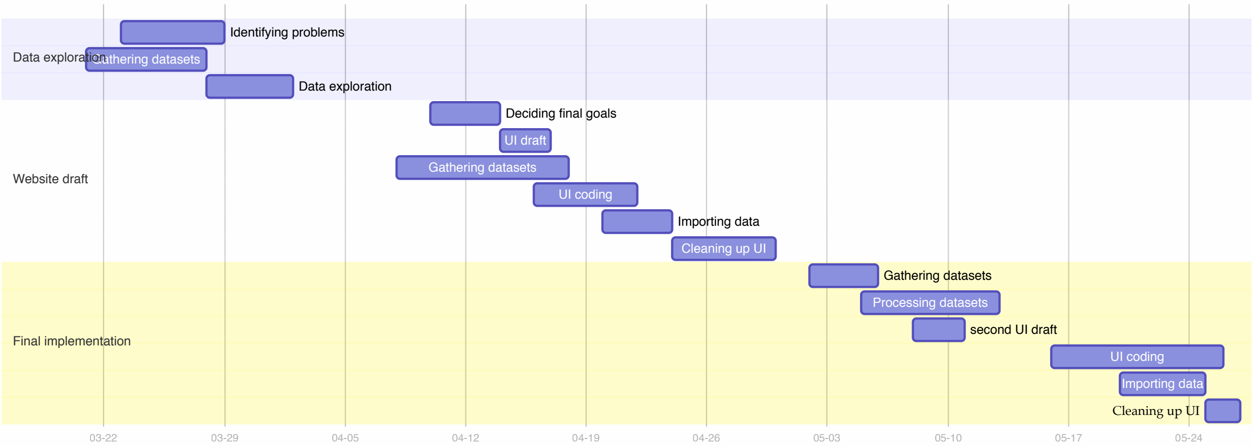
COM480 Process Book

Our Path



Step 1: Determining and exploring data

Early on in the project, we have decided to analyze dataset from Stackoverflow. As a Computer Science student, Stackoverflow has saved our lives countless times.
Unlike many other ways of knowledge sharing like conventional books or lessons, we believe Stackoverflow is unique as it allows instantaneous sharing of knowledge across the world.
We are very interested in this platform and decided to dig deeper and see what information we could find.

There are many available datasets on Stackoverflow, from their official survey, official data dump to Google BigQuery Stack Overflow, and many more, each with their respective collection methods and focuses.

In the end, we have decided to focus on the interactions between users from different countries. Hence, we mainly used data from Google BigQuery Stack Overflow and official survey and integrate them to discover the change in user populations, relationship between users' country and posts (including questions and answers) and dig out the knowledge flow between regions.

We define knowledge flow between countries by the number of questions asked by users from one country, by users of another country.

We performed SQL queries in Google BigQuery Stack Overflow to get the following data:

- Users data:
  - Country of residence
  - ID
  - Name
  - Account created date
  - etc..
- Questions data (2016 - 2019):
  - Question ID
  - Answer count
  - Created time
  - Score
  - Owner's user ID
  - etc..
- Answers data (2016 - 2019):
  - Answer ID
  - Created time
  - Parent (question) id
  - Score
  - Owner's user ID
  - etc..

While most of the raw data we get this way is way too dirty and crude for direct usage, luckily, one of our team members did his final project on Stack Overflow data in the CS 401 Applied Data Analysis 2019. Most of the data has been processed by using more advanced data analysis skills such as matrix factorization, machine learning etc. For this project, our team utilized and amended part of this past work for our use.

In addition, because the data is related to regions of the world, we decided to use the map for better visualization. Therefore, the map dataset is also required and we extract it from GeoJSON.

Step 2: Establish goals

With data and some initial insights, we set up some goals that we want to show on the website:

1. "Knowledge flow" on map

   We plan to visualize knowledge flow as paths between countries with flow directions indicated by animations and flow size indicated by width and color. We will let the user choose which direction he wants to see(knowledge import /export), and how many partner countries to be shown.
2. Sidebar on the right displaying information on the selected country

   We plan to use D3.js combined with SVG to show some statistics about the selected country such as the distribution of user occupations, questions and answers. We would also want to extract the yearly information and show the change in the number of users, questions and answers over different years.

