

# **Enhancing digital community engagement: CERN education, communications and outreach group**

**Author:** *Margarida Baptista Lourenço*

**Supervisor:** *Matthew Chalmers*

## **Abstract**

This report outlines the work carried out during my internship at CERN's Education, Communications, and Outreach group to enhance digital community engagement and align the content of the CERN Courier magazine with current trends in High-Energy Physics. The project focused on three main areas: improving internal communication processes, analysing sex disparity among authors of CERN Courier articles, and developing a tool to identify impactful research papers in High-Energy Physics. The outcomes aimed to improve communication workflows, promote diversity in scientific authorship, and facilitate the identification of significant research contributions. The findings and tools developed during this internship provide valuable insights and resources for future digital engagement strategies at CERN.

Geneva, Switzerland

August 29, 2024

## **Acknowledgments**

I would like to express my sincere gratitude to Matthew for opening this Summer Student position and selecting me for it. Throughout this internship, Matthew made himself available whenever he could to help me and guide me through this experience both with work-related matters and beyond. I'm also very thankful to Mark Rayner for mentoring me through the last two tasks discussed in this report. His contagious enthusiasm and willingness to engage in discussions were invaluable. Lastly, Kate Kahle for her cheerful presence and genuine interest in my endeavours.

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Internship Tasks</b>	<b>4</b>
2.1	Updating Processes in the CERN Internal Communication Group . . . . .	4
2.2	Analysis of Gender Disparity Among Authors of CERN Courier Articles . . . .	4
2.3	Impact Engine: Quest to find the most impactful papers . . . . .	5
<b>3</b>	<b>Conclusion and Future Work</b>	<b>7</b>

# 1 Introduction

During this internship, I worked on several projects which aimed to enhance CERN's digital community outreach and align the content of the CERN Courier magazine with current trends in High-Energy Physics. This report details my involvement in three primary areas: updating internal communication processes, analysing gender disparity among authors of CERN Courier articles, and developing a tool to identify impactful research papers in High-Energy Physics.

## 2 Internship Tasks

### 2.1 Updating Processes in the CERN Internal Communication Group

The tasks involved in this process included the following:

1. Identifying and Eliminating Invalid Email Addresses and Duplicates:

This task began by identifying approximately 800 failing email addresses and duplicates. Specific keywords were flagged in the subject lines of bounced emails. The content of each email was then analysed to extract any strings that resembled an email address. Once the failing emails were compiled, they were cross-checked against the four e-groups used for the CERN Bulletin: the dynamic CERN personnel-only group and the static external and internal groups, available in both French and English versions. Invalid email addresses were subsequently removed from these e-groups.

2. Identifying Non-CERN Email Addresses in Subscriber E-groups:

This task involved identifying non-CERN email addresses within the static e-groups. The focus was on detecting duplicated CERN personnel addresses, private email addresses of CERN personnel, addresses of deceased individuals, and external subscribers to CERN. Additionally, email addresses in one language's static group were matched against those in the corresponding group for the other language.

### 2.2 Analysis of Gender Disparity Among Authors of CERN Courier Articles

To analyse sex disparity among the authors of CERN Courier articles, the articles in the Reviews section were scraped and a comprehensive list of titles, dates, and authors was obtained. The extracted data was then exported to a CSV file for further analysis.

The sex of each author was determined using the *Python* package [gender-guesser](#), which classifies names as male or female based on a large dataset. However, a significant number of articles on the CERN Courier website were unsigned, resulting in an "unknown" classification for the author's sex.

This data is visualised in a histogram, as shown in Figure 1, illustrating the disparity between male and female authors.

