

Web Tools for Genomic Epidemiology - Paraguay 2023

Module Lead: Silvia Argimon & Julio Diaz Caballero

Table of contents

- Background
- Introduction
- Learning outcomes
 - Collect field data with Epicollect
 - Merge data with Data-flo
 - Microreact
- Resources
 - Install and test Epicollect5 on your mobile phone
 - Sign-in to data-flo
 - Sign-in to Microreact
 - Web colours
- Bonus activities

Background

Genomic epidemiology aims to understand the emergence and dissemination of high-risk clones within pathogen populations with the ultimate goal of implementing evidence-based interventions to protect public health. High-risk clones are subpopulations/strains/variants of a pathogen that carry risk elements, such as antibiotic resistance or virulence determinants, and thus pose a potential risk to public health.

The increased genetic resolution afforded by genomic data is useful from global to local geographic scales, and it has proven particularly useful for the investigation of pathogens that exhibit little genetic variation (e.g. Wong *et al.* 2016), and for outbreak investigations (e.g. Hendriksen *et al.* 2011, Eppinger *et al.* 2014). During outbreak investigations (and other genomic epidemiology studies), epidemiological data from patients is collected by healthcare professionals. Antimicrobial susceptibility data, species identification, and any further phenotypic or molecular characterization of the isolates is often generated by the laboratories linked to healthcare facilities and/or by the reference laboratory. Ideally, these different sources of data are stored in a centralised surveillance system and database, such as WHONET. However, these systems rarely incorporate genomic data produced by bioinformaticians. Genomic data may include sequence/assembly quality, genotyping/genoserotyping information, presence/absence of known risk elements, and clustering of the isolates based on genetic similarity –usually in the form of a phylogenetic tree or a minimum spanning-tree.

The job of a genomic epidemiologist often starts by combining data from diverse sources that might not be complete or standardised to facilitate the identification of relevant patterns for meaningful interpretation. Once integrated, the epi, lab, and genomic data can be interrogated over the structure of the tree to identify clusters of interest that will place the isolates in or out of the outbreak.

The **Centre for Genomic Pathogen Surveillance** develops free web applications for data collection, integration, visualisation, and analysis of genomic epidemiology data. **Epicollect5** is a mobile & web application for free and easy data collection. It provides both the web and mobile applications for the generation of forms (questionnaires) and freely hosted project websites for data collection. Projects are created by using the web application at five.epicollect.net, and then downloaded to the device to perform the data collection. Data are collected in the field using multiple devices and all data can be viewed on a central server (via map, tables, and charts). **Data-flo** is a system for customised integration and manipulation of diverse data via a simple drag and drop interface. Data-flo provides a visual method to design a reusable pipeline to integrate, clean, and manipulate data in a multitude of ways, eliminating the need for continuous manual intervention (e.g., coding, formatting, spreadsheet formulas, manual copy-pasting). **Microreact** allows you to upload, visualise and explore any combination of clustering (trees), geographic (map) and temporal (timeline) data. Other metadata variables are displayed in a table. You can specify colours and/or shapes to display on the map, tree and/or timeline. A permanent URL is produced for you to share your Microreact.

Introduction

Aim: To use web and mobile tools for the collection of epi data in the field, its integration with lab and bioinformatics data, and to finally visualise it together to detect a putative outbreak.

Note: This exercise is a work of fiction. Any resemblance to a real cholera outbreak is purely coincidental.

During January and February 2023, fifty-eight patients reported to eight hospitals in Asunción with acute watery diarrhoea and vomiting. *Vibrio cholerae* was isolated from the stool samples of forty patients (all belong to serotype Inaba), and one colony from each confirmed case was sequenced on Illumina MiSeq. An epidemiological investigation revealed that more than half of the culture-confirmed cases regularly visited the Ñu Guazú Park, where water fountains and other water sources are accessible, and which is crossed by the Arroyo Itay. This suggested that the outbreak could be caused by contaminated water in the park. The epidemiological and demographic data is summarised on the file `epi_data.csv`.

A team of disease detectives will be deployed around Ñu Guazú Park today to investigate the potential source of the outbreak using Epicollect5, data-flo,

and Microreact, free mobile and web applications developed by the Centre for Genomic Pathogen Surveillance for easy data collection, integration and visualisation.

Learning Outcomes

At the end of this session the participants will be able to:

1. Collect data in the field with Epicollect5.
2. Merge epidemiological, laboratory and genomic data with data-flo
3. Create a Microreact project to visualise epidemiological, laboratory and genomic data

1. Collect field data with Epicollect

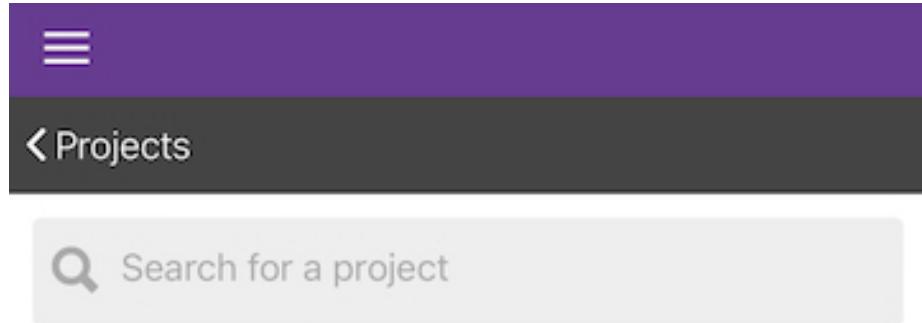
<https://five.epicollect.net/>

The disease detectives will work in pairs, and will be provided with a **collection tube** with a **barcode** sample identifier to collect a sample of water, and will use their **mobile phone** to collect associated metadata in the field.

Make sure your phone is connected to WiFi. Open the **Epicollect5** app on your mobile phone.



Add the project **Asuncion 2023 Cholera Outbreak** by typing into the search box –you will need to be connected to WiFi.



Using the map of the Ñu Guasú park provided, walk to the location of your collection point (water source). Locations are indicated by the red dots labelled a-j. You will not need internet access to collect data in the field.

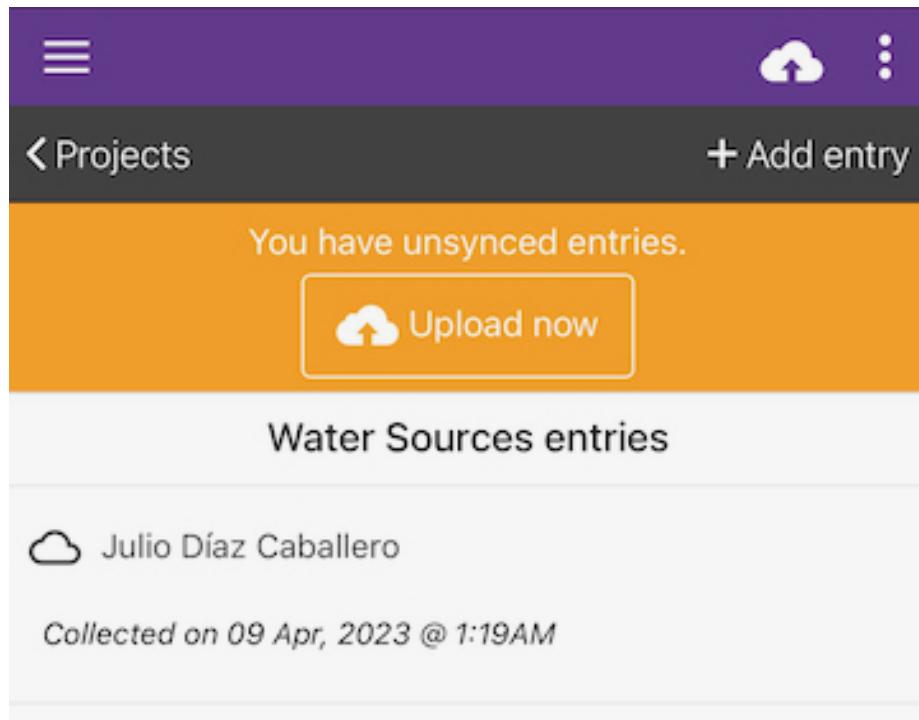
PASTE IMAGE OF MAP WITH LOCATIONS (to be finished)

Open the **Asuncion 2023 Cholera Outbreak** project and add an entry. Follow the form to collect a water sample and associated metadata.

Your team will collect only one water sample, but both of you can collect the associated data on Epicollect5. However, *make sure that only one of you uploads the data later on to avoid duplications.*

Return to the Epidemic Intelligence Center (i.e. the classroom).

One of the pair members only. **Upload your entry (and image)** to the Epicollect5 server –make sure you are connected to the WiFi.



Submit your water sample to the instructor, which will be sent to the reference lab for culture confirmation.

Once all entries are uploaded by the disease detectives we will take a look at the data together on <https://five.epicollect.net/project/asuncion-2023-cholera-outbreak>

Answer the following questions:

1. What is the most common type of water source?
2. Was it possible to collect a sample from all 13 sources? If not, what were the reasons?

2. Merge data with Data-flo

<https://data-flo.io/>

Note: you need to sign-up for data-flo and Microreact. See instructions in the Resources section. Creating your own account will allow you to manage and edit your projects.

The reference lab has sent you the culture results on an Excel file called `lab_results.xlsx`. The lab reported that **2 out of 13** water samples from Ñu Guasú park were positive for *Vibrio cholerae*. This immediately prompted the closure of the **2 water sources**.

Does this confirm that the source of the outbreak was contaminated water from Ñu Guasú park?

