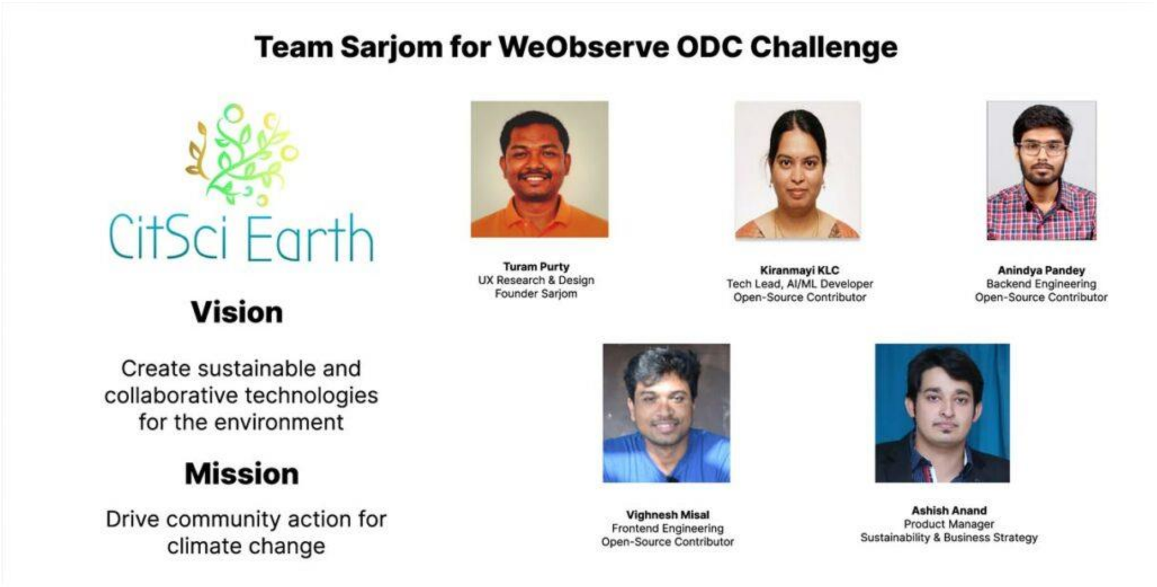
CITSCI MANAGER

WeObserve Open Data Challenge – Final Report

May 15 – July 31

Team members

Turam Purty, Kiranmayi KL Chandra, Vignesh Misal, Ashish Anand (Sarjom, CitSci Earth Lab)



Team Sarjom for WeObserve ODC Challenge

CitSci Earth

Vision

Create sustainable and collaborative technologies for the environment

Mission

Drive community action for climate change

Turam Purty
UX Research & Design
Founder Sarjom

Kiranmayi KLC
Tech Lead, AI/ML Developer
Open-Source Contributor

Anindya Pandey
Backend Engineering
Open-Source Contributor

Vighnesh Misal
Frontend Engineering
Open-Source Contributor

Ashish Anand
Product Manager
Sustainability & Business Strategy

Application functionalities

The application developed for the WeObserve Open Data Challenge is called "CitSci Manager". **CitSci Manager** is a special feature/tool that we integrated into the existing citizen science data aggregation platform, CitSci Earth (www.citsci.earth) using datasets provided by the WeObserve Open Data Challenge 2020. CitSci Manager helps citizen science researchers and volunteers save valuable time while exploring open datasets spread across a variety of data formats and greatly reduce the time required to clean, analyze, and organize datasets.

CitSci Manager has been developed to accomplish the following goals:

1. Developing an open-source infrastructure for data-interoperability in citizen science.
2. Engaging youth in citizen science research and climate change education.
3. Providing recognition and attributions to datasets collected by citizen science volunteers.

The application functionalities are designed using principles of Human-Centered Design (HCD) and Information Design & Architecture to provide a seamless user experience for end-users. (Refer to the demo video in the links shared below) These functionalities are as follows:

1. Users can upload a sample of their research datasets into entities called projects.
2. Users can create a unique name for their project and attach a license to it.
3. Users can select the type of files they want to upload.
4. Currently, **we support only images and CSV files.**
5. **RAW Files** such as images, videos, voice notes, and sensor measurements that have complex encodings are stored separately in a file system.
6. **Document Files** such as CSV, text, word, or excel that contains metadata information about RAW files are stored separately in a flexible document-based database.
7. Both RAW and document files can be processed for metadata retrieval using AI/ML techniques.
8. The metadata information of different document files from different sources(projects) can be fetched into a single screen.
9. Users can quickly perform JOIN and SELECT operations on the uploaded datasets using a user-friendly interface.
10. Users are notified via email with a unique link once the files are processed by the backend server.
11. Users can download their datasets or recombine with other datasets for analysis.

