate the data. The platform is being used and validated by students and academics from our university; in the near future it will be open to other users interested on this topic.**

***Index Terms*—crowdsourcing; collaborative framework; website; characterization in petrology.**

I. Introduction

Certainly, one of the areas involved in the development of the economy of a country is geology, more specifically, mineralogy. Its understanding includes describing and studying the physical and chemical properties of minerals found in a certain region. The classification of rocks is an essential part of modern geology [1], its role in mining is fundamental for the discovery of mineral deposits or oil wells, through the macroscopic and microscopic analysis of rocks [2], which is known as *petrography*.

Most of the classical rock classification approaches are based on manual identification carried out by experts in mineralogy; techniques such as polarized light microscopy, X-Ray diffraction (XRD), Atomic Absorption Spectroscopy (AAS), among others [3], are widely used in the rock classification process. The latest advances in this area show new intelligent techniques carried out by a computer under the supervision of a human expert. Solutions including computer vision have been developed, such as the identification of thin-section images in RGB and HSI color spaces, achieving the identification of 10 different minerals using artificial neural networks (ANN), with an accuracy of 93.53% [4]. More recent studies outperform previous approaches by tackling the classification problem based on the granularity of rocks in images of thin sections; these approaches achieve 98.5% of accuracy using deep Convolutional Neural Networks (CNN) [5]. However, whether to employ manual techniques or supervised techniques using computer vision approaches, a massive and reliable collection of data is required. Hence, one of the problems presented in this field, is the limited and centralized access to this data. In academic institutes often this data is subject to a small number of validations and unfortunately, in most of the cases, this amount of data is not yet digitized. Furthermore, sometimes data is not correctly annotated.

In the current work a crowdsourcing-based platform, named *Petra*, is proposed to collect a large scale of labelled rock images and to generate a digital library in order to tackle the aforementioned problems. The proposed platform is designed under the crowdsourcing philosophy, which allows different users to upload data that is later evaluated by experts before including them in the digital library. This platform was developed and tested in ESPOL in collaboration with the School of Earth Science Engineering. The manuscript is organized as follows. Section II presents works related with the topics tackled in the current work. The proposed solution is detailed in section III. Implementation results are provided in section IV. Finally, conclusions are given in section V.

II. Related works

Generating a massive set of labeled images, with the corresponding meta-data, becomes a time consuming and expensive task. Traditionally, datasets are constructed by a single research group and are intended to solve a specific problem [6]. In this section state of the art strategies on data acquisition, management and visualization are summarized.

A. Crowdsourcing

Crowdsourcing is a business practice that means literally to outsource an activity to the crowd [7]. In general, crowdsourcing is a model for distributed problem-solving that utilizes a group of individuals (crowd), to provide solutions to problems. This approach started to get popular by Amazon connecting internet users by a platform called Mechanical Turk (a.k.a. MTurk), where researchers pay workers to complete surveys, participate in experiments, and conduct content analysis. Today, most of the results from MTurk have been used in published studies in the social science disciplines, marketing,

978-1-7281-3764-3/19/$31.00

psychology and political science [8]. Later, some other platforms that accomplish tasks with the same philosophy were developed, such as Samasource, Upwork, Knowxel just to mention a few (e.g., [9] [10], [11]). Similar to these tools, which address some specifically tasks, the current work aims to address the acquisition, processing and management of rock data, hence some platforms related with these tasks are reviewed in the next sub-section.

B. Data acquisition, labelling and management systems

As mentioned above, there are several crowdsourcing-based platforms for data acquisition, labelling and management, in this section just some of them are reviewed. Regarding data acquisition, in [10] a general-purpose crowdsourcing platform for mobile devices is proposed. Since the platform is intended to be used with mobile devices, all the embedded sensors could be used for the crowd for data acquisition (e.g., images, geo-localized images, sound, user trajectory, etc.). The users, distributed all around the world, were asked to do a task that was collected by a server.

Once data is acquired, the crowdsourcing paradigm can be used for data processing. For instance, Russell et al. [6] proposed LabelMe, a web based tool that allows easy image annotations and instant sharing of such annotations. It allows to generate object categories from a given scene. Basically, object class recognition where the labeling is semi-automatic but mostly achieved by users feeding the database, the most representative functionality is to draw in the scenes the borders of any object and giving them the respectively categorization. This solution was developed in order to provide large datasets for computer vision scientists, similar to CALTECH-101 [12]. Interactive Vascular Modeling Environment (IVME) is another imaging characterization platform, this is a solution for medical education, research and clinical purposes. It affords editing, manipulation, quantification and labeling of vascular-brain models and angiography data in 2D and 3D [13], including direct manipulation of the 3D vascular models and examination of 2D-3D correspondence.