One colony from each source was sequenced on Illumina MiSeq by the reference lab. A maximum likelihood phylogenetic tree (`tree.nwk`) was inferred from the genomes of the 34 clinical samples and 2 environmental (water) samples. Six genomes from a previous outbreak (Jan-Feb 2020) were also included in the tree inference and their associated data added to the `epi_data.csv` file.

The disease detectives now have the information needed for the investigation in the following formats:

- `epi_data.csv` Epi data from 34 clinical cases and 6 cases from previous outbreak
- `Epicollect project` Metadata of 13 water sources from Ñu Guasú Park
- `lab_results.xlsx` Culture and serotyping results
- `tree.nwk` Phylogenetic tree of 34 clinical cases, 2 culture-positive water samples, and 6 cases from previous outbreak

The files are located in the VM folder **FOLDER NAME**.

We will combine data from these different sources with a data-flo workflow that takes the files above and the data from the Epicollect project as input, and creates as an output a Microreact project where the data can be visualised.

Open the data-flo workflow (<https://data-flo.io/run?prnvpvmKzq8rDKbPD9ZXAs>)

The screenshot shows the Data-flo web application interface. At the top, there is a navigation bar with the Data-flo logo, followed by links for TRANSFORMATIONS, DOCUMENTATION, and SIGN OUT. Below the navigation bar, the title 'Asuncion 2023 Cholera Outbreak' is displayed, along with creation and update timestamps: 'Created: 03/03/2023, 10:27:36 Updated: 09/04/2023, 01:39:59'. There are also edit, copy, and download icons. A tabs menu at the bottom of this section includes INFO (selected), RUN, and API.

Description

This is an exercise designed for the WGCAC Genomics and Epidemiological Surveillance of Bacterial Pathogens (Asuncion, Paraguay) 16 - 21 April 2023

Inputs

- epicollect project url ([text](#))
- lab results excel ([file](#))
- epi_data csv ([file](#))
- tree ([file](#))

Outputs

- url ([url](#))

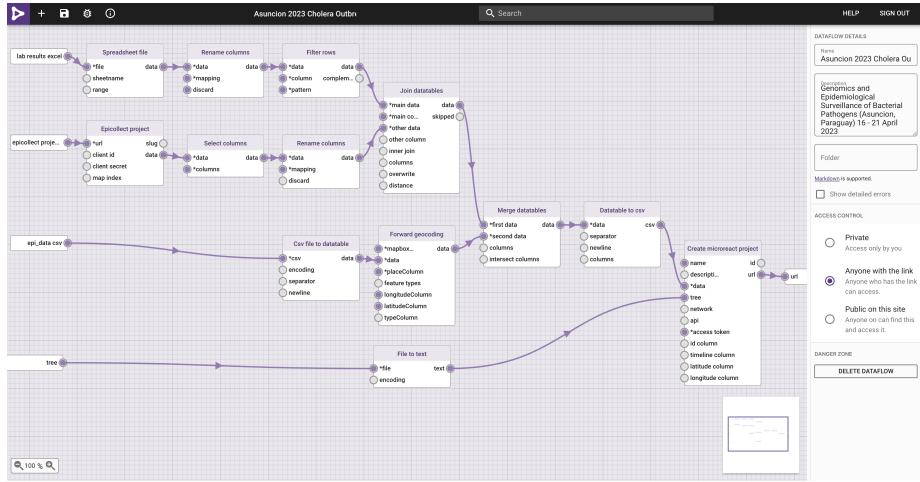
Copy the workflow to your own dataflo account.

The screenshot shows a copy of the 'Asuncion 2023 Cholera Outbreak' workflow. The title and timestamps are identical to the original: 'Created: 03/03/2023, 07:27:36 Updated: 17/04/2023, 09:44:23'. The edit, copy, and download icons are visible. The tabs menu shows 'INFO' selected.

Description

This is an exercise designed for the WGCAC Genomics and Epidemiological Surveillance of Bacterial Pathogens (Asuncion, Paraguay) 16 - 21 April 2023

This will open a copy of this workflow in your dataflo account.

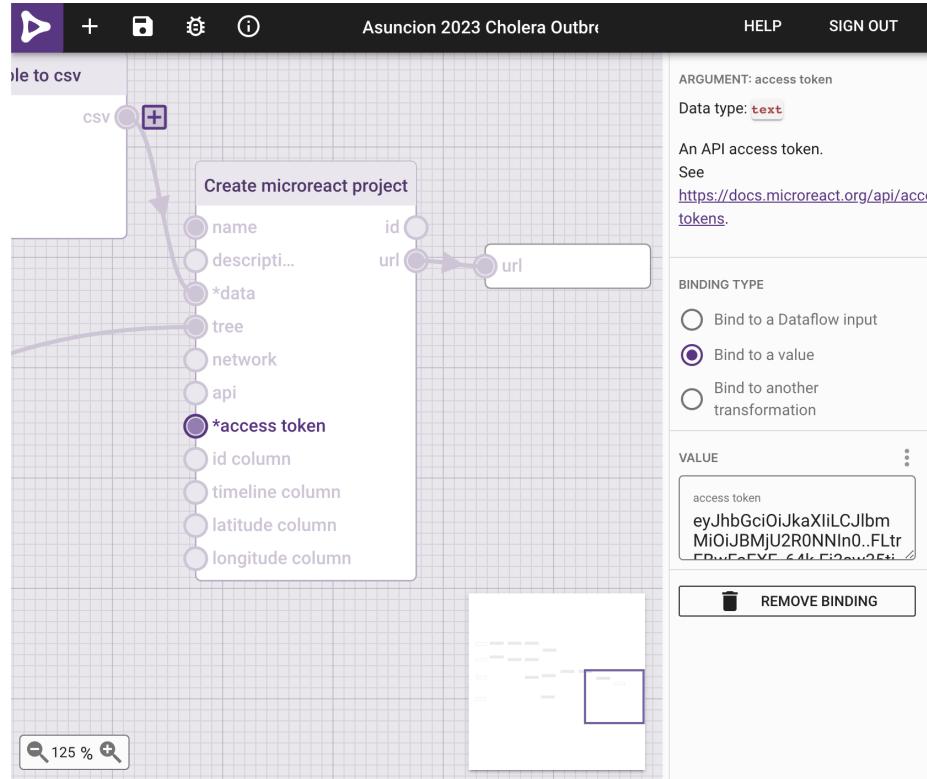


On a different browser tab, get your microreact API access token at <https://microreact.org/my-account/settings> (you must already have created your microreact account).

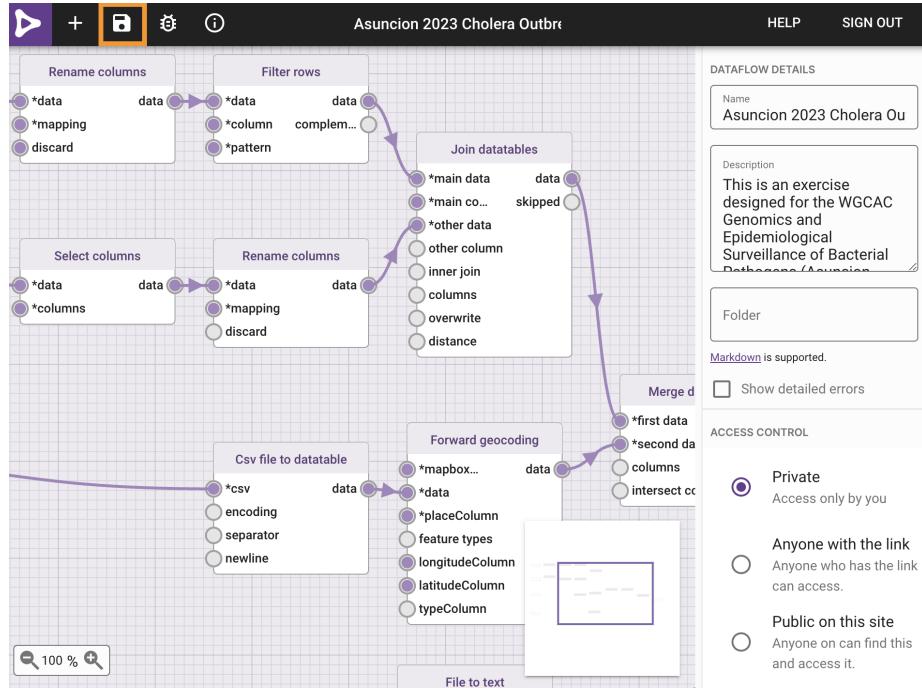
The screenshot shows the Microreact account settings page. Under 'API Access', there is a box labeled 'Access Token' containing the value 'eyjh'. To the right of the box is a copy icon. Below the token box is a 'REQUEST ACCOUNT DELETION' button.

Edit the workflow to include your own microreact API access token.