3. Choropleth map showing temporal evolution in terms of the user number and user activities

   We also consider showing the evolution of changes in population and users' activities by showing different Choropleth maps over the years. We will add a slider for user to drag to exact years they want to visualize
4. Dynamic map

   Ideally, we would want to show the label of the country name so that users have a better idea of which country's statistics they are viewing. However, since there are so many country labels to display, the map gets messy quickly. Hence, we would want a way to dynamically show and hide these labels so that we can show the most information to the user without overwhelming them

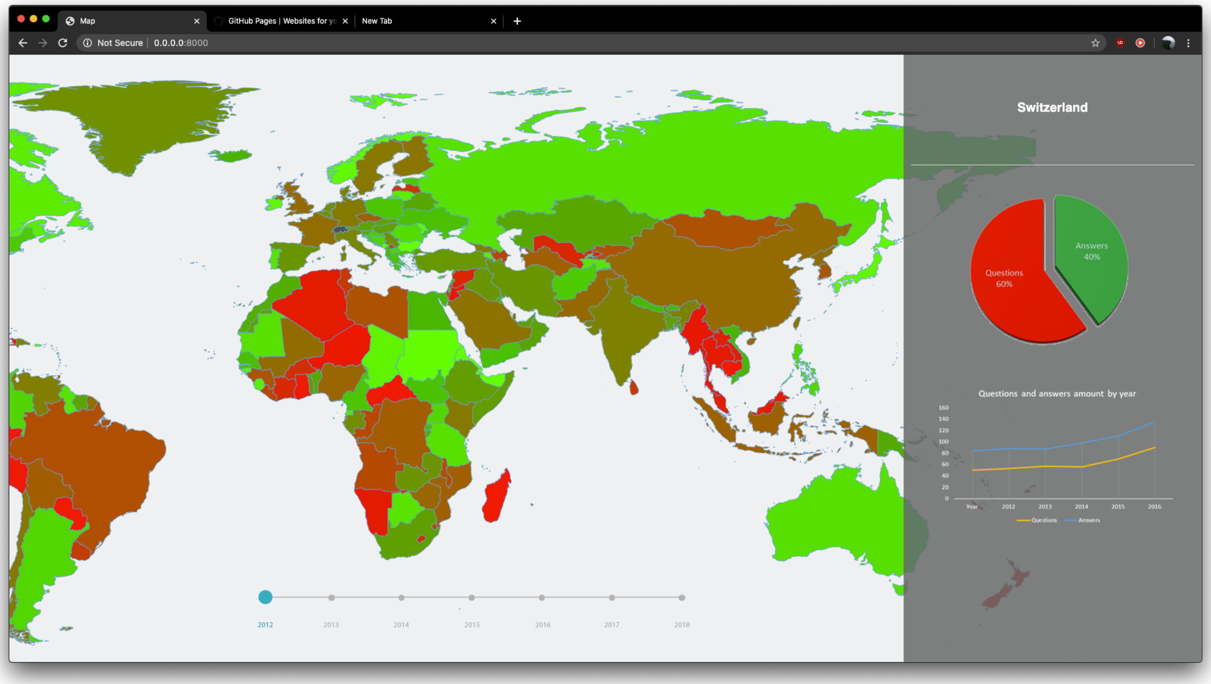
Step 3: *Data preprocessing*

As we have established the goals, we found that data we collected and preprocessed in step 1, cannot fulfill the requirement for all the functions. Therefore, we also did another data collection and processing in order to achieve our end goal. We also offload and store some of the data processing beforehand to reduce the loading time of the page.
Here are some of the extra data we collected and process in this step

1. Since we need to map the country data from different data sources, and display them in the correct location in the map, we set up a dictionary to record the ISO 3166-1 alpha-3 code and name of each country. In the same dictionary, we also pre-computed and stored the total number of users, questions and answers within our dataset
2. Since we collected country data from different sources (GeoJSON, BigQuery etc.), there are quite a number of discrepancies in names of countries (`United States` vs `United States of America`, etc.). We identified these discrepancies and aligned them using their ISO alpha-3 code.
3. Since we also want to show the changes of user population over the years, we split the data from Stackoverflow's official survey by year and counted the number of users in different years for each country.
4. To speed up the query time, questions and answers data between country are processed and the top n flows between each countries are stored to allow for the visualization of knowledge flow on the website.

Step 4: *Building the demo site*

According to the goals we have set in step 2, we determined the general arrangements of different control and visual elements and built a demo to visualize the general UI layout of our website. In the demo, we drew a base map on the website, created a sidebar on the right and a slider at the bottom.



Step 5: Function implementation

With both the UI design and dataset in hand, we began linking the data with the interface and implementing the user controls. We split the tasks into 3 main parts (Flow visualization, Side pane and Choropleth) and distributed them among the three of us and we implemented them individually. In the end, the three parts are combined and alignments are made to create the final website.