Structural Health Monitoring (SHM) platform [14] is a crowdsourcing implementation for data managing; it integrates mobile sensing (including an iOS app) and web based computing affording the citizens to measure structural vibrations, such data is stored as modal properties of structures (e.g., bridges) after being automatically processed. The field test results showed that the vibration data acquired by citizens without expertise are useful for identifying structural modal properties with high accuracy and concluding it may be convenient for long-term monitoring of structural integrity of spatially distributed urban infrastructure.

Moreover, the applications mentioned above it has been demonstrated that crowdsourcing methods brings several benefits in the performance of building and gathering data for digital libraries by completing microtasks (simple tasks such as images tagging), and macrotasks (special skilled tasks) together. Those benefits include achieving goals of library faster, building new virtual space between libraries and communities, by seizing the knowledge expertise among people and making data widely discoverable. [15]

Under the crowdsourcing framework, in the current work the Petra platform is proposed. In a collaborative approach, Petra takes all the concepts reviewed above to generate massive reliable data collection of annotated rocks. It uses a group of individuals (including experts, researchers and academics of petrology and mineralogy) that feed the digital library of rocks, validating the data with a cross-evaluation system between the pools, therefore enhancing the confidence in the quality of the data, as showed in Fig. 1. The proposed architecture, including the managing system and evaluation methodology, as well as all the implications, are detailed in next section.

III. Proposed approach

This section presents details of the implemented system for the proposed digital rock library generation. First, details on the software architecture are provided; then, the front-end design is summarized showing the different user profiles and options; finally, the crowdsourcing methodology implemented to validate the uploaded data is presented.

A. Software architecture

The system architecture of Petra, showed in Fig. 2, is implemented on the Django framework based on Python, spanning a single three layers structure: Model, Template, View (MTV); The ”Model” is the layer for accessing the database, contains all the information about the rocks and the relationships between the data. The ”Template” layer contains the display logic and all the decisions related with the presentation, such as web pages or other resources. And the ”View” layer containing the business layer and all the logic for accessing the models and connect them with the templates. Python has been chosen because is a very suitable language for developing web applications and has a very large community of developers giving support, without mentioning that it is free and open source [16]. On the other hand, Django offers more portability, maintainability, efficiency, usability, reliability and functionality than other python-based frameworks of its type [17].

Regarding the data management and storage, MongoDB has been used. It is a distributed NoSQL database system that takes advantage of its dynamic scheme. This is due to the scalability of Petra system to allow variable characteristics and attributes according to the type of rocks that are upload (macroscopic or microscopic attributes) and possibly future variations. Finally, MongoEngine is used as a Document-Object Mapper for work with MongoDB, which uses a simple declarative API, very similar to the Django ORM.

B. Front-end design

Petra has been designed with three main modules: acquisition, visualization and evaluation. As a crowdsourcing-based platform, there are groups of users (pools) that work in a collaboration-based system. Therefore, each group of collaborators has to specifically access only the properly