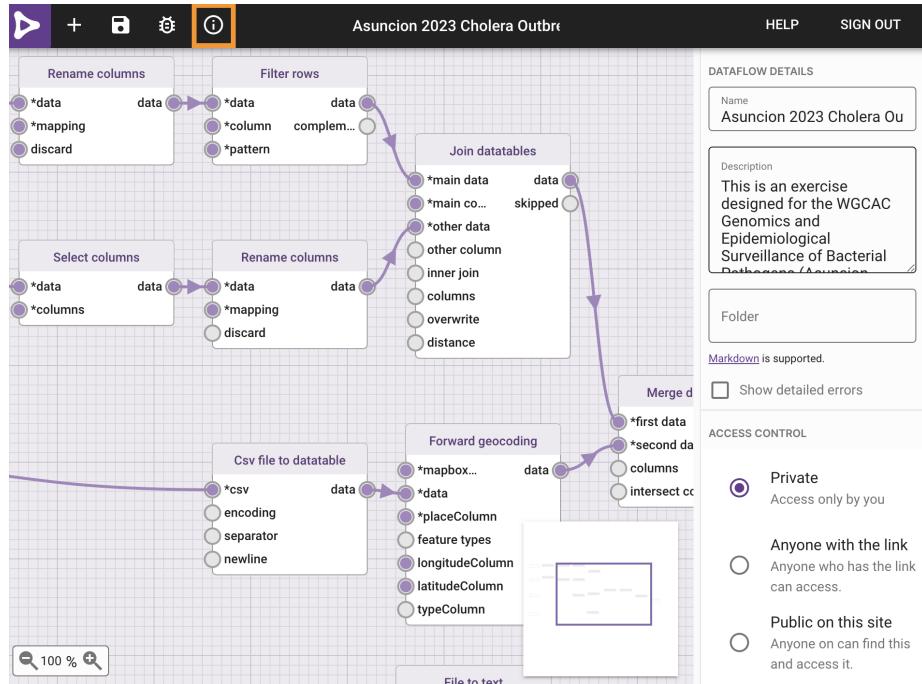
1. Click on `*access token` in the `Create microreact project` box
2. From the options on the right, select `Bind to value`
3. Paste your API access token in the `VALUE` box



Save your newly edited dataflo project by clicking on the **save** icon.



Now lets go to the implementation page by clicking on the *i* icon.



To run the workflow paste the url of the Epicollect5 project (<https://five.epicollect.net/project/asuncion-2023-cholera-outbreak>) and upload the files.

The screenshot shows the Data-flo interface for running a workflow. The top navigation bar includes links for TRANSFORMATIONS, DOCUMENTATION, and SIGN OUT. Below the navigation is a header for the 'Asuncion 2023 Cholera Outbreak' project, with creation and update dates. A toolbar with edit, refresh, and download icons is also present. The main area is divided into sections:

- Inputs:** Contains three file inputs:
 - 'epicollect project url': Set to <https://five.epicollect.net/project/asuncion-2023-cholera-outbreak>
 - 'lab results excel': Choose file lab_results.xlsx
 - 'epi_data csv': Choose file epi_data.csv
- tree:** An empty input field.
- RUN:** A purple button with white text, highlighted with an orange border.
- Outputs:** A section currently empty.

On the left side, there is a vertical numbered list:

- 1 Inputs
- 2 Run
- 3 Outputs

Click on **Run**. The Outputs box now shows the url of a Microreact project created by data-flo.

The screenshot shows the Data-flo web application interface. At the top, there is a navigation bar with the Data-flo logo, TRANSFORMATIONS, DOCUMENTATION, and SIGN OUT links. Below the navigation bar, the main title is "Asuncion 2023 Cholera Outbreak". Underneath the title, it says "Created: 03/03/2023, 10:27:36 Updated: 09/04/2023, 01:39:59". There are three tabs at the top: INFO (selected), RUN, and API. The main content area has a flowchart with three nodes: "Inputs" (checked), "Run" (checked), and "Outputs" (3 items). Below the flowchart is a "RUN AGAIN" button. At the bottom, there is a URL link: "url. <https://microreact.org/project/kdLCjjazR1oxsrYdCLjbrx-asuncion2023choleraoutbreak>".

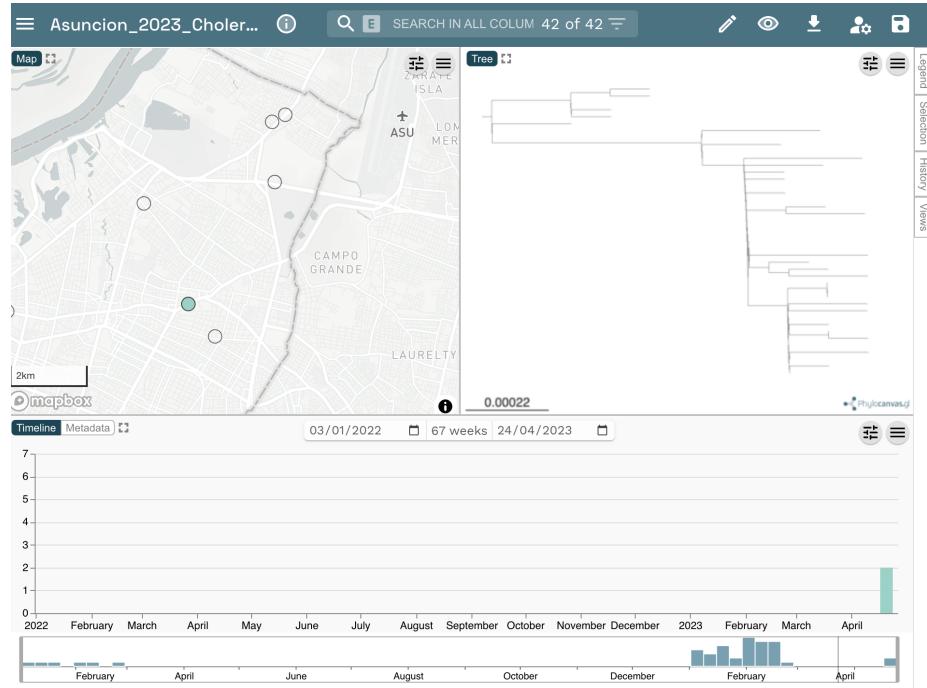
Before you open the Microreact project answer the following questions:

1. If later on we wanted to add additional entries (i.e. water samples) to the Epicollect project, could we use the same data-flo? If so, how would this be beneficial?
2. Do you use/know of any other methods to join data? If so, how do they compare with data-flo?

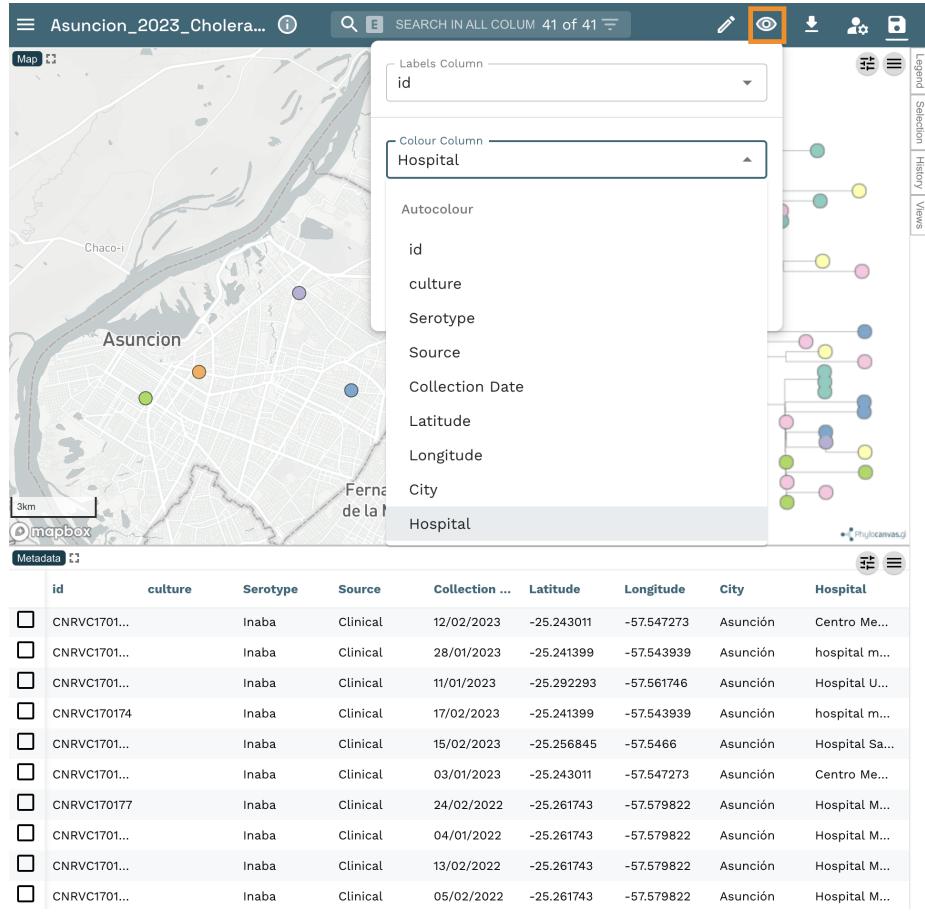
3. Microreact

<https://microreact.org/>

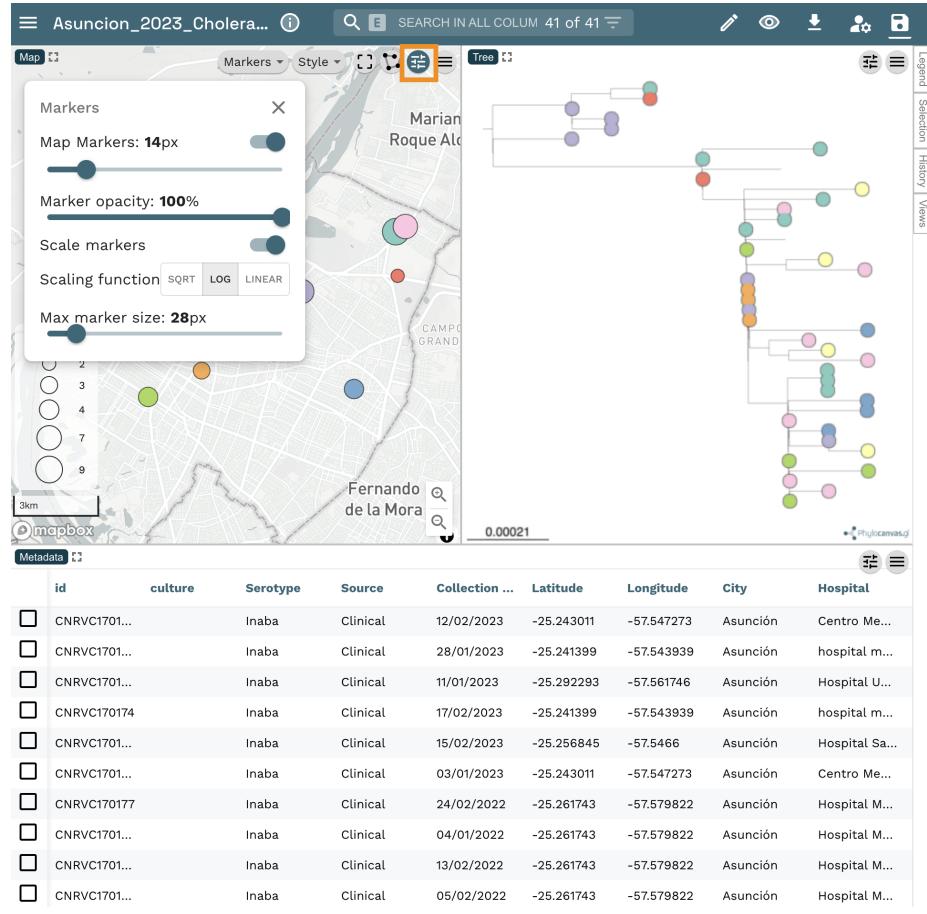
Open the Microreact link in the output from data-flo. It should look something like this:



Now change the colour column to “Hospital”. Click on the eye icon at the top left of the page. Select “Hospital” from the **Colour Column** dropdown list.



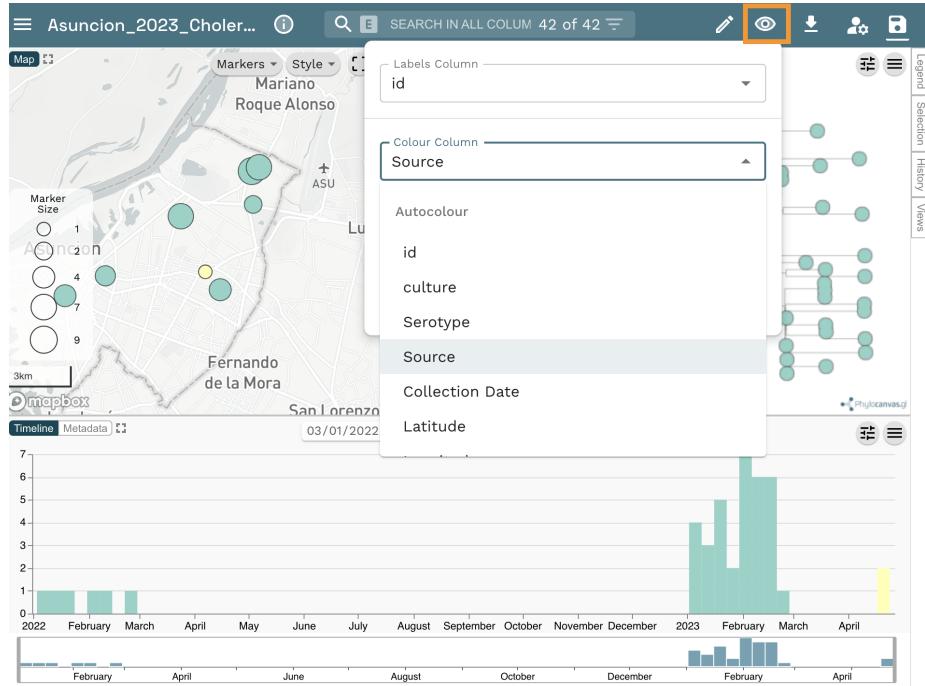
Scale the markers on the map by clicking on the sliders icon, then in the opening options click on **Markers**. Finally, toggle the **Scale markers** option on.



Answer the following questions:

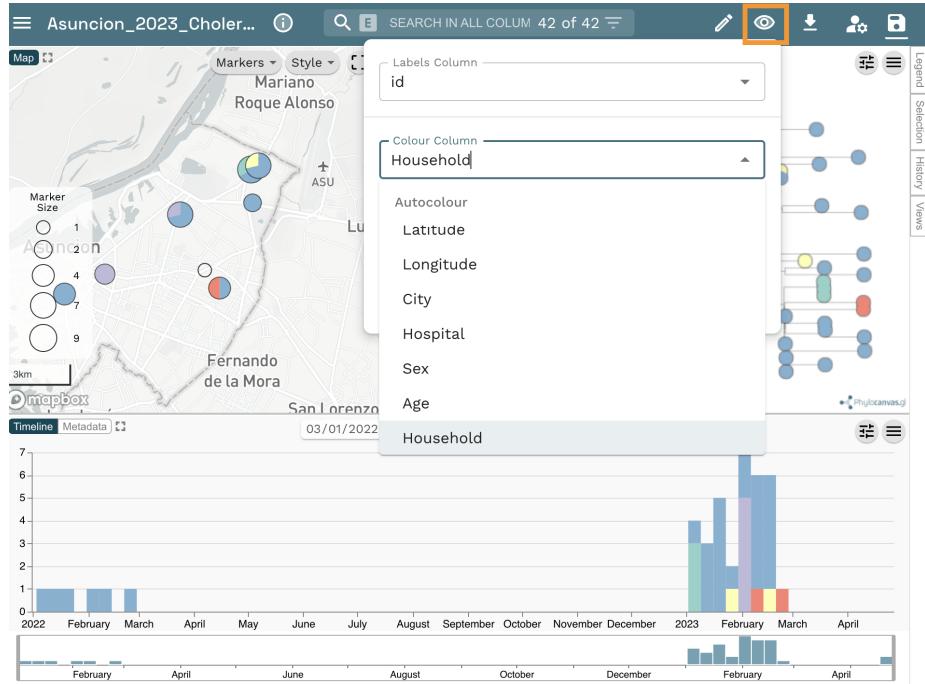
1. What do the markers on the map represent?
2. Which hospital reported the index (first) case? Tip: Click on the earliest sample on the timeline.

Change the colour column to “Source”. Click on the eye icon at the top left of the page. Then select “Source” from the Colour Column dropdown list.

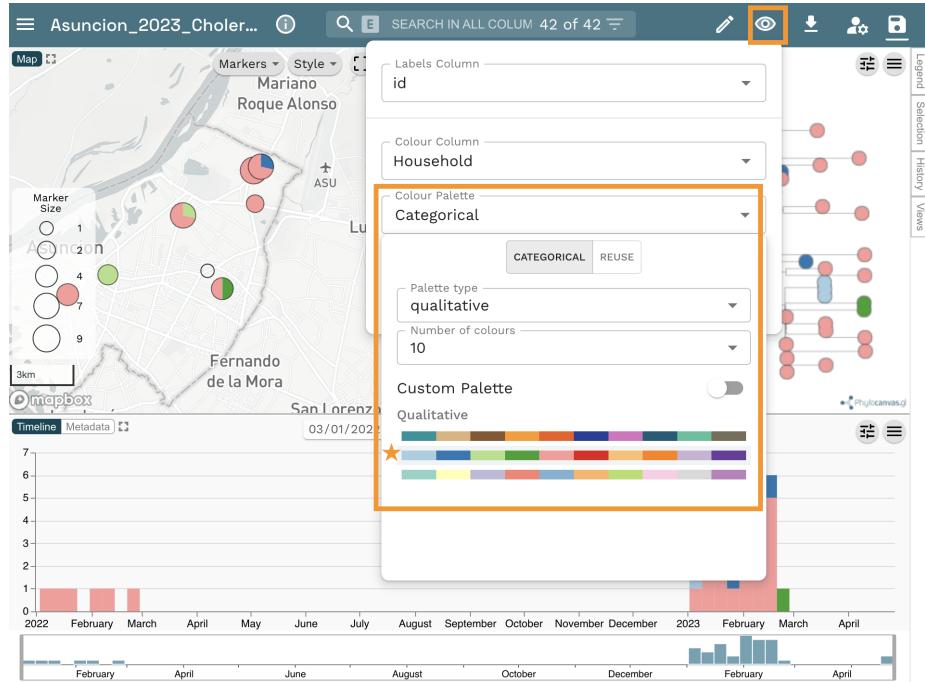


3. Can you confirm that the source of the outbreak is the contaminated water from Ñu Guasú Park?
4. Why is it useful to include genomes from a previous outbreak in the analysis? Is the current outbreak related to the past outbreak from Jan-Feb 2020?

