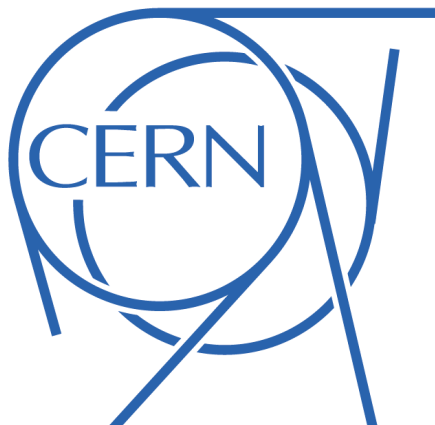




Conseil Européen pour la  
Recherche Nucléaire

European Organization for  
Nuclear Research

Research & Computing Sector  
Scientific Information Service Department

## PROJECT REPORT

# CERNipedia: Wikipedia of CERN

Hridyesh Gupta

Dr. A. P. J. Abdul Kalam Technical University, Lucknow

Supervisor: [Benjamin Bergia](#)

Co-Supervisor: [Jens Vigen](#)

CERN Summer Student Program

CH-1211 Geneve 23, Suisse

September 12<sup>th</sup>, 2025

# Abstract

CERNipedia was developed as an internal knowledge platform for CERN after Wikipedia proved unsuitable due to conflict of interest warnings, notability requirements, and content deletion issues for organization-specific information. The project implemented a local MediaWiki instance with automated content generation capabilities, utilizing the InspireHEP API to populate approximately 6,400 scientist profiles, Greybook for experiments pages and CERN's official website for terminology pages. Custom Python scripts and bot accounts were developed to automate page creation from CSV data, with built-in safeguards against content overwriting and duplicate handling. The resulting platform provides CERN with complete editorial freedom for collaborative knowledge sharing while maintaining structured guidelines for consistent content organization. This projects effectively addresses the gap between CERN's internal knowledge and public wiki publishing limitations, ensuring that CERN's insights are accessible worldwide.

# Acknowledgments

I would like to express my sincere gratitude to all those who contributed to the successful completion of the CERNipedia project during my summer student program at CERN.

First and foremost, I thank my supervisor Benjamin Bergia and co-supervisor Jens Vigen and the RCS-SIS team for their guidance, support, and valuable feedback throughout the project development. Their expertise and mentorship were instrumental in overcoming the technical challenges encountered during implementation.

I acknowledge the InspireHEP team for providing the comprehensive API that enabled automated extraction of author data, forming the backbone of our scientist profiles database. Special appreciation goes to the CERN Scientific Information Service for maintaining the abbreviations database that served as a crucial resource for terminology pages.

I am grateful to the MediaWiki community for their extensive documentation and API resources, which facilitated the local installation and customization of the platform. The open-source nature of MediaWiki made this project feasible and scalable.

Finally, I extend my appreciation to CERN for providing this opportunity as part of the summer student program, offering access to the necessary resources and infrastructure that made CERNipedia a reality.

---

# Contents

|          |                                               |           |
|----------|-----------------------------------------------|-----------|
| <b>1</b> | <b>Project Introduction &amp; Objectives</b>  | <b>3</b>  |
| <b>2</b> | <b>Initial Approach and Challenges</b>        | <b>3</b>  |
| <b>3</b> | <b>Issues Found and Solutions Implemented</b> | <b>4</b>  |
| <b>4</b> | <b>Implementation Workflow</b>                | <b>5</b>  |
| <b>5</b> | <b>Technical Challenges and Solutions</b>     | <b>6</b>  |
| <b>6</b> | <b>Visual Diagrams</b>                        | <b>7</b>  |
| 6.1      | Data Flow Diagram . . . . .                   | 7         |
| 6.2      | Detailed Flow . . . . .                       | 7         |
| <b>7</b> | <b>Results and Achievements</b>               | <b>8</b>  |
| 7.1      | Quantitative Results . . . . .                | 8         |
| 7.2      | Technical Accomplishments . . . . .           | 8         |
|          | <b>Bibliography</b>                           | <b>10</b> |

## 1 Project Introduction & Objectives

The goal of this project is to create a centralized knowledge platform at CERN, where users can search and explore information related to the organization. The platform aims to cover (but is not limited to):

- Information on researchers, authors, and their publications
- Information about CERN experiments, departments, and groups
- Information about various abbreviations and terminologies

It should allow free-form, collaborative contributions internally, without being restricted by public publishing standards like Wikipedia.

## 2 Initial Approach and Challenges

I initially tried to directly publish these articles on Wikipedia. However, this quickly led to:

- **Conflict of interest warnings** (since I was writing about CERN-affiliated people and work)

- 
- **Notability issues**, where the subjects (e.g., specific scientists or internal acronyms) were flagged as "not famous enough"
  - **Page deletion or tagging** due to lack of secondary sources

This made Wikipedia an unreliable place for storing internal but useful CERN-specific information.

### 3 Issues Found and Solutions Implemented

**Problem 1: Public publishing violations** Wikipedia's guidelines prevent hosting of internal or non-notable information, leading to content deletion and conflicts of interest.

**Solution:** A local instance of *MediaWiki* was installed and configured within CERN's environment. This provides:

- Full editorial freedom without external restrictions
- A controlled environment for internal collaboration

**Problem 2: Large volume of authors and researchers** CERN involves thousands of scientists, and maintaining their data manually was infeasible.

**Solution:** