

UNIVERSITATEA POLITEHNICA BUCUREŞTI
FACULTATEA DE AUTOMATICĂ ȘI CALCULATOARE
DEPARTAMENTUL CALCULATOARE



**Interfata UI pentru incarcare de seturi de date
corespunzatoare articolelor si proiectelor stiintifice**

Coordonator științific

Prof. dr. ing. Ciprian Mihai Dobre

Studenti
Sergiu Isopescu
Claudiu Marcel Nedelcu

BUCUREŞTI

2021

CUPRINS

CUPRINS	1
SINOPSIS	2
ABSTRACT	2
1 INTRODUCERE	3
2 SOLUȚIA PROPUȘĂ	3
2.1 Incarcare set de date	3
2.2 Cautare seturi de date	4
2.3 Informatiile unui set de date	6
2.4 Informatiile utilizatorului autentificat	8
2.5 Instanta CKAN	9
2.6 Baza de date	9
2.7 Sistem de raportare și vizualizare a datelor	10
3 IMBUNATATIRI VIITOARE	10
4 CONCLUZII	11

SINOPSIS

Universitatea Politehnica Bucuresti necesita sprijin in administrarea activitatilor de cercetare, urmarind automatizarea procesului de raportare periodica a rezultatelor de catre cercetatori. Astfel, platforma CRESCDI ofera suport pentru incarcarea de articole stiintifice ale studentilor si profesorilor sau extragerea metadatelor de pe Google Scholar si Brainmap pentru articolele deja existente, asociate studentilor sau profesorilor care activeaza in universitatea noastra. Fiecare utilizator al platformei are acces la toate publicatiile sale si la statistici anuale privind proiectele proprii, dar si la legaturile sale sociale afisate pe o hartă a colaboratorilor cu care a interactuat.

Proiectul curent urmareste extinderea platformei prin posibilitatea de a putea fi incarcate si seturile de date utilizate in intocmirea articolelor stiintifice. Acestea sunt importante pentru ca ofera posibilitatea altor cercetatori de a verifica rezultatele curente si de a reutiliza datele in scopul unei posibile continuari a procesului de cercetare.

ABSTRACT

Politehnica University of Bucharest needs support in the administration of the research activities. It aims to automate the process of periodic reporting of researchers results. The CRESCDI platform offers support for uploading scientific articles written by students and teachers or extracting metadata from Google Scholar and Brainmap for the already existing articles, associated to students or teachers that are representing our university.

Each user of the platform is able to access all his publications and annual statistics about his own projects, but also his social connections displayed on a map of the collaborators with whom he interacted. The current project aims to expand the platform through the possibility of uploading the data sets used in the preparation of scientific articles. These are important because they provide the opportunity for other researchers to verify the current results and reuse the data for a possible continuation of the research process.

1 INTRODUCERE

In domeniul cercetarii este importanta partajarea datelor si a informatiilor atat pentru a se face vizibile rezultatele finale, cat si pentru a spori si imbunatatiti procesele ulterioare de cercetare.

Pentru a imbunatatiti platforma curenta, proiectul curent urmareste conceperea unei platforme anexe prin intermediul careia utilizatorul sa poata incarca seturile de date utilizate in cadrul proiectelor proprii.

Implementarea si functionarea sistemului format din platforma CRESCDI si platforma anexa ce se doreste a fi implementata urmareste oferirea posibilitatii cercetatorilor din cadrul Universitatii Politehnica Bucuresti de a-si incarca si descarca publicatiile si seturile de date folosite in cadrul acestora si de a avea acces atat la fisile de autoevaluare a activitatii stiintifice si didactice, cat si la legaturile sale sociale afisate pe o hartă a colaboratorilor cu care a interactionat.

2 SOLUȚIA PROPUȘĂ

Proiectul curent urmareste implementarea unei platforme anexe prin intermediul careia studentii si profesorii cercetatori din universitatea noastra vor putea incarca si gestiona ulterior seturile de date utilizate in cadrul articolelor si proiectelor de cercetare.