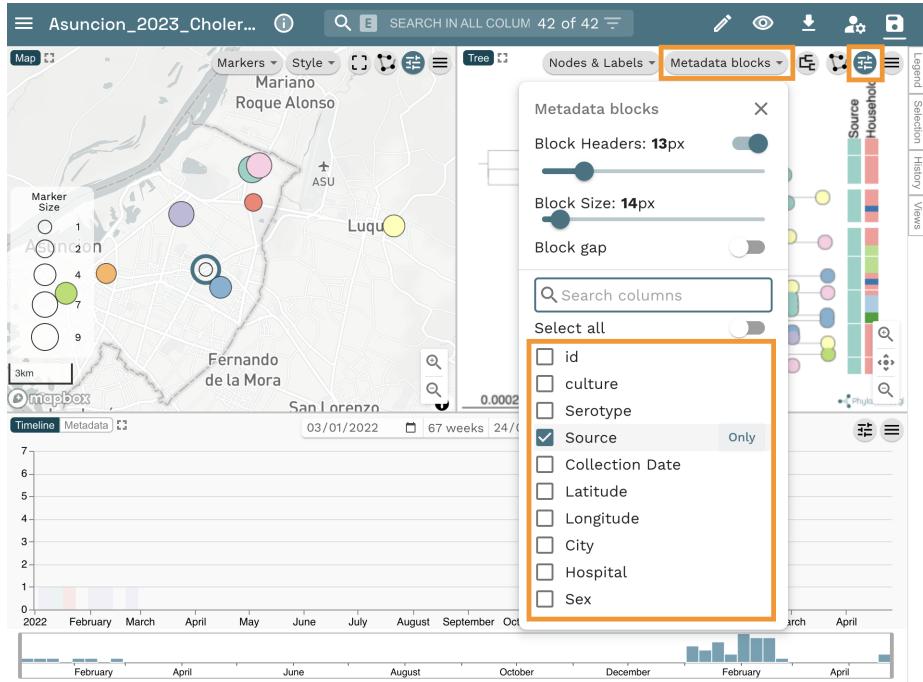
Microreact assigns colours automatically to every column in your metadata table, but you can customise them. Let's change the colour column to "Household". Click on the eye icon at the top left of the page. Then select "Household" from the Colour Column dropdown list.



Now Select “Categorical” from the **Colour Palette** option. Next Select “qualitative” from the **Palette type** dropdown list and “10” from the **Number of colours** dropdown list. Finally, select the second option from the list of palettes.



Change the colour column back to “Hospital” and add the columns “Source” and “Household” as metadata blocks. For this, select the slider icon from top right of the map panel, then, click on the **Metadata blocks** button. Finally, select “Source” and “Household” from the list of options.



5. Do any of the cases share a household? Taking into account the incubation period for cholera (a few hours to five days), does the data support intra-household transmission?

Resources

Install and test Epicollect5 on your mobile phone

Epicollect5 is available on Android and iOS. To install it on your mobile device, get it from Google Play or the Apple App Store. To test Epicollect5 on your phone, open the application and click on the pre-loaded EC5 Demo Project. Click on **+ Add entry** and follow the questionnaire. Please answer all the questions, including taking a photo with your phone's camera. Once you've answered the questions, save and upload the entry. Please note that you must be connected to the internet (or to a mobile network) to be able to upload an entry.

Sign-in to data-flo

<https://docs.data-flo.io/introduction/getting-started-sign-in>

Sign-in to Microreact

Head over to <https://microreact.org/my-account>, and sign up using your email address or any other authenticating option



Email

Sign in with Email

or

Sign in with Google

Sign in with Facebook

Sign in with Twitter (Legacy)

Web colours

Web colours are used on web pages, such as the microreact.org projects, and are usually specified in hexadecimal format preceded by a number sign (or hashtag sign if you were born before the Jurassic period). For example, #FFFFFF is the

hexadecimal code (or hex code) for the colour “white”, #000000 is the hex code for the colour “black”, and #FF00FF is the hex code for the colour “magenta”. For more information on the format of the hexadecimal code see Wikipedia page https://en.wikipedia.org/wiki/Web_colors

Useful websites to work with web colours:

ColorBrewer

<https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3>
Colour palettes for maps

ColorHexa

<https://www.colorhexa.com/> Get complementary colours, colour gradients, etc.

Coolors

<https://coolors.co/palettes/trending> Colour palette generator

Wes Anderson colour palettes

<https://github.com/karthik/wesanderson> Hipster colour palettes inspired by Wes Anderson movies

Bonus activities

1. Download data from the Epicollect5 project on the web.

- a. Go to the **epicollect** link provided above.
- b. Select **VIEW DATA**.

c. Click on Download and select your format of choice.

View	Delete	Edit	Title	Created At	Collector's Name and Last Name	Location Type	Please describe location	Type
			Julio Diaz Caballero	9th Apr, 2023	Julio Diaz Caballero	Outdoors		Ho

2. Create a microreact project using a metadata in csv format and a phylogenetic tree.

a. Click on the menu icon at the top left of the screen



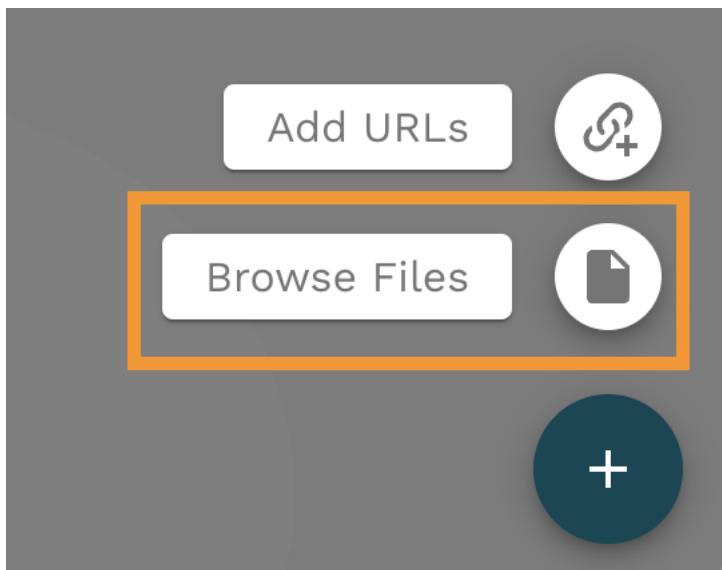
b. Click on Upload

A screenshot of the Microreact dashboard. At the top right, it says "Version 231". Below the logo, there are several menu items: "Showcase", "Upload" (which is highlighted with an orange border), "Documentation", "My Account", "Sign out", and "Send Feedback". To the right of the menu, there is a dark sidebar with the word "isua" partially visible. The main area of the dashboard shows a map of Europe with various colored dots representing data points.

c. Click on the + symbol at the bottom right of the screen.



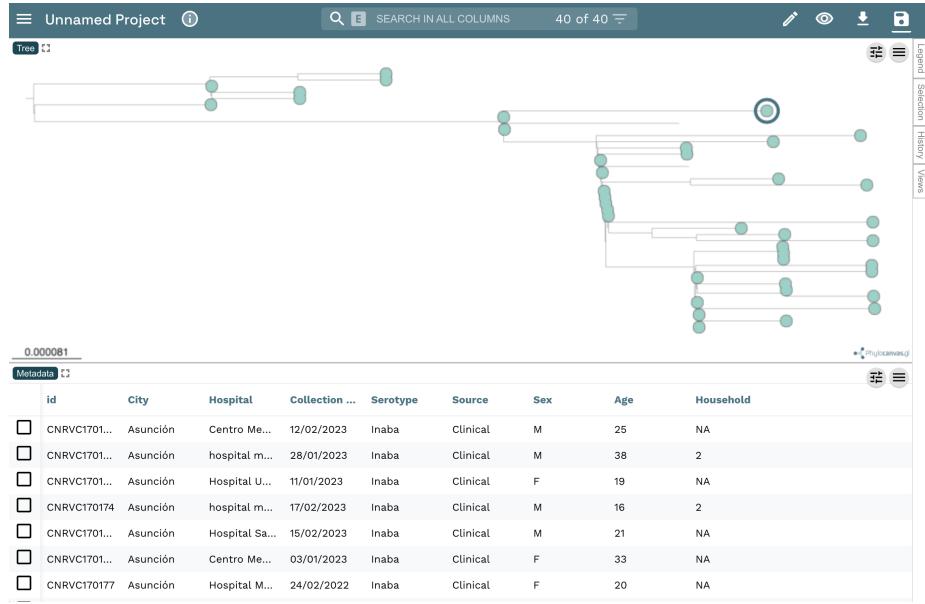
d. Select **Browse Files**



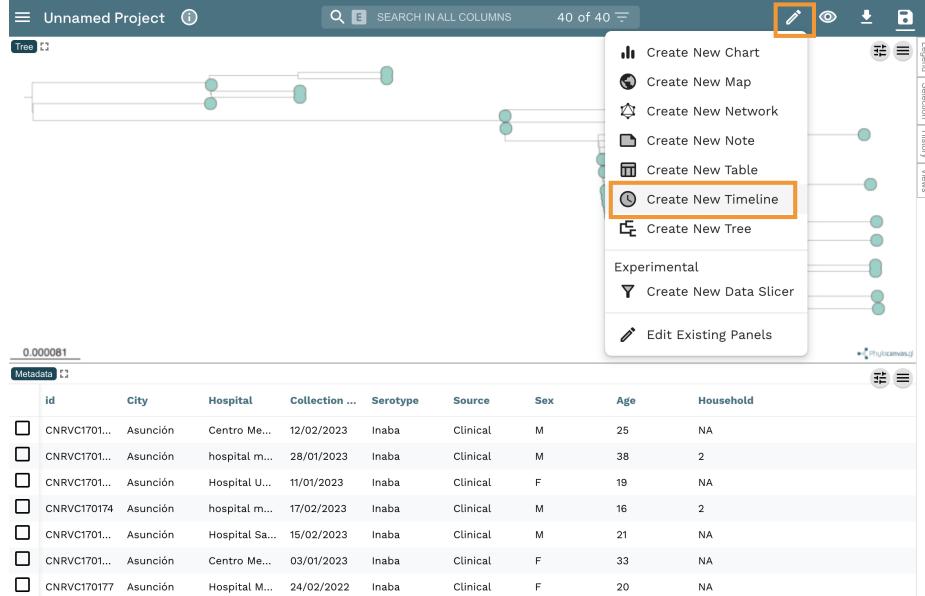
e. Choose the files `tree.nwk` and `epi_data.csv`, and click on the **Open** button
(These files were downloaded in the dataflo section).

f. Click on **Continue** in the next prompts.