For the minimum viable product (MVP) submission, CitSci Manager was developed using Python, Jinja Templates, and MongoDB. The code repository for the submission can be accessed below:

- **Summer Submission** - https://github.com/WeObserve/OpenDataChallenge-CitSciManager-Alpha-version

The current version of CitSci Manager is closely integrated with the open-source platform, CitSci Earth and the technology stack for the MVP has been upgraded to Java (Spring framework), React (Gatsby), and MongoDB.

The code repository for the production instance can be accessed below:

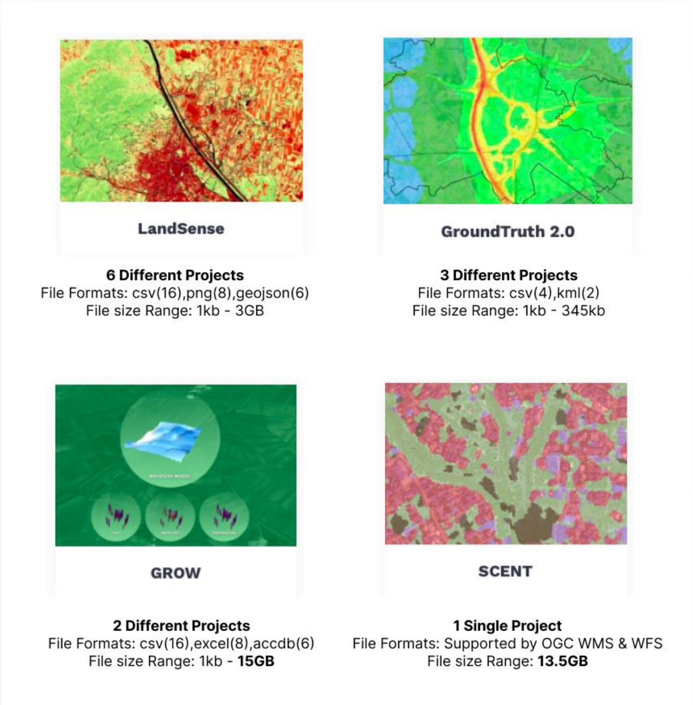
- **Front End Repository** - https://github.com/WeObserve/OpenDataChallenge-CitSciManager-Production-FrontEnd
- **Back End Repository** – https://github.com/WeObserve/OpenDataChallenge-CitSciManager-Production-BackEnd

A demo of the application can be accessed from the links below:

Summer MVP Demo - https://www.youtube.com/watch?v=oSU707XhHpQ
CS SDG Conference Demo - https://www.youtube.com/watch?v=N-gIDvn42Ok

Implementation considerations

CitSci Manager is a cloud-based application. Our application is designed in such a manner that it captures different types of files and datasets from the front end and stores them in the most optimal manner for the most optimal storage/archival and recovery. The design for the application was accomplished after analyzing all the datasets that were provided for the Open Data Challenge. (Figure- 1).



LandSense

6 Different Projects
File Formats: csv(16),png(8),geojson(6)
File size Range: 1kb - 3GB

GroundTruth 2.0

3 Different Projects
File Formats: csv(4),kml(2)
File size Range: 1kb - 345kb

GROW

2 Different Projects
File Formats: csv(16),excel(8),accdb(6)
File size Range: 1kb - **15GB**

SCENT

1 Single Project
File Formats: Supported by OGC WMS & WFS
File size Range: **13.5GB**

Figure 1: Meta-data Analysis of WeObserve Datasets across projects
Image Source: WeObserve ODC 2020 Website

We conducted a metadata analysis and found critical issues with the citizen science datasets that were spread across a variety of projects, file formats, schemas, and sizes. For a 1-month innovation challenge, it took us 2 weeks just to analyze the metadata of all the datasets and try to come up with a possible solution. We realized that this was a problem itself in a lot of citizen science datasets and defined our problem statement as,

"How can we make citizen science datasets accessible across a diversity of projects?"