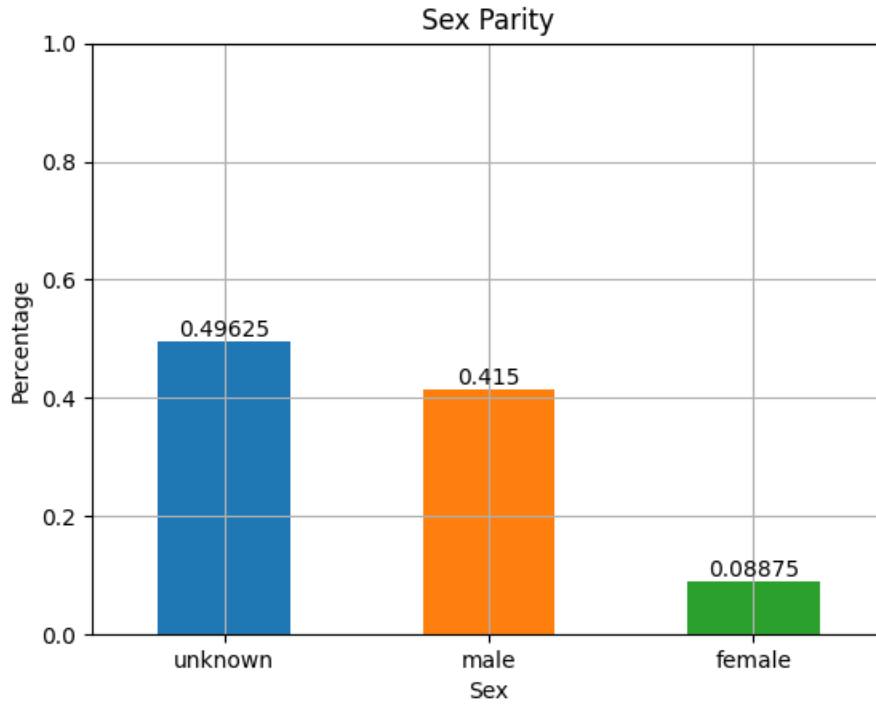


Figure 1: Comparison of authors' sex.

### 2.3 Impact Engine: Quest to find the most impactful papers

**Impact Engine** is a command-line tool developed to help identify the most impactful papers in a specific period of time published in INSPIRE HEP. This tool was designed to fetch citation data, generate CSV files, and create visualisations for easy analysis. It supports various plot types and can handle both individual timeframes and series of timeframes. Additionally, users can fetch and display or save papers that cite a specific paper.

Using this tool it is thus possible to obtain a CSV file with the most cited papers published during a user-specified timeframe. The top 10 papers from this list are then plotted in a bar chart. As an example, figure 2 displays the papers first published between November and December of 2023, ordered by citation count in ascending order. The length of these timeframes is customisable, with the requirement that it must be a divisor of 12. In this case, a two-month timeframe was used.

The command below generates CSV files for every 3-month timeframe between the specified start and end dates and plots the data for the last timeframe in the series (see figure 3).

```
python3 cli.py --start-date "2020-04-01" --end-date "2024-04-01"
--step-back 3 --generate-csvs --timeframe-plot bar
```

The algorithm then analyses the references within each paper, counting how often they appear within each timeframe. This process is repeated for every timeframe in the series, with papers ordered based on the latest timeframe. In the end, it is possible to see the evolution

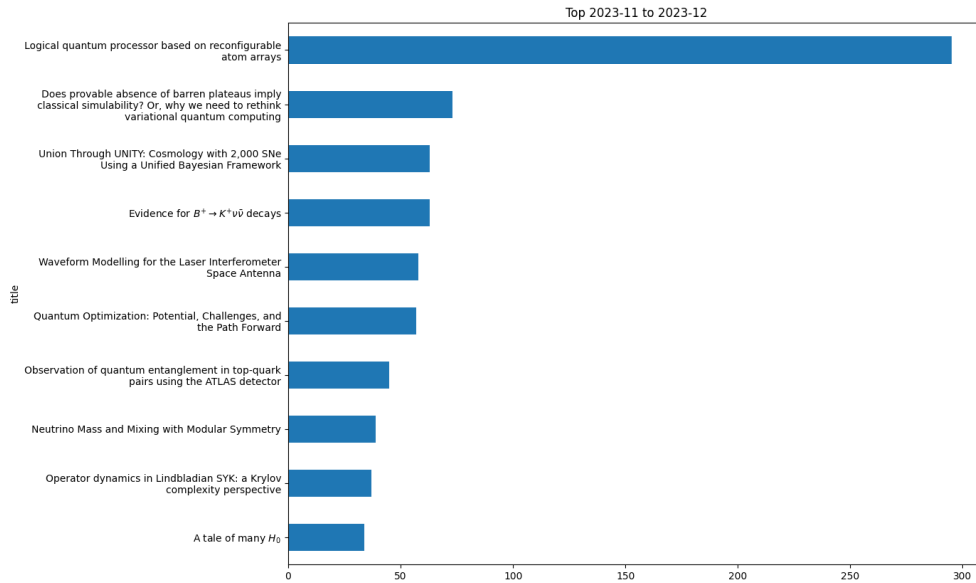


Figure 2: Top-ranking papers first published between November and December of 2023.