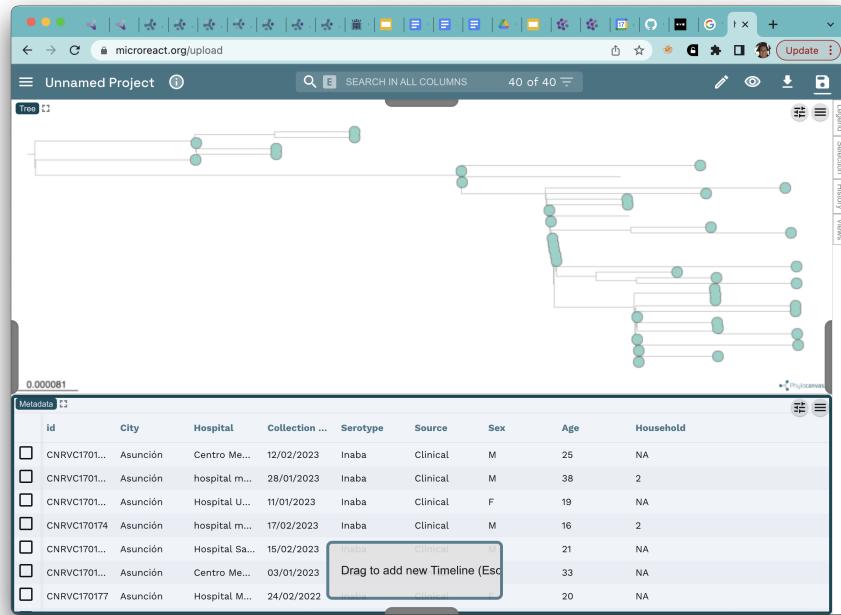
g. Your view will look like this:



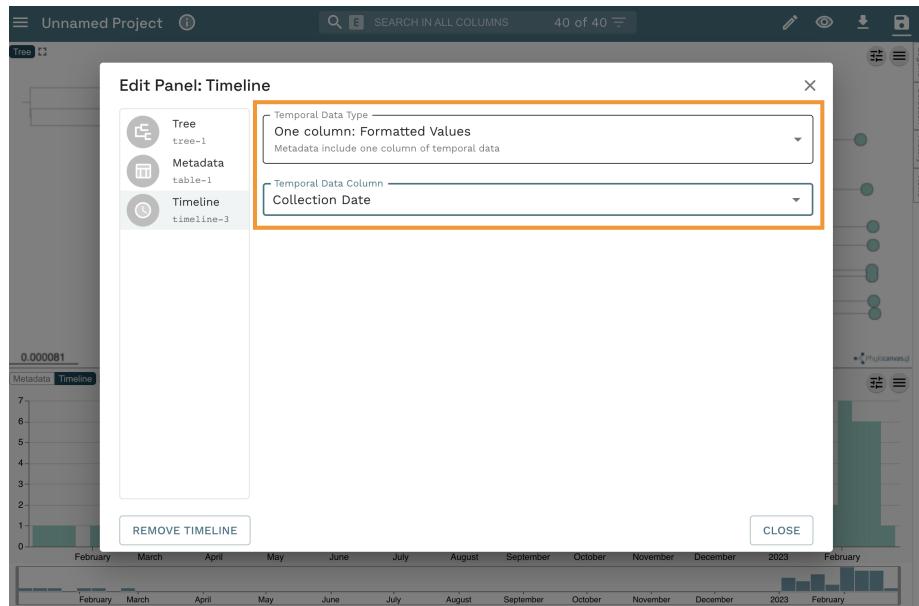
h. You can add a Timeline by clicking on the pencil icon at the top left of the screen, and clicking on **Create New Timeline**



i. Then position the new panel by dragging the pointer to the bottom panel.



- j. Finally, select “One column: Formatted Values” from the Temporal Data Type dropdown list, and “Collection Date” from the Temporal Data Column dropdown list. Click on CLOSE



3. Create a microreact project from a Google Spreadsheet

- a. Open this Google spreadsheet

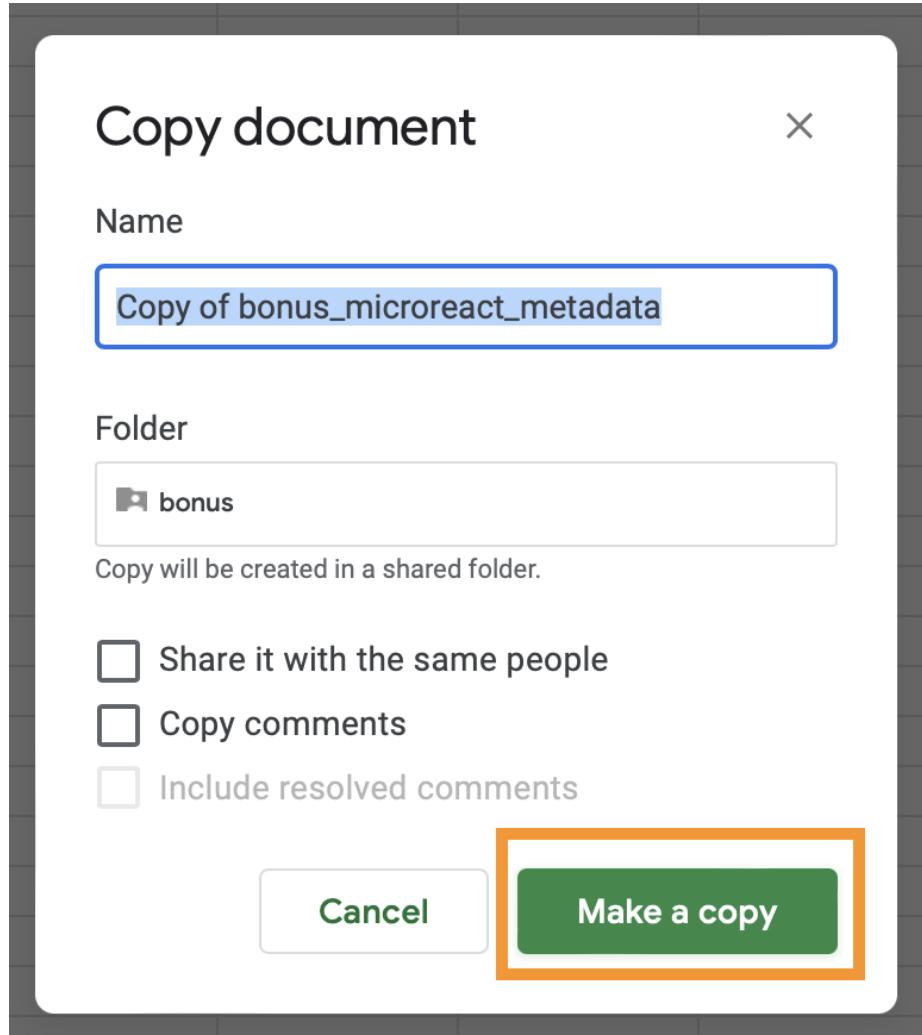
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	filename	NAME	DATE	COUNTRY	ACCURACY	STRAIN	SOURCE	ACCESSION	STUDY ACCESS	SANGER LANE	PATIENT GEND	PATIENT AGE	PATIENT TYPE	SPECIMEN T
2	G18251430.fasta	G18251430	7th October 2012	India							Female	42	O	bl
3	G18251432.fasta	G18251432	22nd January 2012	India							Male	22	I	bl
4	G18251435.fasta	G18251435	14th September	India							Female	49	O	bl
5	G18251437.fasta	G18251437	23rd March 2018	India							Male	19	I	bl
6	G18251438.fasta	G18251438	10th April 2018	India							Male	37	I	bl
7	G18251439.fasta	G18251439	11th April 2018	India							Male	40	O	bl
8	G18251440.fasta	G18251440	16th April 2018	India							Male	51	I	bl
9	G18251443.fasta	G18251443	23rd July 2018	India							Male	33	I	bl
10	G18251444.fasta	G18251444	22nd January 2012	India							Female	39	I	bl
11	G18251445.fasta	G18251445	4th December 2012	India							Male	29	I	bl
12	G18251446.fasta	G18251446	20th January 2012	India							Male	24	I	bl
13	G18251447.fasta	G18251447	22nd January 2012	India							Female	16	I	bl
14	G18251448.fasta	G18251448	10th December 2012	India							Female	25	O	bl
15	G18251450.fasta	G18251450	30th April 2018	India							Female	49	I	bl
16	G18251451.fasta	G18251451	1st July 2012	India							Male	56	I	bl
17	G18251452.fasta	G18251452	6th July 2012	India							Male	68	I	bl
18	G18251453.fasta	G18251453	10th November 2012	India							Female	26	O	bl
19	G18251454.fasta	G18251454	10th September 2012	India							Female	26	O	bl
20	G18251455.fasta	G18251455	29th December 2012	India							Female	33	I	bl
21	G18251459.fasta	G18251459	4th May 2018	India							Female	12	I	bl
22	G18251497.fasta	G18251497	24th March 2017	India							Male	10	I	bl
23	G18251498.fasta	G18251498	22nd August 2012	India							Male	13	O	bl
24	G18251499.fasta	G18251499	29th June 2017	India							Male	35	I	bl
25	G18251492.fasta	G18251492	24th February 2012	India							Female	12	I	bl
26	G18251494.fasta	G18251494	8th January 2012	India							Female	33	I	bl
27	G18251495.fasta	G18251495	11th June 2018	India							Female	17	O	bl
28	G18251496.fasta	G18251496	8th March 2018	India							Female	32	O	bl
29	G18251498.fasta	G18251498	16th February 2012	India							Male	13	I	bl
30	G18254202.fasta	G18254202	26th February 2012	India							Female	25	I	bl

- b. Make a copy of this in your own Google account by selecting **Make a copy** from the **File** menu item. When prompted, click on **Make a copy**. This will open a separate tab where the copy will be available.