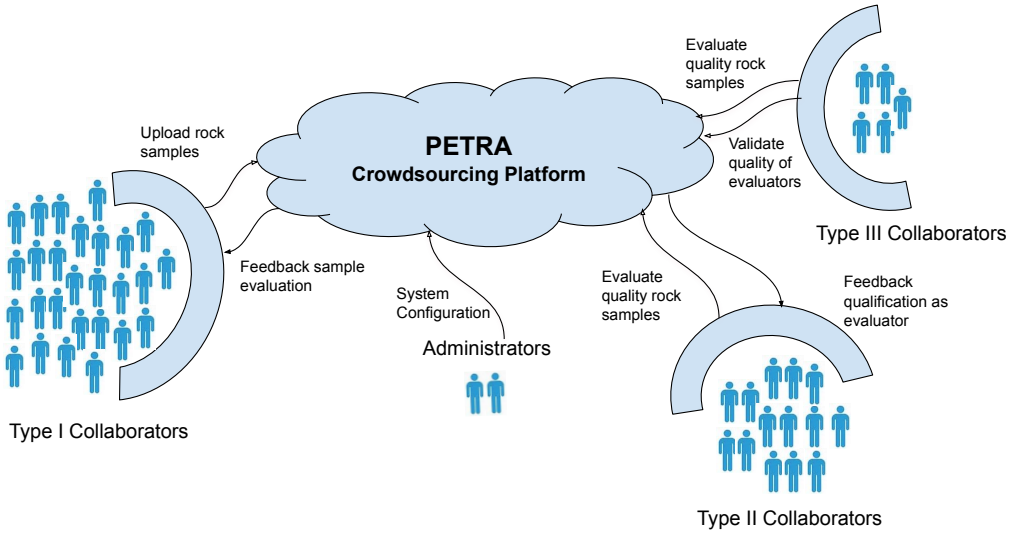

Fig. 1: Petra crowdsourcing platform structure: *Type I* collaborator, whose function is limited to the entry of new samples, generally an undergraduate student. *Type II* collaborator, performs the validations of the entered samples, typically advanced users such as researchers in the area. *Type III* collaborator, such as seniors researchers, have the possibility to decide on a sample didn't reach an agreement in the type II collaborators decision, also can decide the final status of a sample in case of a wrong evaluation in the previous phase.

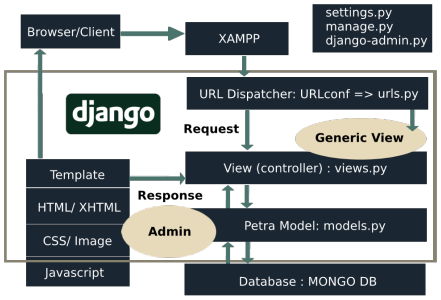

Fig. 2: Architecture of Petra.

authorized modules, for that reason, an authentication system that performs the corresponding access validation has been incorporated, management by an administrator profile. It has been defined three different accessibility levels which are: *Type I* collaborator, who fulfills the role of feeder for dataset of rocks, whose function is limited to the entry of new samples, generally an undergraduate student. *Type II* collaborator, who performs the validations of the entered samples, typically advanced users such as researchers in the area (the data validation strategy is explained below in sub-section C). And finally the *Type III* collaborator, who is usually an expert in the area, such as a senior researcher or a full professor with a large experience on this topic. Type III collaborators have the possibility to approve any sample entered even without the validation of a Type II collaborator, furthermore, in case a sample contains a few wrong annotations, Type III collaborator can correct these annotations and approve the sample.

Acquisition Module: This module consists of an interface that allows to upload information of different samples, this is done by Type I collaborator. In this module the attributes and characteristics of the rock are defined and divided into groups: identification, macroscopic and microscopic attributes. The distribution of the attributes was made based on standardized templates for rock characterization, used in ESPOL.

This module, as shown in Fig. 3, also allows the acquisition of images by configuring and selecting an image capture device, such as a camera, or loading images from a local file. The parameters of the rock are mostly standardized, thus avoiding duplicity in the categories, some of these parameters are: name of rock (basalt, granite, quartzite, etc.), the type, which defines whether they are igneous, metamorphic or sedimentary, and other parameters such as age, fracturing, crystallinity, grain size, among others.

C. Quality validation

Method	n	%
Expert review	46	77%
Photo submissions	24	40%
Paper data sheets submitted along with online entry	20	33%
Replication or rating by multiple participants	14	23%
QA/QC training program	13	22%
Automatic filtering of unusual reports	11	18%

TABLE I: Most common validation mechanisms from survey [18] where n is the number of projects who handle it and % is the relative percentage, notice that one single project may be handling more than one mechanism

Crowdsourcing-based approaches employ multiple strategies to ensure data quality and validation, involving not only fitness for the data, but also completeness, validity, consistency, precision, and accuracy. Petra, like most citizen science platforms (crowdsourcing type), relies on adequately large