2.1 Incarcare set de date

La incarcarea unui set de date, utilizatorul autentificat trebuie sa furnizeze cateva date de incadrare a setului de date, precum: titlu, titlul articolului asociat, anul publicarii setului de date, domeniul de care apartine, tag-uri reprezentative din domeniul selectat, tara (se va realiza o mapare a tarii la coordonatele geografice ale acesteia pentru a putea reprezenta locatiile seturilor de date pe o hartă intuitivă) si o scurta descriere a setului de date. De asemenea, se poate seta confidentialitatea setului de date (public sau privat), un link de GitHub catre codul care utilizeaza setul de date respectiv si un link folosit pentru descarcarea setului de date daca acesta se afla deja pe o alta platforma. Mai mult, utilizatorul trebuie sa completeze campurile asociate intregitati si reutilizarii datelor si accesul continuu la setul de date ce urmeaza sa fie incarcat.

In cazul in care domeniul dorit de utilizator nu se regaseste in lista de domenii existente, acesta are posibilitatea sa il adauge folosind optiunea Other si completand numele acestuia. Acelasi avantaj se regaseste si la nivelul tag-urilor. Utilizatorul poate adauga tag-uri noi, reprezentative setului de date, fie pentru domeniul selectat, fie pentru domeniul nou care urmeaza sa fie creat.

Odata completat acest formular (Figura 1), se selecteaza fisierul ce urmeaza sa fie incarcat, fiind premise doar anumite tipuri de fisiere. Acestea sunt expuse utilizatorului, iar validarea formatului fisierului se efectueaza atat in front end, cat si in back end, informatia fiind extasa din antetul fisierului. Mai departe, toate metadatele sunt colectate in baza de date a platformei, la index-ul *datasets*, apoi se creeaza in instanta Ckan un pachet corespunzator setului de date, iar fisierul este incarcat sub forma de resursa si atribuit pachetului mentionat. Dimensiunea maxima acceptata a unui fisier este configurabila, momentan fiind setata la valoarea de 100MB.

The screenshot shows a web-based dataset upload interface. At the top, there's a header with the University Politehnica of Bucharest logo, the CRESCTI logo, and user information (Claudiu, Logout). On the left, a sidebar menu includes links for About project, CKAN Instance, Dashboard, Privacy Policy, Sponsors, Developers, and Contact us. The main form area has several input fields:

- Private**: A toggle switch.
- Dataset title:** A text input field labeled "Dataset title".
- Authors:** A text input field labeled "Dataset authors".
- Show more info**: A blue button.
- Article title:** A text input field labeled "Article title". A note below says "This field is optional".
- Year:** A text input field labeled "Year of the publication".
- Select a country:** A dropdown menu labeled "Select Country".
- Select a Domain:** A dropdown menu labeled "Select Domain".
- Select tags:** A text input field labeled "Select tags".
- Short Description:** A text input field labeled "Please enter a short description...".
- GitHub link:** A text input field labeled "GitHub link". A note below says "This field is optional".

 At the bottom, there are two buttons: "Download Link..." and "Upload dataset". The "Upload dataset" button has a file input field labeled "Choose File" with the placeholder "No file chosen" and a note "-zip, tar, gz, rar, 7z, json, pdf, jpg, csv".

Figura 1. Incarcare set de date

2.2 Cautare seturi de date

Cautarea unui set de date pe platforma noastra se realizeaza intr-o maniera foarte simpla. Intrucat la incarcarea unui set de date sunt cerute cateva metadate (informatii de incadrare precum domeniu, tag-uri sau titlu), cautarea se realizeaza cu ajutorul unor filtre (Figura 2) prin care utilizatorul poate selecta domeniul, tara asociata articolului, formatul datelor asociate articolului si tipul de descarcare. De asemenea, utilizatorul poate cauta setul de date dupa numele si anul publicarii acestuia, dupa numele autorilor si dupa cateva taguri specific domeniului selectat.

The screenshot shows a search interface with the following elements:

- Top navigation: "All domains", "All countries", "All Data Formats", "All Downloads".
- Search input fields: "All tags", "Author", "Year", "Dataset title".
- Sort options: "Sort By" dropdown, a numeric input field set to "7", and up/down arrow buttons.
- Action buttons: "Search" button with a count of "31" and a "Quick" link.

Figura 2. Filtre de cautare

Optiunile pentru selectarea domeniului, a tag-urilor, a tarii, a formatului datelor si a tipului de descarcare sunt incarcate din baza de date si reprezinta toate variantele existente in momentul respectiv. Pentrufiltrele asociate autorilor, anului publicarii si titlului setului de date, cautarea se bazeaza pe expresii regulate pentru a regasi fragmente din cautarea utilizatorului in metadatele existente. Mai mult, daca utilizatorul aplica o cautare bazata pe mai multe nume de autori, acestea trebuie separate prin caracterul „,”, urmand a fi despartite in back end si a fi cautate toate fragmentele in liste de autori existente.