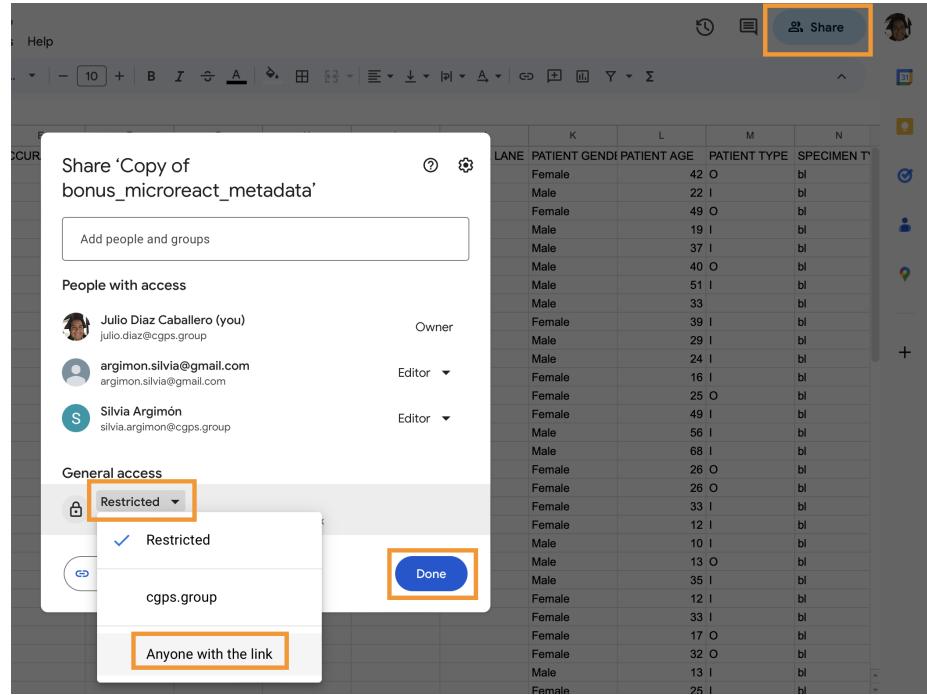
Screenshot of a Google Sheets interface showing the 'bonus_microreact_metadata' spreadsheet. The 'File' menu is open, and the 'Make a copy' option is highlighted with an orange box.

The spreadsheet contains the following data:

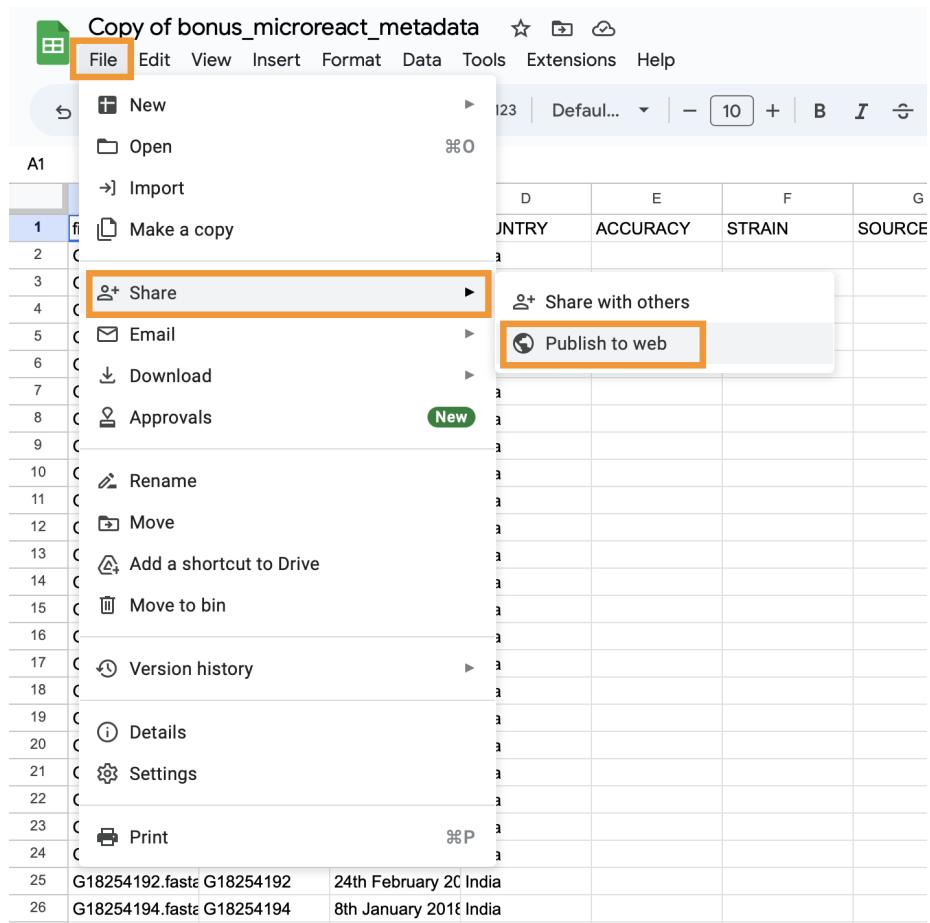
		D	E
		JNTRY	ACCURACY
1	file		
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25	G18254192.fasta	G18254192	24th February 2018 India
26	G18254194.fasta	G18254194	8th January 2018 India
27	G18254195.fasta	G18254195	11th June 2018 India
28	G18254196.fasta	G18254196	8th March 2018 India
29	G18254198.fasta	G18254198	16th February 2018 India
30	G18254202.fasta	G18254202	26th February 2018 India



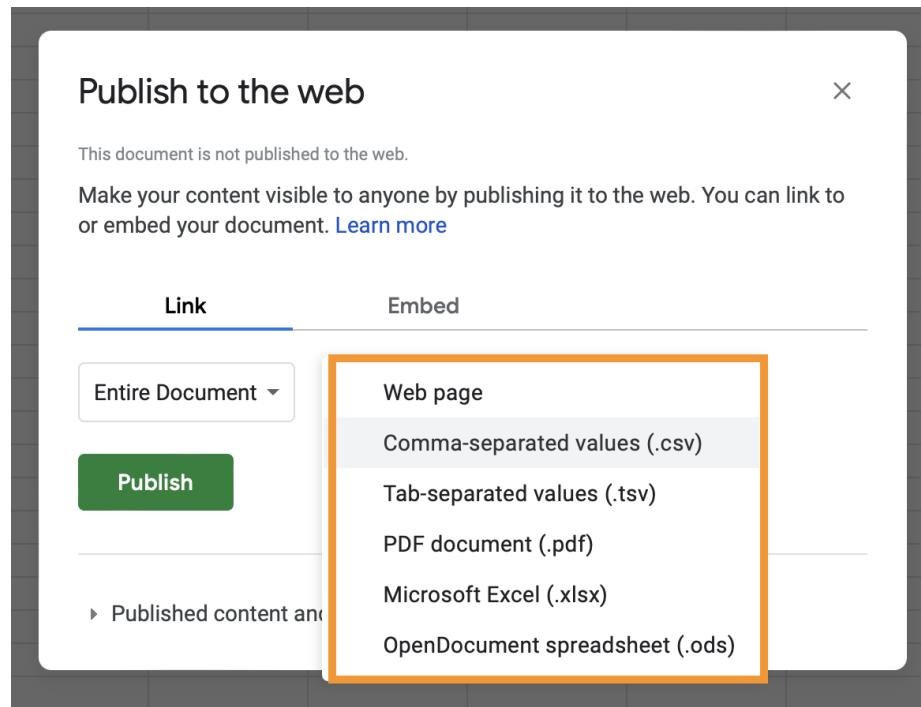
- c. Set access to shareable by clicking on the **Share** button at the top right of the screen. A dialogue screen will popup, here, click on the **Restricted** button, select **Anyone with the link**, and click on **Done**

