Enhancing Accessibility and Navigation of Biological Pathway Data: A Comprehensive Sitemap Approach

Personal Background

- Name: Asma Daoud
- GitHub username: semsoum712
- Email: asma.daoud.abbassi@gmail.com
- Phone: +216 25523784
- Links:
  - LinkedIn: www.linkedin.com/in/asma-daoud-3b78b82bb
  - GitHub: https://github.com/semsoum-712
- Location:
  - Country: Tunisia
  - City: Sousse
- Affiliation:
  - Higher Institute of Biotechnology of Monastir
  - Higher School of Health Sciences and Techniques of Tunis
  - Bourguiba Institute of Living Languages
  - Red Crescent

- Background and relevant work experience:
  Unfortunately, I lack experience in work, but I see the Google Summer of Code as a valuable opportunity.

Relevant Skills

- **What are your programming languages of choice and how do they relate to the project?**

  My programming languages of choice are JavaScript, HTML, CSS, XML, and Python. These languages directly align with the requirements of the project, which involve creating a sitemap for a website called Pathway Commons. JavaScript will be instrumental in implementing interactive features and functionality on the sitemap. HTML and CSS will allow me to structure and style the sitemap effectively, ensuring it is visually appealing and user-friendly. XML will be used to manage and organize the data within the sitemap efficiently. XML is crucial for SEO (Search Engine Optimization), as it provides a standardized format for search engines to crawl and index the website's content. Additionally, with my knowledge of Python, I can utilize it for automation tasks to further organize and process the data. With my expertise in these languages, I am well equipped to tackle the challenges of this project and deliver a high-quality solution.

- **Any prior experience with open-source development?**

  Unfortunately, no.

- **What do you want to learn or accomplish this summer?**

  This summer, my primary goal is to contribute significantly to the development of the sitemap project for Pathway Commons as part of Google Summer of Code. I aim to gain a comprehensive understanding of the intricacies involved in mapping the pathways and interactions within the Pathway Commons website. Specifically, I hope to learn advanced techniques in data visualization and representation, as well as explore innovative methods

for organizing and navigating complex biological pathways. Being passionate about both informatics and biology, this opportunity allows me to merge my interests and work on projects that bridge the gap between these fields. By the end of the summer, I aspire to deliver a robust and user-friendly sitemap solution that enhances the accessibility and usability of Pathway Commons for researchers and scientists worldwide. Moreover, I am motivated to earn money through this opportunity to continue funding my studies and pursue further academic endeavors.

- **Any prior exposure to biology or bioinformatics?**

  Absolutely, I have a multifaceted exposure to biology and bioinformatics. Initially, my academic journey began with a mathematical baccalaureate, which provided me with a solid foundation in scientific principles. Subsequently, I pursued studies in nutrition at the Higher School of Health Sciences and Techniques of Tunis, where I delved into subjects like chemistry and biology, expanding my understanding of biological systems.
  My transition to the Higher Institute of Biotechnology of Monastir for personal reasons further reflects my evolving academic interests. In my free time, I avidly engage with bioinformatics through watching educational videos and exploring online resources. This diverse background has equipped me with a unique perspective and a strong desire to contribute to projects in the field of bioinformatics.

Project Proposal

- **Link to the original project idea and potential mentors:**

  Project link: https://github.com/nrnb/GoogleSummerOfCode/issues/218
  Potential mentor: Augustin Luna

- **Project overview, expanded from the original idea:**

  The proposed project aims to enhance the accessibility and utility of Pathway Commons (PC), a vital resource aggregating biological pathway data from numerous curated public databases. By creating a comprehensive sitemap for PC Search pathway and interaction network pages, this endeavor seeks to streamline researchers' exploration of biological networks and interactions, crucial for understanding disease mechanisms and therapeutic targets. The project entails identifying key biological network data items within PC Search, including pathways and interactions, and structuring them into a well-organized sitemap. Additionally, the inclusion of top gene queries and metadata enriches the indexing of PC Search pages by search engines like Google, thus broadening the audience of researchers and facilitating accelerated research discovery. Stretch goals such as incorporating network snapshots and utilizing services like Systems Biology Layout & Rendering Service (SyBLaRS) further enhance the usability of the platform, particularly for visualizing complex biological networks. Overall, this project not only improves the indexing of PC Search pages but also contributes to advancing biological research by facilitating easier access and reuse of curated knowledge, ultimately fostering innovation and discovery in the field.

- **Project details - Multiple sections describing key aspects of the project:**

❖ Our project focuses on enhancing the discoverability of biological network knowledge by creating an XML sitemap file, which is essential for search engine accessibility. To achieve

this, we will employ Python for automation due to the large volume of URLs involved. Utilizing Python's libraries, such as xml.etree.ElementTree, we will programmatically generate the XML structure required for the sitemap. Our process includes data collection, URL processing, error handling, testing, integration with the website, and submission to major search engines. It is worth noting that the generated sitemap will adhere to Google's recommendations, ensuring optimal indexing and accessibility of the biological network content.