Odata cu modificarea oricarui filtru, se executa din nou cautarea si se actualizeaza numarul de rezultate din badge-ul butonului de cautare. De asemenea, in cazul returnarii mai multor rezultate in urma aplicarii filtrelor, utilizatorul va putea pagina raspunsul prin selectarea numarului de rezultate afisate pe pagina si va putea sorta rezultatele.

In urma selectarii filtrelor dorite, datele se trimit catre server, unde sunt validate, iar apoi se construieste o expresie de match a seturilor de date pe baza filtrelor setate. Rezultatul cautarii este apoi usor prelucrat pentru a fi trimis in front end si afisat utilizatorului. Fiecare obiect din rezultat contine datele specifice setului de date asociat, conform Figurii 3. Odata accesat titlul setului de date, utilizatorul este redirectionat catre pagina acestuia.

Title:	Paths of Glory Dataset		
	2000		Putin
Title:	The Hunted Dataset		
	2014		admin
Title:	Man of Extremes Dataset		
	2008		admin
Title:	A Style-Based Generator Architecture for Generative Adversarial Networks Dataset		
	2014		admin

Figura 3. Rezultatul cautarii

2.3 Informatiile unui set de date

Odata cu accesarea titlului unui set de date, utilizatorul este redirectionat catre o pagina care cuprinde toate informatiile corespunzatoare setului de date selectat (Figura 4). Printre acestea se numara toate metadatele asociate, momentul de timp al ultimei actualizari si informatii despre rating si despre detinatorul setului de date. De asemenea, daca setul de date are o resursa interna atasata, se ofera un fingerprint al fisierului (md5 checksum) pentru asigurarea integritatii acestuia.

The screenshot shows a web interface for the RESCDI platform. At the top, there's a navigation bar with the University Politehnica of Bucharest logo, a search bar, and user account information (Claudiu, Logout). The main content area features a large banner for a dataset titled "California Traffic Collision Data from SWITRS Data", which is marked as "Public" and was last updated 2 minutes ago. Below the banner, there's a sidebar with links like "Search datasets", "About project", "CKAN Instance", "Dashboard", "Privacy Policy", "Sponsors", "Developers", and "Contact us". The main content area has tabs for "About" and "Download". The "About" tab displays detailed metadata: Domain (IT), Rating (0), Data format (zip); Authors (Shivam Bansal, Programming Technology Automation); Article title (California Traffic Collision Data from SWITRS); Resource Type (Resource checksum: 1ae47ebc147d3c463af339c29e9f32de, Gitlink: www.github.com); Country (Romania); Created (2020); and a short description: "Short description for California Traffic Collision Data from SWITRS should be here". It also includes sections for Data integrity and authenticity (Is true, Continuity of access, Of course, Data reuse, Always), and a "Download" button. A pencil icon is visible in the top right corner of the main content area.

Figura 4. Pagina setului de date

O alta functionalitate oferita de aceasta pagina este actualizarea setului de date. Acest proces se efectueaza doar de catre detinator si permite atat actualizarea metadatelor, cat si a resursei. In Figura 5 se observa posibilitatea de conversie a resursei la oricare din cele 3 tipuri suportate: None, Internal sau External.

