Visualisation microbial data D3.JS

Stijn De Busser & Jef Ceulemans
Applied Computer Science – Application Development

2024 - 2025

Table of contents

1. Introduction

We are currently doing a foreign internship in Bordeaux, at LaBRI (Laboratoire Bordelais de Recherche en Informatique), a research institute focused on computer science. As part of our project, we are developing an application aimed at supporting microbiological research. The goal of this application is to allow users to view, edit, and analyse microbial data through interactive visual representations.

Users will be able to upload CSV files, which will dynamically load both the raw data and corresponding visualisations. The application will also include data export functionality, making it easy to retrieve and share processed information.

The dataset contains biological data with taxonomic classifications, excluding the Genus and Species levels. Additionally, the application will provide a clear and intuitive overview of all patients and their classifications, presenting aggregated values in a structured and accessible way.

2. Background

The Laboratoire Bordelais de Recherche en Informatique (LaBRI) is a computer science research institute based in Bordeaux. It is part of the Université de Bordeaux and collaborates with the CNRS and Bordeaux INP. LaBRI conducts research in several areas within computer science, including Algorithms, Architecture, Calculations, Artificial intelligence, Software, Networking, Robotics, Security, and Visualisation.

For this project, the collaboration focuses in particular on visualisation. LaBRI supports us by providing expertise on the biological data we are visualising. They have provided guidelines on what the visualisation should look like, and it is now up to us to realise this. In addition, we can turn to them for technical questions and further clarification.

To provide some background on the data; it is derived from DNA sequencing, a method used to detect genetic material. This process retrieves microbial data that is proportional to each sample. In the context of our project, these samples represent patients. Microbial data is organised hierarchically according to taxonomic classifications. Taxonomic levels are used to organise organisms into hierarchical groups, ranging from broad to specific categories. In our case, the hierarchy will not include levels like Genus and Species. Each taxonomic level is associated with an aggregated value, where parent nodes in the hierarchy represent the sum of their child nodes' values. As one moves higher up the taxonomic tree, the aggregated values closer approach a proportion of 1.

3. Objective

The primary objective of this project is to develop an application that enables users to view, modify, and analyze microbial data through interactive plot representations. This application will address the inherent challenges in analyzing microbial data, particularly its complex hierarchical structure. Researchers currently spend significant time manually searching through datasets to identify patterns or correlations, and drawing meaningful conclusions across taxonomic levels is difficult without visual aids.

The application will provide a clear and intuitive overview of all patients and classifications, displaying aggregated values accordingly. It will support multiple taxonomic levels, simplifying complex data structures to facilitate easier understanding of relationships between organisms, and allowing users to explore and compare patterns more efficiently. Key functionalities will include the ability for users to upload CSV files, dynamically loading both the data and corresponding visualisations, and to export data. The application's core components will include data import, data export, and data display in plots/other representations.

The plots and visualisations will consist of:

- **Icicle plot:** A hierarchical, tree-like visualisation that displays taxonomic data as nested rectangles. It helps users easily explore and understand the structure and distribution of classifications.
- **Parallel coordinate plot:** A multi-dimensional chart that allows comparison of values across different taxa. The lines called ‘polylines’ consist of the patient data.
- **Boxplots:** Statistical well-known charts showing the distribution, median, quartiles and outliers of the data. In this project’s specific use case, they will be placed directly underneath the axes of the parallel coordinate plot.
- **Spreadsheet:** Tabular view of the original data in CSV-like format. This standard representation allows researchers to examine the raw data directly. This view reflects the exact format used during export.

4. Business case & stakeholders

Business Case

The development of this microbial data visualization application presents a compelling business case by directly addressing significant inefficiencies and complexities in current biological data analysis. Researchers currently face considerable manual overhead and difficulty in interpreting the hierarchical structure of taxonomic classifications. This often leads to extensive time spent manually searching through datasets and hinders the ability to draw meaningful conclusions or detect trends across various taxonomic levels without effective visual aids.