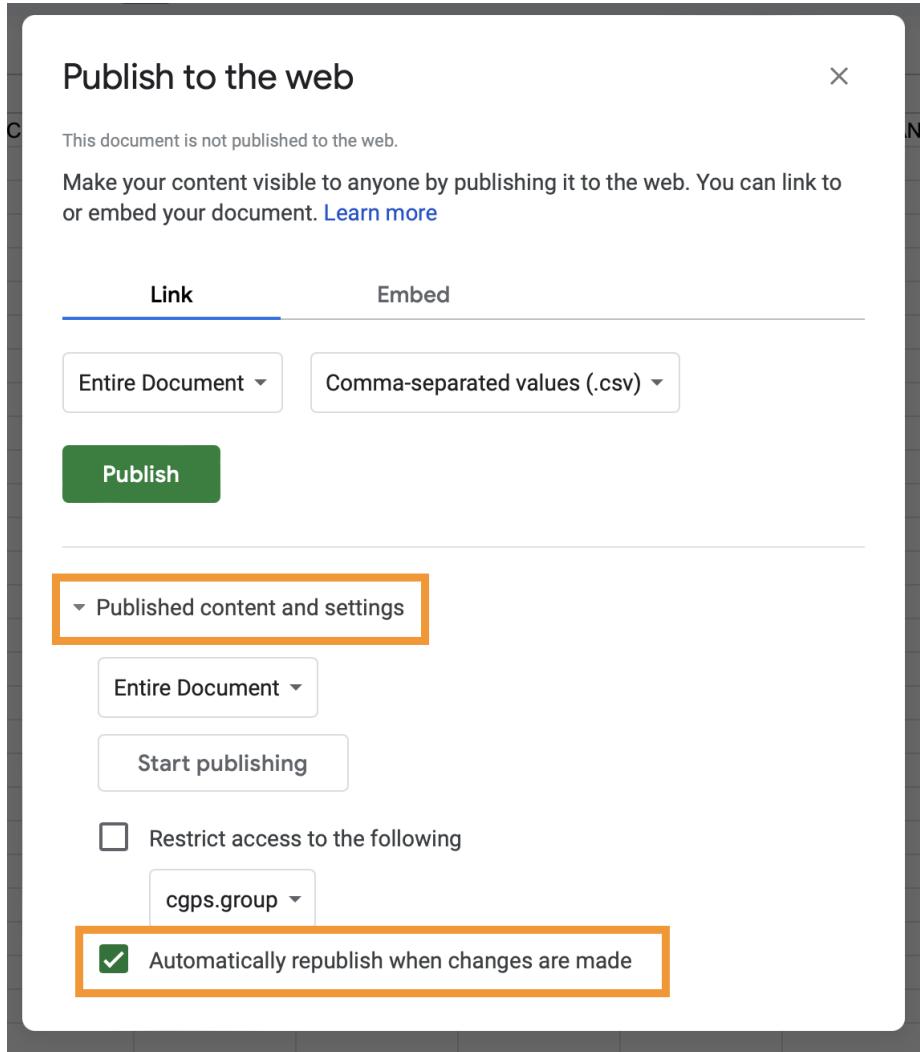


- d. Now to publish the google spreadsheet click on the **File** menu item, select **Share**, and click on **Publish on web**

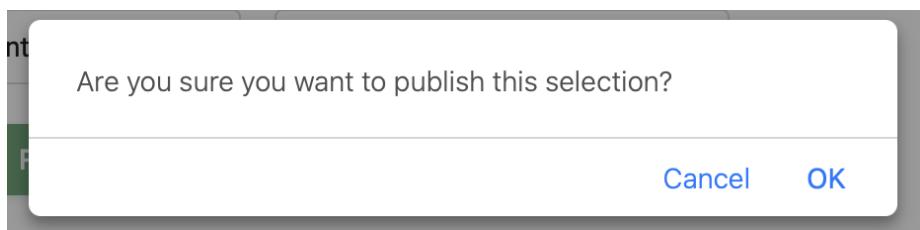


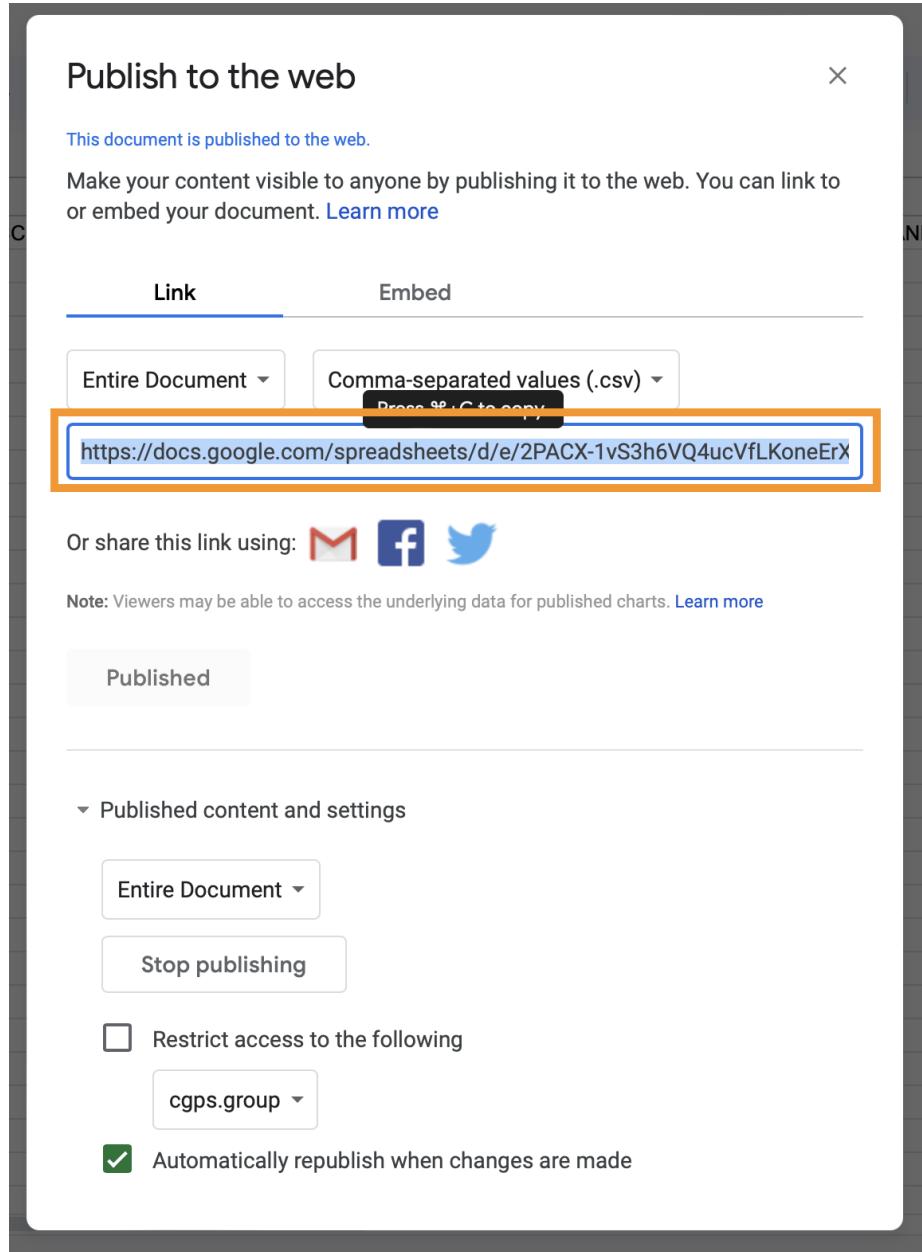
- e. In the popup message click on Web page, and select Comma-separated values (.csv). Also, make sure the Automatically republish when changes are made option is turned on under the Published content and settings section





- f. Confirm your choices in the popup message clicking on OK. This will provide more details about your Google spreadsheet, copy the url.





g. Go to **micrometeor.org**, and select **Upload** from the main menu.

Version 231



Showcase



Upload



Documentation



My Account

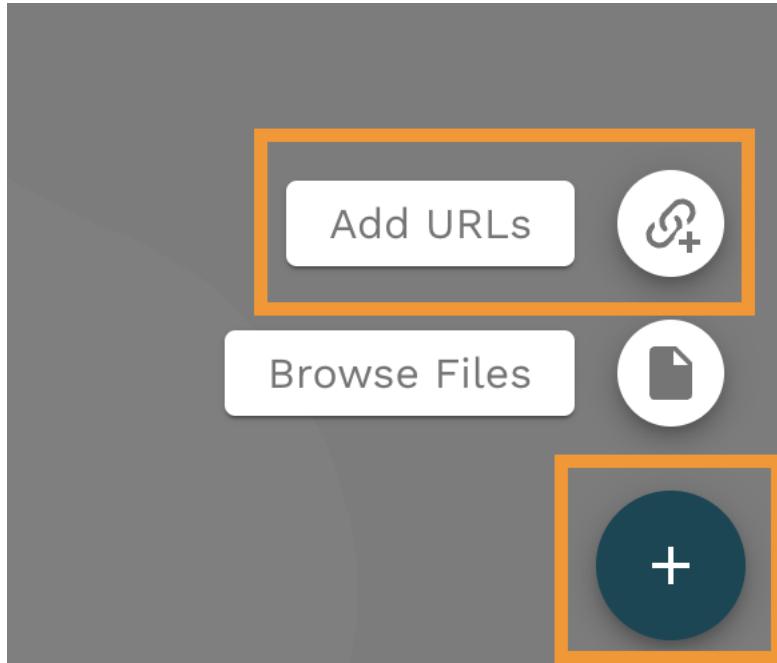


Sign out

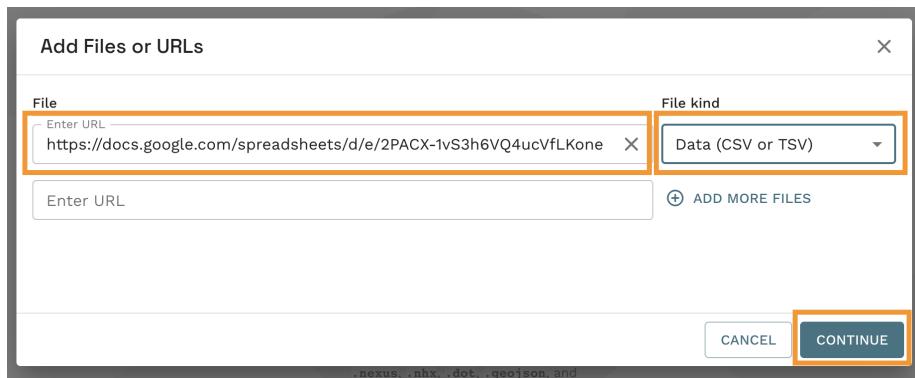


Send Feedback

h. Select the plus icon at the bottom right, and click on Add URLs



- i. Paste the url from step f and select Data (CSV or TSV) under File kind. Finally click on CONTINUE to see your microreact.



- j. You should get a screen like this!