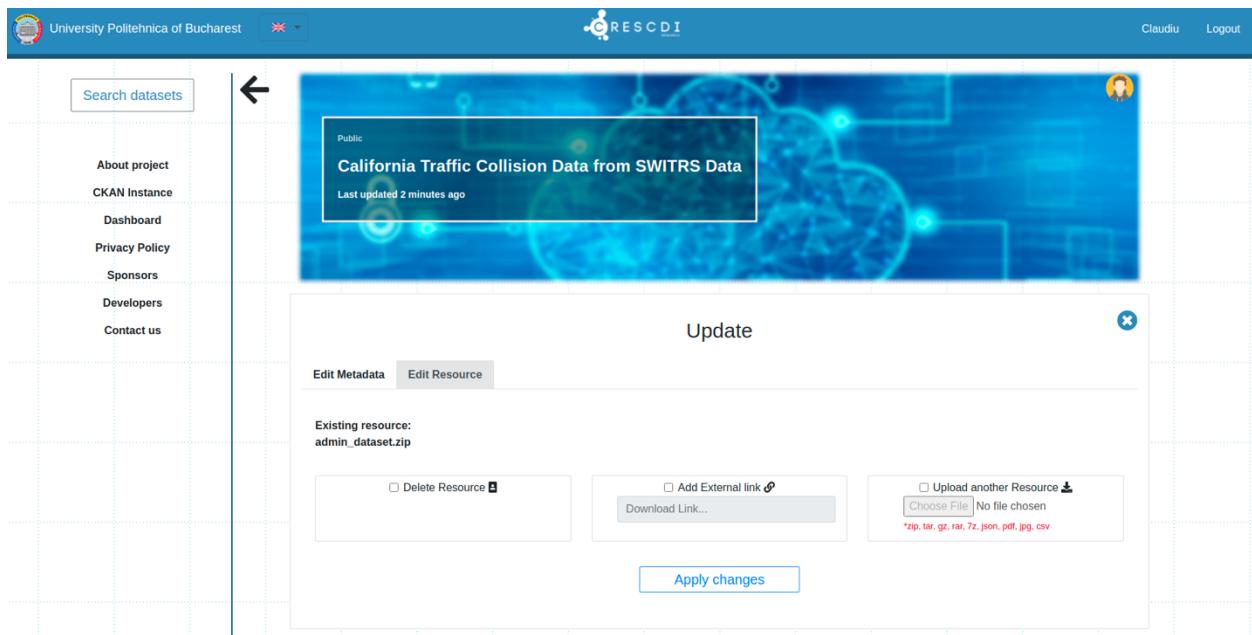


Figura 5. Actualizarea resursei

De asemenea, conform Figurii 6, aceasta pagina ofera utilizatorilor posibilitatea de a trimite un feedback despre setul de date curent prin acodarea unui rating (de la 0.5 la 5), insotit sau nu de un comentariu. In urma evaluarii, se trimit calificativul in back end, iar informatiile despre rating ale setului de date sunt actualizate: valoarea medie a rating-ului si numarul total de rating-uri acordate.

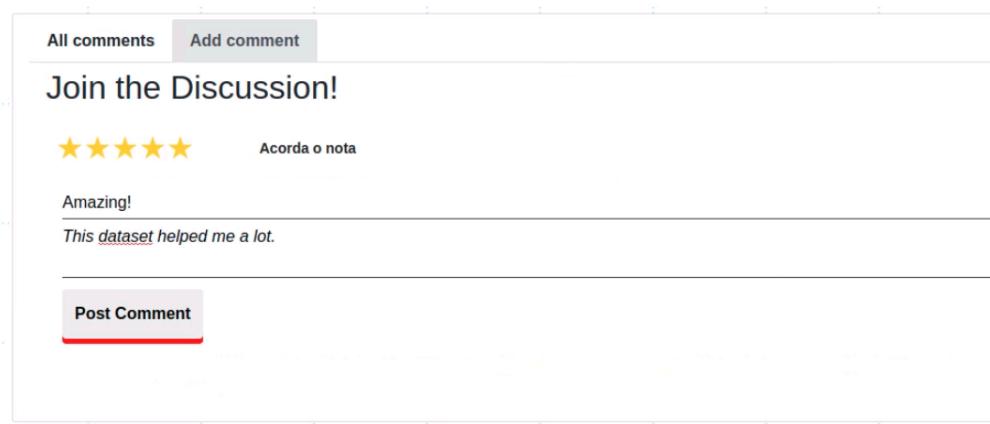


Figura 6. Acordarea unui calificativ

2.4 Informatiile utilizatorului autentificat

La autentificarea utilizatorului pe platforma, se genereaza un buton cu numele acestuia, care are rolul de a directiona utilizatorul catre o pagina care cuprinde toate informatiile ce ii sunt asociate (Figura 7).