This project offers a clear solution by providing an intuitive and interactive platform for visualizing microbial data. The value proposition of this application lies in its capacity to transform a time-consuming and difficult analytical process into an efficient and accessible one for researchers. The strategic benefits for the scientific community, and by extension for LaBRI as a research institute, include:

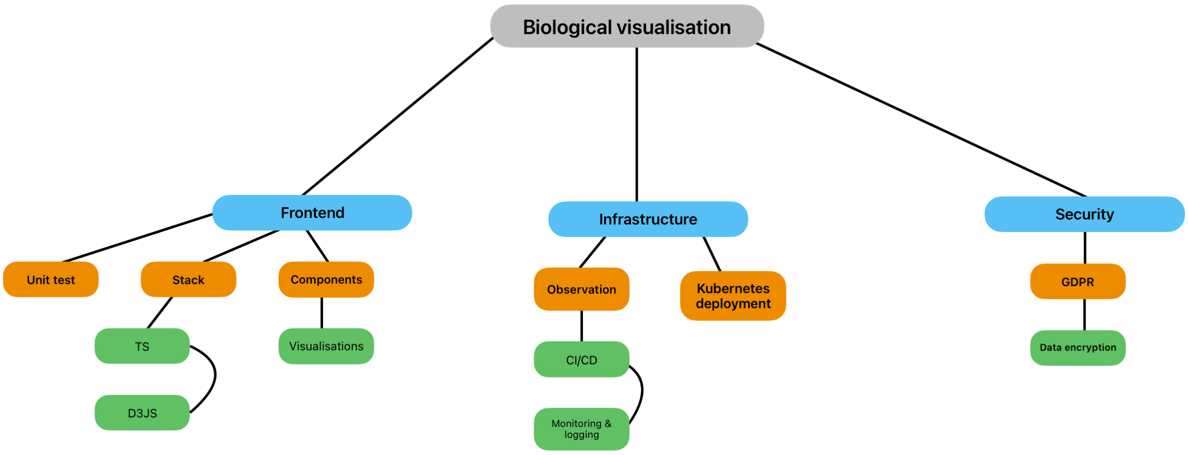
- **Clearer Conclusions:** Visual tools will enable quicker, more effective interpretation of data from CSV files.
- **Improved Efficiency:** The application will reduce the time spent on manual data analysis.
- **Ease of Use:** A user-friendly interface will make complex data accessible to both researchers and non-specialists.

Stakeholders

The key stakeholders involved in this project are:

- **Laboratoire Bordelais de Recherche en Informatique (LaBRI):** As a computer science research institute based in Bordeaux, and part of the Université de Bordeaux, collaborating with CNRS and Bordeaux INP, LaBRI is the primary collaborating research partner. They conduct research in several areas within computer science, with their collaboration for this project focusing particularly on visualisation. LaBRI supports the project by providing expertise on the biological data we are visualising and has supplied guidelines on what the visualisation should look like. They also serve as a resource for technical questions and further clarification.
    - Project mentors (LaBRI):
        - Patricia Thebault
        - Romain Bourqui
        - Antonin Colajanni
- **Jef Ceulemans:** App Developer (Applied Computer Science – Application Development)
- **Stijn De Busser:** App Developer (Applied Computer Science – Application Development)

Product breakdown structure



5. Project timeline

Project Start

- Week 1:
    - Creation of basic graphs and plots.
- Week 2:
    - Meeting about project requirements.
    - Research and project start.
    - Built backend (not needed).
- Week 3:
    - Loading of data.
    - Display of graphs.
    - GUI proposal.
- Week 4:
    - Changed client-side model.
    - Side panel.
    - Parallel coordinate plot.
- Week 5:
    - Display classifications properly.
    - Side panel reflections from taxonomy tree.
    - Patients filtering.
- Week 6:
    - Logarithm scale.
    - Moves app.
    - Import functionality.
    - ~ 5 smaller items.
- Week 7:
    - Hover & selection.
    - Export functionality.
    - Sorting in spreadsheet.
    - ~ 10 other smaller items or bugfixes.
- Week 8:
    - Different views & styling.
    - Export with selection.
    - Instant data reflections.
    - ~ 5 smaller items & bugfixes.
- Week 9:
    - Shift option spreadsheet interface.
    - Separate message when filtered out.
    - Many bugfixes.
- Week 10:
    - Boxplot colors & interaction.
    - Export dialog.
    - Checkboxes keep zero-values.