❖ In this project, one of our main focuses revolves around enhancing research efficiency by compiling a comprehensive list of interaction IDs identified by HGNC symbols. These interactions play a pivotal role in understanding the complex dynamics within biological systems. Our strategy involves meticulous curation and analysis, utilizing top gene queries sourced from PC Search and pertinent insights from "Hot Genes" extracted from Gene Cards. By amalgamating these resources, we aim to provide researchers with a valuable toolset for exploring gene interactions, thereby facilitating deeper insights into molecular mechanisms and biological pathways. Through this initiative, we aspire to empower scientific inquiry and drive innovation in the field of genetics and molecular biology.

❖ We will also implement an HTML sitemap for pathways and interactions, employing carefully selected criteria to ensure effective organization into meaningful categories. By identifying the most relevant and beneficial categorization methods, we aim to enhance user navigation and facilitate seamless exploration of the Pathway Commons database.

For pathways, we will categorize them based on various biological processes such as cell cycle regulation, signal transduction, metabolic pathways, apoptosis, DNA repair, and transcriptional regulation. For instance, pathways associated with cancer, neurological disorders, cardiovascular diseases, immunological disorders, metabolic disorders, and infectious diseases will be grouped under the respective disease categories. Additionally, pathways can be categorized based on their molecular functions, such as enzyme catalysis, receptor signaling, protein synthesis, and DNA replication.

Similarly, interactions will be categorized into distinct groups based on their types and biological contexts. For example, protein-protein interactions, gene regulatory interactions, metabolic interactions, and signaling pathway interactions will each have their dedicated categories. Furthermore, interactions can be categorized based on cellular localization, regulatory network motifs, and environmental responses.

Our design approach will seamlessly integrate with the existing aesthetic of the Pathway Commons website, ensuring a cohesive and visually harmonious user experience across all aspects of the platform.

- ## Implementation plan and timeline

**** N.B: a sprint is a two working week adapted from the Agile development methodology. ****

- **Sprint 1: Environment Setup and Code Familiarization (May 27 - June 8)**
  - Set up the development environment including necessary tools and dependencies.
  - Familiarize myself with the PC search tool codebase.
  - Study the existing pathways file and understand its structure.

    - Write Python scripts to read and handle the pathways file.
    - Generate an initial XML sitemap for the pathways.

- **Sprint 2: Interaction File and XML Sitemap Integration (June 9 - June 22)**
    - Create the interactions file format for storing interaction data.
    - Develop Python scripts to generate XML sitemap dedicated to interactions.
    - Merge the pathways and interactions XML sitemaps into a single comprehensive sitemap.

- **Sprint 3: Mid-term Evaluation Preparation (June 23 - July 12)**
    - Perform bug fixes and optimizations as necessary.
    - Document my code thoroughly, including inline comments and README files.
    - Conduct a code review with my mentor and address any feedback.
    - Prepare for the mid-term evaluation by organizing my progress and achievements.

- **Sprint 4: Network Identification and Metadata Collection (July 13 - July 31)**
    - Identify the networks to be included in the HTML sitemap.
    - Research and collect metadata sources relevant to the networks.
    - Develop scripts to extract and integrate metadata into the sitemap.

- **Sprint 5: HTML Sitemap Design and Functionality (August 1 - August 16)**
    - Design the layout and structure of the HTML sitemap.
    - Implement the HTML sitemap using appropriate web technologies (HTML, CSS).
    - Utilize JavaScript to make the HTML sitemap fully functional, including user interactions like clicking or hovering.

- **Sprint 6: Finalization and Evaluation (August 17 - August 26)**
    - Perform thorough testing to identify and fix any remaining bugs.
    - Ensure comprehensive documentation covering all aspects of my project.
    - Conduct a final code review with my mentor and address any last-minute feedback.
    - Prepare all necessary materials for the final evaluation, including a summary of achievements and future plans.
    - Submit the final evaluation by the deadline on August 26th.

- **Links to other projects or products that illustrate my ideas:**
    - https://www.xml-sitemaps.com/details-apps.pathwaycommons.org-eedd553ad.html
- **Hurdles and questions that will require more research and planning:**
    - How would we collect the interactions IDs for the sprint 2?
    - What is the most suitable way to categorize the networks in the HTML sitemap?

My Availability

- **Do you have any other time-consuming activities scheduled during the coding period?**
  During the coding period, I don't have any other time-consuming commitments or activities scheduled, as it coincides with my summer holiday. This allows me to dedicate my full attention and effort to the project without any distractions or conflicting priorities.

- **Do you have a full- or part-time job or internship planned for this summer?**
  No.

- **How many hours per week do you have available for GSoC?**

  I plan to dedicate four hours each day, from Monday to Friday totaling 20 hours per week. I believe this structured approach will allow me to maintain focus and make steady progress on the project. Furthermore, I intend to take the weekends as an opportunity to recharge and spend quality time with my family. This time away from work will rejuvenate me, ensuring that I return to the project with renewed energy and enthusiasm for the tasks ahead.

- **Where will you be located during GSoC? Are you traveling during the summer?**

  I will stay in my country “Tunisia” throughout the summer, as I am not planning any travel. I am fortunate to reside in a beautiful country that is renowned as a top destination for tourists. Thus, I will take advantage of the opportunity to explore and appreciate the natural beauty and cultural richness that my homeland offers during this time.