The screenshot shows the CRESCDI user profile for Sergiu. At the top, there is a header with the University Politehnica of Bucharest logo, a search bar, and a user menu with 'Sergiu' and 'Logout'. Below the header is a large portrait of Sergiu. To the left of the portrait, his user information is listed: Username: Sergiu, Country: Romania, Email: sergiu@upb.ro, Public datasets: 2, and Private datasets: 0. To the right of the portrait is a search bar with dropdowns for 'All domains', 'All countries', 'All Data Formats', and 'All Downloads', and a 'QSearch' button. Below the search bar are filters for 'All tags', 'Author', 'Year', and 'Dataset title', along with a 'Sort By' dropdown set to '7'. The main content area displays two datasets: 'Multi-step ACR tasks run the world dataset' (Title: Multi-step ACR tasks run the world dataset, Date: 2020, Type: IT, Format: tar.gz, Location: Romania, Download Link, Delete button) and 'Small Change Dataset' (Title: Small Change Dataset, Date: 2004, Type: CHEMISTRY, Format: zip, Location: Canada, Download Link, Delete button). At the bottom of the page are links for 'Upload dataset', 'Search datasets', 'Made with ❤ at UPB', and 'SUPPORT'.

Figura 7. Pagina utilizatorului

In partea stanga se regasesc cateva informatii despre utilizator, precum email, tara de origine si numerele de seturi de date publice, respectiv private.

Partea centrala cuprinde panoul de cautare prezentat anterior si personalizat pentru cautarea seturilor de date proprii, indiferent de confidentialitatea acestora. Fiecare element din rezultat ofera posibilitatea de a fi sters. Totodata, este sters si pachetul asociat din instanta Ckan impreuna cu resursa, daca exista.

La nivelul platformei, se poate seta proprietatea SOFT_DELETE care reprezinta un flag pentru a marca varianta de stergere. Daca flag-ul este setat pe true, stergerea unui set de date va avea ca rezultat doar marcarea acestuia in baza de date drept sters. Daca flag-ul este setat pe false, stergerea unui set de date va avea ca rezultat disparitia acestuia din baza de date (hard delete). De asemenea, in cazul activarii mecanismului de stergere soft, am definit un cron-job care se executa zilnic si are rolul de a sterge hard toate seturile de date care au fost sterse soft cu mai mult de o perioada de timp configurabila. Posibilitatea de stergere soft are rolul de a oferi pentru o anumita perioada de timp posibilitatea de restaurare (rollback) in cazul stergerii accidentale a unui set de date. De asemenea, job-ul este configurabil prin intermediul proprietatilor asociate.

2.5 Instanta CKAN

Pentru platforma noastra am ridicat o instantă CKAN locală pentru a gestiona fisierile utilizatorilor. În cadrul acesteia am definit o organizatie (Figura 8), Crescdi și am creat un service account folosit în cadrul platformei. La încarcarea unui set de date pe platforma noastră se creează un pachet în cadrul instantei care conține câteva metadate de bază. Dacă setul de date conține o resursă internă, aceasta se asociază pachetului, ambele id-uri fiind salvate în baza de date pentru asocierea cu setul de date.

The screenshot shows the CKAN interface for the organization 'Crescdi'. At the top, there's a navigation bar with links for 'Datasets', 'Organizations', 'Groups', 'About', and a search bar. Below the header, the organization's logo (a circular emblem for Politehnica University Bucharest) is displayed, followed by the organization's name 'Crescdi' and a note: 'There is no description for this organization'. It shows 0 followers and 8 datasets. The main content area displays a summary of datasets: '8 datasets found' with one entry: 'Multi-step ACR tasks run the world dataset' (No description). There are tabs for 'Datasets', 'Activity Stream', and 'About', and a 'Manage' button.

Figura 8. Organizatie CKAN

De asemenea, apelurile de actualizare și stergere a unui set de date pe platforma noastră au efect și asupra pachetului asociat din cadrul instantei CKAN.

2.6 Baza de date

Am ales să utilizam o baza de date NoSQL pentru a putea profita de o indexare și căutare rapidă a datelor. Astfel, baza de date utilizată de platforma noastră se regăsește într-un container Docker construit pe baza unei imagini ElasticSearch preluate din Docker Hub. La nivelul bazei de date avem un index corespunzător componentei login, un index corespunzător metadatelor din seturile de date, îndecsi asociate domeniilor, tagurilor și comentariilor și un index la care este regăsită maparea dintre țari și coordonate geografice.

Pe baza unui script de initializare a bazei de date, se populează indexul de locații cu maparea fiecarei țari la coordonatele geografice ale centrului acesteia, indexul seturilor de date este

initializat cu 35 de intrari mock-uite, corespunzatoare unor seturi de date de test, iar indexul pentru autentificare este initializat cu 2 intrari specifice unor conturi de test.