The application was designed and implemented within this scope. Amazon Web Services (AWS) was used to store large datasets, such as the 15GBs GROW dataset in CSV format in S3 Buckets (Amazon File Storage System is called S3). Processing such large files took us a lot of computation time for the application even when using Big-Data Processing tools such as DataBricks and Apache Spark. Hence, a decision was taken to create sample CSV files from large files. These files would have a limited number of rows, but with all attributes of the datasets and hence, would be useful for our use case as we would be easily able to fetch the header attributes of the datasets into our front-end for users to manipulate. Sample CSV data files were processed from the GROW Observatory provided by the WeObserve team and the external weather datasets obtained from the National Oceanic and Atmospheric Administration (NOAA). An interface was designed where a user could fetch the attributes (CSV header) of the datasets and **join columns that were unique** and **select additional columns for analysis**. This was a task that we regularly had to do in order to understand the various WeObserve datasets provided for the challenge. CitSci Manager became a generic tool that could be easily customized to a wide range of datasets spread across different formats in citizen science projects. Once the user submitted the request, depending on the size of the sample files, a job would run and finish the data merging process. An automated email is sent by a scheduler to the user with a unique S3 URL to download the transformed CSV file for analysis. This would significantly reduce the time and the burden on the end-user to analyze datasets across projects and our architecture (Figure 2) directly addresses the data-interoperability issue of citizen science projects in an efficient manner.