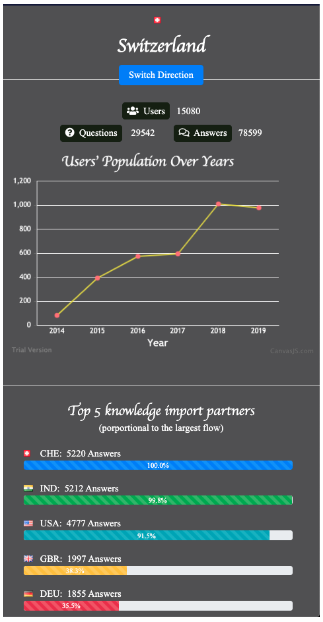
To speed up development and allow for easy maintenance, we split the data, visual elements and the update scripts. We load and store the data on page load use jQuery to perform updates to the individual visual elements when necessary.

In addition, some functions are also made changes and are different from the draft when we implemented. For example, we expected to have a pie chart on the side pane, but many countries actually occupy a very small percentage and thus the pie chart does not seem to have any difference. Therefore, we discard this chart. On the contrary, flow information is added in the sidebar in order to tell the user the top 5 flow in an intuitive way. In addition, the style of the flow is also improved, which is not only a line but also contains a bubble. A color bar is also added to the map to show the number mapped with the color.

Furthermore, we also add an abstract page to introduce our website and the dataset. Team information is also added. This two page aims to help users to learn more about our website and our team member. We add a navigation bar at the top to link all of these pages.

Step 6: Overall arrangement and style adjustments

We have gone through several iterations of the design and made some changes and modifications to the style and arrangement of the site comparing to the first draft. Icons from fontawesome and boostrap such as national flags, users' icon etc, are used to improve the style of our website. In addition, font style, colors, margin and padding are also considered when we arrange our pages. In the following, we will show you part of our UI to help illustrate the changes:



Switzerland

Switch Direction

Users 15080

Questions 29542 Answers 78599

Users' Population Over Years

Top 5 knowledge import partners
(porportional to the largest flow)