2.7 Sistem de raportare si vizualizare a datelor

Am decis sa concepem un dashboard intuitiv pentru a putea vizualiza in timp real datele stocate in baza de date. Astfel, am creat un container Docker pe baza unei imagini Kibana, care este linkat la containerul bazei de date ElasticSearch pentru a permite accesul la informatiile stocate.

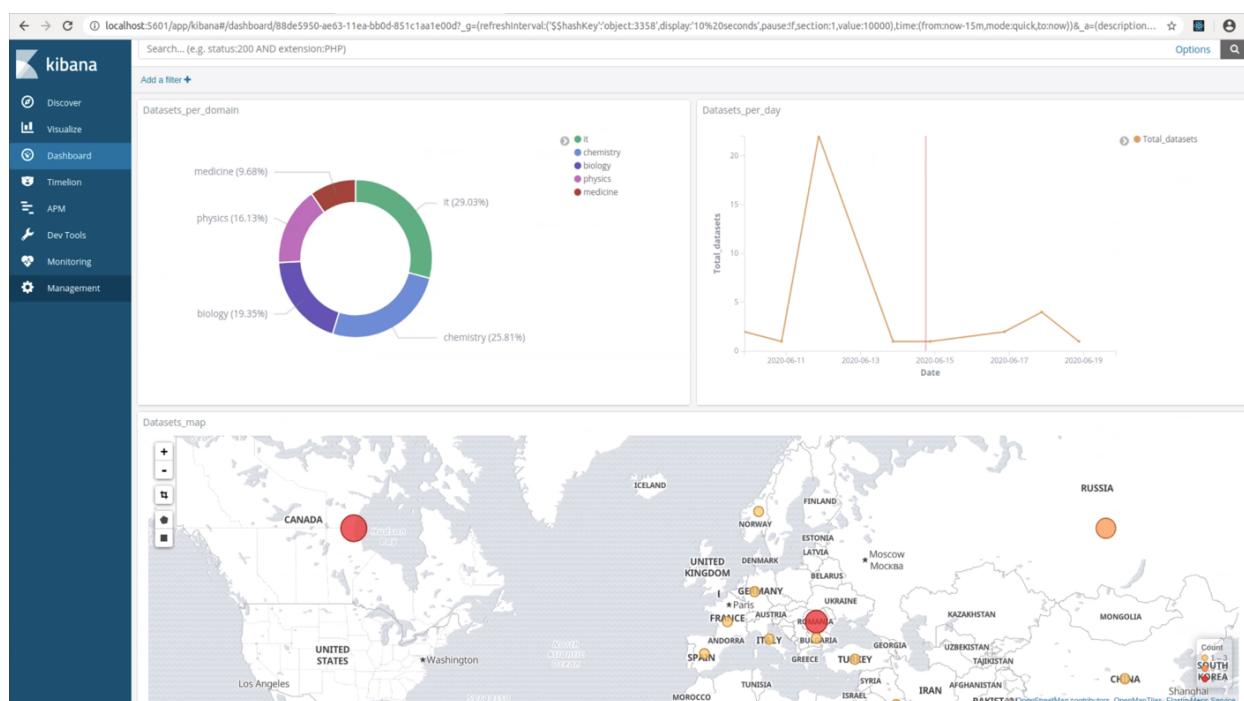


Figura 9. Dashboard Kibana

Dashboard-ul implementat se poate observa in Figura 9. Am creat 3 diagrame prin care evidențiem procentul de seturi de date incarcate din fiecare domeniu, o statistică privind evoluția numărului de încarcări zilnice pe platforma noastră și o hartă globală care ilustrează tarile de proveniență a seturilor de date și numărul de încarcări la nivel de țară.

3 IMBUNATATIRI VIITOARE

- CKAN harvesting research

4 CONCLUZII

Proiectul curent urmărește extinderea platformei CRESCDI prin posibilitatea de a putea fi încarcate și seturile de date utilizate în întocmirea articolelor științifice. Acestea sunt importante pentru că oferă posibilitatea altor cercetatori de a verifica rezultatele curente și de a reutiliza datele în scopul unei posibile continuari a procesului de cercetare.

În momentul de fata, atât din perspectiva back end, cât și din perspectiva front end, aceasta platformă auxiliară oferă o funcționalitate completă pentru gestionarea seturilor de date ale utilizatorilor, urmand să fie extinsă în urmatoarea perioadă prin furnizarea aspectelor menționate la punctul 3.