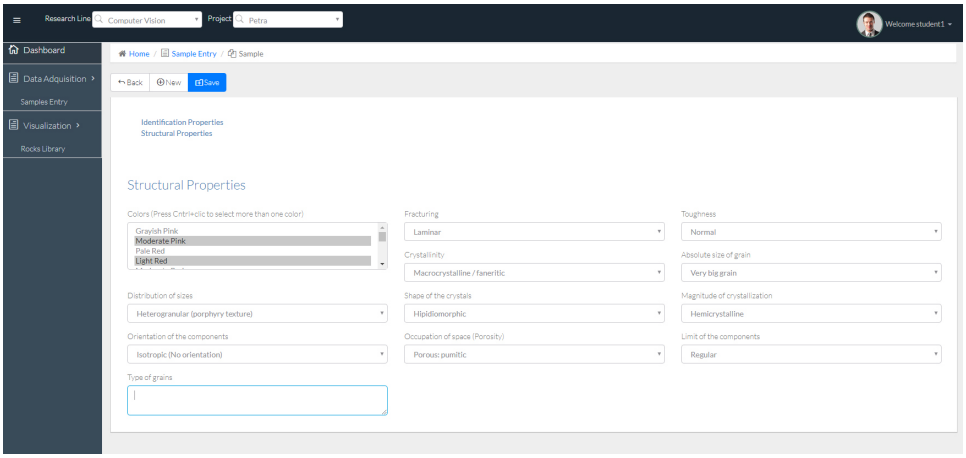

Fig. 3: View of the Petra's acquisition module—structural attributes annotations.

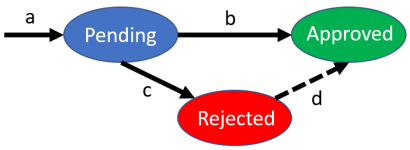

Fig. 4: Status evaluation flow: a) sample data uploaded, it is pending; b) if 2/3 evaluators approve all attributes, it changes to *Approved*, c) else if 2/3 evaluators reject at least one attribute, it changes to *Rejected*; d) if more than 50% of attributes are correctly annotated it could change to *Approved* by an expert reviewer.

numbers of collaborators, including a wide variability in skills and expertise between collaborators. From a survey including approximately 280 projects and 560 individuals connected to additional crowdsourcing platforms, most common mechanisms are showed in Table I, where strategies based on expert review have given good results and have been widely used in the literature as a validation mechanism. Most of projects also choose not only to have one validation mechanism but a combination of them, for example, expert review plus automatic filtering, expert review plus photos, expert reviews plus paper data sheets, among others combinations. Following the framework suggested in [18]. *Participant training*: it is defined as a "before, during process", which is considered as a preventive mechanism. On the other hand, *Expert Review*: it is defined as an "after process" for validation mechanism. Expert Review can be subdivided in three categories: professionals, experienced collaborators and multiple parties experts.

Evaluation Module: This module consists of a combination of all mechanism previously detailed for data and performance validation which was implemented in Petra platform. An interface has been developed for sample evaluation process.

The entered samples are assigned with one of the following status: **pending** (the sample has not yet been completely evaluated by a pool of three *Type II* collaborators), **approved** (the sample was completely evaluated and approved by the pool of three *Type II* collaborators), and **rejected** (the sample was completely evaluated and rejected by the pool of three *Type II* collaborators), as shown in Fig. 4.

Once a sample is entered into the system, it is assigned to a pool of three *Type II* collaborators, the assignment is randomly performed and weighted depending on the workload for each collaborator, the criteria with which the system updates the status of a sample entry is the following: A sample is defined with status **approved** only if the evaluation of the sample has an agreement of the pool for all the attributes entered for that sample. For example, if the sample was entered with a total of 20 attributes, it will be approved by the system if every attribute gets at least two out of three approvals. In the counterpart, if the condition mentioned above is not reached, but the pool still reach an agreement (following the example), if at least one of the attributes has an agreement of the pool as rejected, then the sample get the status as **rejected**.

The *Type III* collaborator performs in the system as a second validator when a sample has not reached the **approved** status but has an agreement upper than 50%, then he/she could decide to review the evaluation results of each member in the pool and approve or definitely reject the sample. Also, as an expert, this collaborator is able to change the values of the attributes of a sample to make it a valid sample. This evaluation system establish a double validation for *expert review* mechanism to ensure the quality of data.

The *Rating Participant Performance* is the second mechanism used for quality validation. For this mechanism, the rating performance for each collaborator is evaluated based on the other collaborator evaluations. For a *Type I* collaborator the evaluation is done once each sample entered by that collaborator change its state from **pending** to **approved** or **rejected**. If the collaborator completes three **rejected** samples with an agreement lower than 50%, he/she will be automatically banned from uploading more samples; for each sample, once is completely evaluated, he/she will have the option to obtain feedback from the evaluators, this way, the system will prevent the library to be feed with junk data and malfeasance.

For a *Type II* collaborator, the evaluation takes place also once each sample entered by that collaborator change its state.