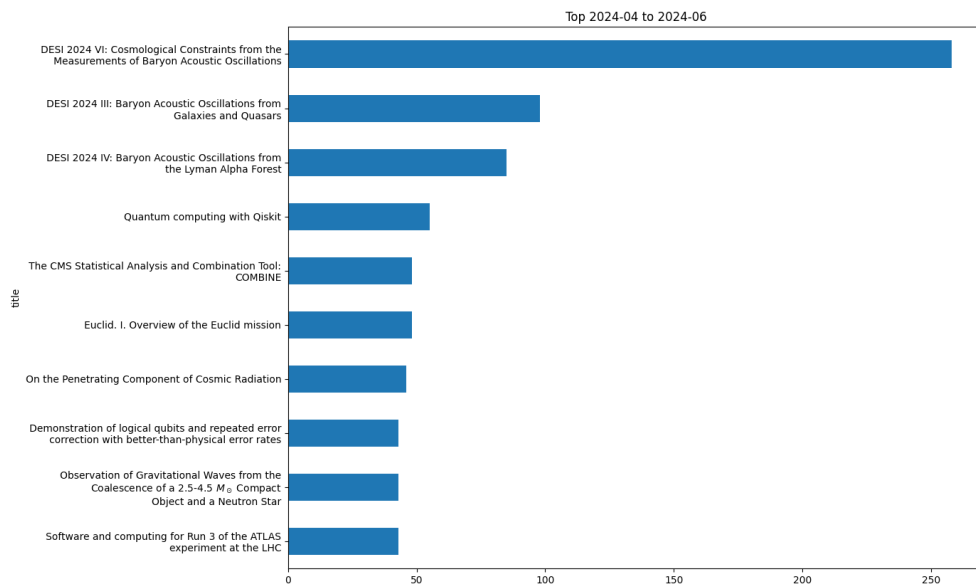


Figure 3: Top-ranking papers first published between April and June of 2024

of the most cited references in that last timeframe through out time, as shown in figure 4. By default, this scatter chart is plotted in the end. To disable it, it's enough to add the argument `--final-plot none` to the previous command.

Having this information, if one is interested in knowing the citing papers that information can be obtain using the command:

```
python3 cli.py --recid "123456" --save-citing-paper
"citing_papers.csv"
```

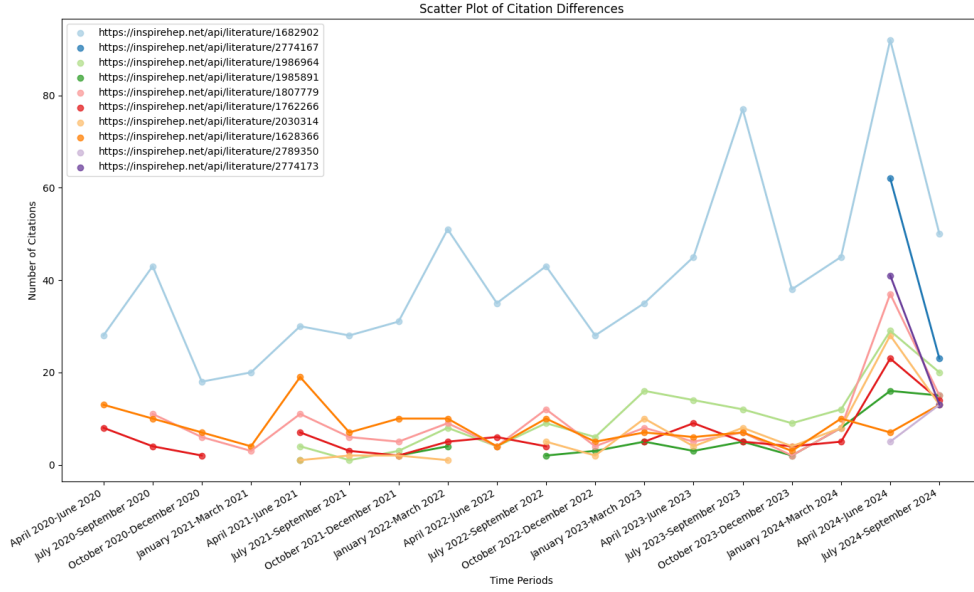


Figure 4: Evolution of the most cited references between July and September of 2024 since April-June of 2020.

which will also save the information to the specified CSV file name.

Additionally, the user can plot data from an existing CSV file by selecting the appropriate plot type — bar plots for top-cited papers within timeframes, and bubble or scatter plots for combined citation data.

### 3 Conclusion and Future Work

There are several areas where the [Impact Engine](#) can be enhanced to improve its performance and robustness. The main ones are the following:

1. **Performance Optimisation:** Currently, the tool takes several seconds to retrieve and process data from INSPIRE HEP. Future work should focus on optimising data retrieval and processing.
2. **Robustness and Testing:** To ensure the tool's reliability, extensive testing is needed. This includes validating the accuracy of the data, and identifying and fixing any bugs.