    - Unselecting & disabling when filtered out.
    - Excel behaviour spreadsheet.
    - Presentation.
    - ~ 5 other smaller items.
    - Bugfixes of course.
- Week 11:
    - Presentation.
    - Realisation document.
    - Reflection document.
    - Project plan.
    - Bugfixes.
- Week 12:
    - Delivering realisation, reflection, and project plan.
    - Code documentation.
- Week 13:
    - Final deliverables.
    - Code documentation.
    - Final presentation.

Project Finish

6. Project Scope

This section explains what our project is about and what the app needs to be able to do. First, we go over the must-have features, the things that absolutely need to be in there for it to work. After that, we list the should-have, could-have, and would-have features, stuff that would make the app better, more useful, or just cooler if we have time to add them.

Must-Have Features

- We need to have a CSV import function to load the file with the taxonomic data.
- There must be an export option where users can choose to export either the selected data or the full current view.
- We need a graph view that shows the data. This should include:
  - an icicle plot
  - parallel coordinates
  - box plot
  - spreadsheet view
- The graphs should show whatever data is currently selected. The way this selection works still needs to be figured out.
- There also needs to be interaction between the different graphs, for example, if you select a quartile in the box plot, it should also be selected in the other views.

Should-Have Features

- A filter and search option to help users find and focus on specific data more easily.

Could-Have Features

- A user account that remembers which CSV files were used, so the user can continue with their last session without reuploading.

Would-Have Features

- More advanced analysis options, like an AI system that suggests interesting patterns or things to look into based on the data.

7. Risk analysis

Here we look at what could possibly go wrong on the front-end and how to handle those problems.

Since we're using D3.js for graphs and data visualisations, one risk is that it might get slow if the datasets are really big. Rendering all that data can cause performance issues, and updating the visuals in real time, like when graphs interact with each other can get complicated.

To fix this, we can try to optimize how D3 renders the data. If it still doesn't work well, we could switch to a different library or add some Angular-specific performance tweaks to speed things up.

Another risk is with the spreadsheet and data editing views. If there's too much data, the page might struggle to display it all at once. To deal with that, we could add things like pagination, scrolling, or even validate the data to keep things running smoothly.

Another thing that could go wrong is that we transform the data incorrectly in the app. Since we need to do some aggregation and the data has a hierarchical structure, it's important that our app understands this properly. To make sure everything is correct, we'll need to have some talks with the biology experts so we interpret the data the right way in the application.

8. Communication

REGULAR UPDATES

Effective communication and reporting will be maintained throughout the project to ensure all stakeholders are regularly informed of progress and key developments. Weekly updates for the internship supervisor are given and delivered at the end of every week. Besides these weekly updates, which are summarised in a document, extra regular meetings are scheduled within the host organization, LaBRI. These meetings are scheduled every week, usually on a Thursday, from the beginning of the internship to the end. In addition to these meetings designed to update mentors and supervisors, there are also regular BKB lab meetings where students or other researchers present and discuss their work.

KEY MEETINGS

Key meetings are structured to facilitate project progression and evaluation. A kick-off meeting was planned for week 2 or 3, where the internship supervisor chaired the meeting and students presented the initial project plan on Friday, March 21st. The first formal meeting took place in week 4, involving a 10-minute presentation of the current project plan, incorporating remarks from the kick-off meeting.

The improved project plan was submitted in week 5, at latest one week after the first meeting, taking into account the feedback received. An internship evaluation by the internship mentor was scheduled before week 7, to be completed on the internship portal in a session between the mentor and internship student.
A second meeting was planned for week 8, featuring a 10-minute presentation covering the realisations until that point, the first version of the realisation document, and remaining work. An intermediate internship evaluation with the internship supervisor also took place in week 8.

Optionally, internship documents (project plan, evidence for realisation, summary, reflection, and additional agreed documents) could be submitted for review no later than May 19th. Another internship evaluation by the internship mentor was scheduled for sometime in May, conducted between the mentor and internship student and completed on the internship portal, with the internship supervisor present if requested. Finally, two project presentations are scheduled: the first for the BKB lab meeting on May 15th, and the second for biologists on a different campus, organized by our mentor, on May 27th.