CHE: 5220 Answers 100.0%

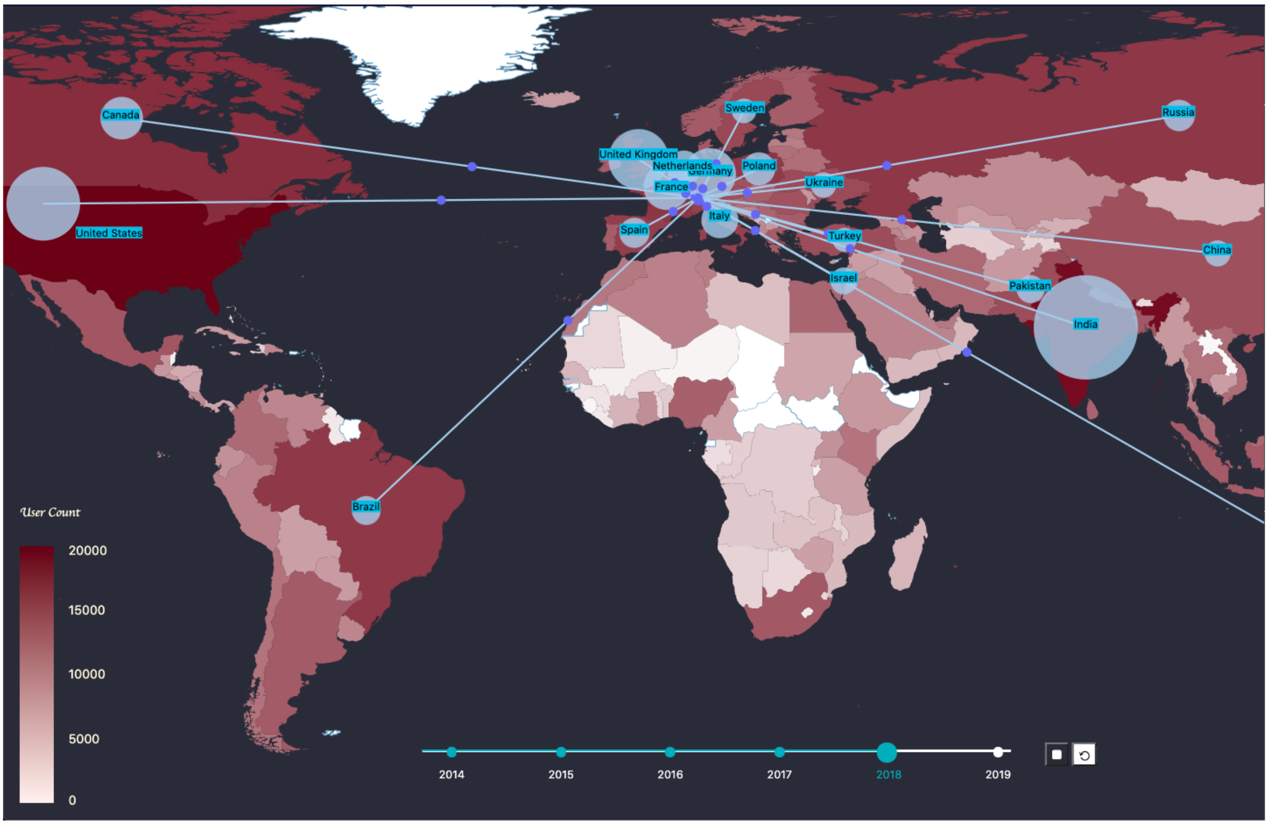
IND: 5212 Answers 99.8%

USA: 4777 Answers 91.5%

GBR: 1997 Answers 38.3%

DEU: 1855 Answers 35.5%

- **Right side pane**
  1. We have removed the pie chart and replaced it with statistics and a list of the top 5 partners. We believe that this information is more relevant to the user compared to the proportion of questions and answers.
  2. We have added country flags to give a better photo representation of each country
  3. We have added a control button to allow users to switch the knowledge import/export direction.
- **Main map**
  1. We have changed the Choropleth map color scheme to make it more visually pleasing
  2. We have added a bubble around each country to indicate the size of the flow
  3. We have added additional control buttons to allow the user to control the temporal evolution



Challenges

Saibo:

- To find some geographical data was more difficult than I thought. For example, the centroid data of countries on the internet was usually not complete. I think most web developers just use some APIs like google map and the centroid data is already embedded. For people who draw maps from scratch like us, this caused some worries. But I found the open data finally on the website of a university's geography department(quite a surprise).
- I need to balance the information volume to display and the size of the website. if the information is too dense, the user experience may not be good as they got an information shock. But the visualization must convey some information we dug from the data. Thus I need to make a choice between the two.

Yueran:

- When I implement the base map, I have struggled for a while on the SVG and d3.js at the beginning. Although I have experience in developing website but I am very new to SVG and d3.js. This library is very flexible but also complicated. First of all, the position arrangement seems to very different from the natural html/css. Html elements can adjust their layer by z-index attribute. However, with SVG, elements which are added later will overlap the elements added previously. It really takes me much time to figure it out.
- Dynamic way to show the labels also takes me a long time to go over it. As the label with the flow will not be the same as the one in natural way, I need to think about all events to make it change or revert to the normal style. Similarly, this is also because of the unfamiliarity of SVG and d3.js. Maybe because the labels are text elements of SVG, I have tried to use jquery to implement it but failed.

Frank:

- I am responsible for developing the Choropleth for the graph. I ran into quite a number of difficulties when trying the implement the progress bar and linking the data.
- Being mostly a Mobile Developer, I am used to the privilege of newer languages such as Kotlin and Swift which has a much more cleaner and direct ways of implementing things. Being such an old language, JavaScript has many quirks and alternative implementing ways that I found challenging. I keep finding old and deprecated ways to implement things which in the end caused me much trouble.
- Moreover, I also found the lack of libraries for control elements challenging. When implementing the bottom progress bar for the Choropleth, I thought there would be libraries available which I could simply modify for my own use. But in the end, I found that there is no suitable library and have to result in a hacky way of implementation

Peer assessment

- Saibo
  - To Yueran: Yueran designed the information holder located on the right side of the webpage and the global layout of the website. As an experienced web developer, she gave guidance during the project. Her work is done on time, code is clean and well documented.

    - To Frank: Frank built the interactive geo map, a dynamical time evolution effect. He is open minded and has creative ideas about the project.
    - Both: Easy to cooperate. Good listeners.
- Frank
    - To Yueran: Yueran is responsible for the right side pane. While I am very new to Web development and made a lot of mistakes at first, you are so kind and patient to guide me during this project. Your code is clean and I have learned a lot just by viewing your code.
    - To Saibo: Saibo is reponsible for visualizing the knowledge flow. You are very creative and are not afraid to share your ideas. You have contributed a lot of brilliant ideas for this project. Thank you for doing a lot of hard work behind-the-scene to provide us with clean and organized data.
- Yueran
    - To Saibo：Saibo is in charge of the part of the flow. He is very creative and very enthusiated in making a flow in a very cool way. He also gives out many fancy ideas to this project and make the whole website more cooler than we thought.
    - To Frank: Frank is responsible for the interactive Choropleth map and time-vary slider. He really has done a great job with fast speed in finishing the task. Additionally, he also proposed many wonderful ways to improve the functions and UI of our website.

Task distribution

- Background map: Yueran
- Interaction on map: Frank
- Knowledge flow: Saibo
- Sidebar: Yueran
- Top 5 flow destinations: Saibo
- Temporal evolution: Frank
- Website layout: Yueran