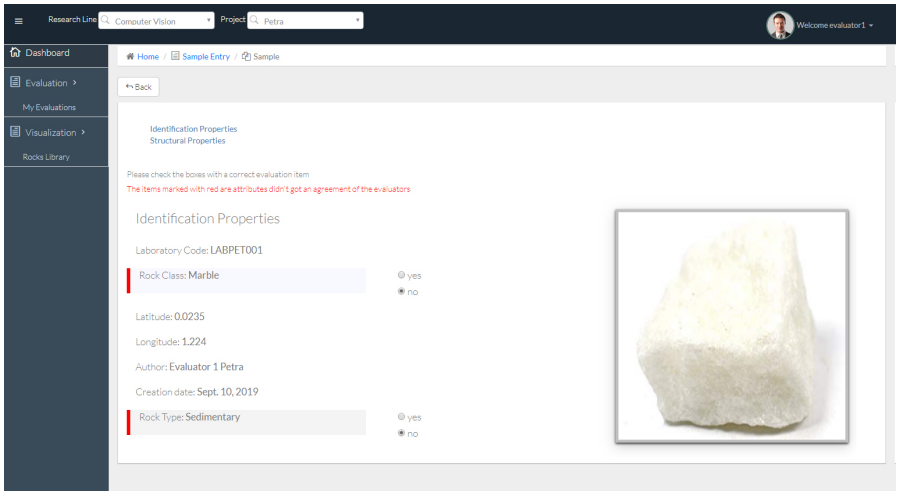

Fig. 5: View of the Petra's evaluation module.

It means, once all the members from the pool of collaborators have already evaluated the sample. If the collaborator completes three evaluation samples with an agreement lower than 50% (agreement with other collaborators of the same pool), the collaborator will be automatically banned from evaluating more samples in the system, he/she will lose his condition of expert and will become a *Type I* collaborator, for each sample, once it is completely evaluated, he/she will have the option to check the evaluations from the other members in the pool.

To achieve this validation, Petra provides an interface in evaluation module, as shown in Fig. 5, where each of the attributes for the sample can be validated by a simple checkbox. This interface is used for both *Type II* and *Type III* collaborators, with the only difference that *Type III* collaborators can also change the values of each parameter and save the sample again with an **approved/rejected** status.

IV. Implementation Results

As mentioned above, the main purpose of Petra is to build a digital, valid and high-quality library of rocks for academic purposes; this task is accomplished through the proposed crowdsourcing approach with different type of collaborators, leads to the platform to show the final results of the collaborators efforts. Some snapshots of the interface showing illustrations of the platform are depicted in this section. The platform is not only used to upload samples, but it is also used for the academic community to visualize samples—in the case of undergraduate students to learn about distinctive characteristics of each category.

Visualization Module: This module presents a collection of the samples that have **approved** status, as shown in Fig. 6; it is the main view for a query and cannot be modified through this view. It is just the result for querying the database. This view shows a brief summary of the annotated rocks, the location and collected date, the author that has uploaded the sample and the type of rock. Also, it points to a detailed information of each sample. Same as many other digital dictionaries, the samples can be filtered and sorted to have a better understanding of the data. This module provides the academics with a very important source of information to improve their knowledge by having a ground truth for their researches and projects.

Once the prototype of the application has been launched, it can be evidenced that the use of the application has been constantly growing as can be seen in the Fig. 7, since the amount of rock samples accepted has been increasing, and also maintaining a constant samples waiting to be evaluated and finally it can be seen that the number of rejected samples has been decreasing as users gain more experience in the specialized field of rocks, improving their analyzing ability.

V. Conclusions

This paper presents in detail the implementation of a crowdsourcing-based platform for rock data collection and characterization. Petra architecture is designed to support multiple user profiles operations simultaneously, depending on the functionalities that will be authorized. The data validation strategy, based on the crowdsourcing paradigm, is explained by showing the construction of specialized knowledge that is obtained over time with the help of collaborating experts using the proposed solution. Finally, snapshots of different platform's windows are depicted showing the design and usability of the platform. Currently the platform is available just for students and academics of ESPOL, as a future work it is expected to be open for the whole community trying to increase the amount of data collected and validated under Petra framework.

References

[1] M. Mynarczuk, A. Grszczyk, and B. lipek, "The application of pattern recognition in the automatic classification of microscopic rock images," *Computers & Geosciences*, vol. 60, pp. 126–133, 10 2013.

[2] L. B. Gonalves and F. Leta, "Macroscopic rock texture image classification using a hierarchical neuro-fuzzy