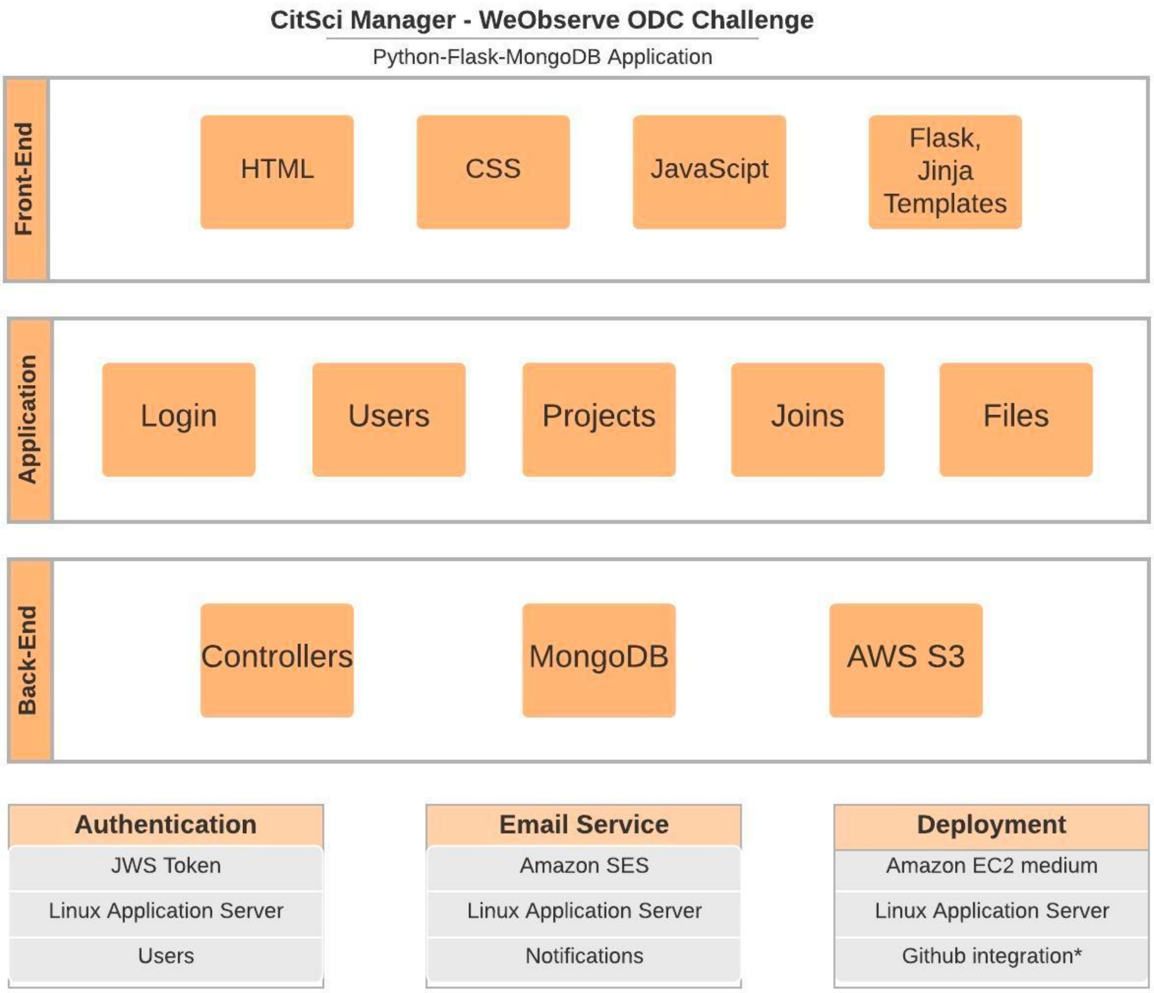


Figure 2 - Software Architecture of CitSci Manager.

Some additional general implementation considerations are as follows:

1. Backend services are separated from the frontend to the greatest extent possible.
2. MongoDB database is used to store the mapping of the users, files, and projects.
3. With scaling and security in mind, the backend is divided into multiple layers, i.e., controllers, service layer, and data access layer.
4. The authentication is implemented using the JWT token system.
5. All file uploads are handled from the client-side to reduce the load on the backend.
6. The upload file size is limited to 2GBs from the front end.
7. A scheduler implemented using PySpark that runs every minute to transform the files and upload the transformed files to S3.
8. Frontend form validation is implemented with Flask-WTF and WTForms libraries.
9. The application is hosted using AWS EC2. The repositories are open sourced on Github.

10. GROW dataset is used with an external weather dataset from the National Oceanic and Atmospheric Administration (NOAA) to demonstrate the working of the application.

Expected impact

CitSci manager will enable researchers and organizations to host citizen science projects and publish a sample of their datasets and make their findings available for the citizen science community. Students, youth, and citizen scientists can then explore samples of datasets on our platform from a diverse set of projects and conduct their own hyperlocal analysis. The application will save crucial time and efforts for users to analyze citizen science datasets and open up avenues for civic engagement and collaborations.

The application is built with open source licensing and other people/organizations can pick up our code and build libraries to manage various forms of file formats used across citizen science projects. The main actors/beneficiaries of the application will be:

1. Citizen Science, Climate Change, and Environmental monitoring agencies and researchers who spend a lot of time and effort to go through citizen science datasets spread across projects.
2. Citizen Science volunteers, amateur enthusiasts, and students who wish to participate in citizen science projects can generate new knowledge and data analysis.
3. Government and policy researchers can access datasets across a variety of projects that can help decision making in areas of biodiversity conservation, urban planning, infrastructure development, and industrialization.

The impact of our work directly contributes to the United Nations Sustainable Development Goals (SDGs) such as,

1. Providing **quality education** to youth and researchers by improving the interoperability of datasets.
2. Contributes to the **sustainable industry, innovation, and infrastructure** for environmental monitoring through citizen science.
3. Contributes to the development of **sustainable cities and communities** that can make data-backed decisions about their local environment and biodiversity.
4. Contributes to a better understanding of **climate change.**
5. **Builds partnerships** amongst researchers, policymakers, citizen science communities, and students working towards achieving sustainable development goals.

Future outlook

Citizen Science data can become very instrumental in civic action and data-backed decisions about the environment, biodiversity conservation, urban planning, infrastructure development, and industrialization. Our work will directly impact existing data contributions, improve awareness, and participation in citizen science while strengthening community partnerships, improving data quality, and civic decision-making. These mutually beneficial endeavors will strengthen local community resilience by building collective and individual capacity to address climate change threats, be they environmental or public health in nature.

We aim to expand upon our summer work and collaborate with research agencies across the world in developing a new suite of open source collaboration software called “Community Resource Planning” (CRP), that will empower not-for-profits and citizen science research observatories to easily set up a scalable, customizable, and robust cloud-based digital ecosystem and effectively manage projects, datasets, and community engagement activities/discussions. Our goal is to design the infrastructure in such a way that it promotes open data, engagement, and interoperability between a diversity of citizen science projects across the world.

The collaboration suite will leverage Artificial Intelligence/Machine Learning (AI/ML) models to improve data quality, gamify community participation for hyperlocal environmental monitoring, develop Application Programming Interfaces (API) for citizen science data aggregation, collaboration, and publication, and contribute to the development of citizen science data interoperability standards in collaboration with Open GeoStandard Consortium (OGC).

- END OF DOCUMENT -

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no 776740.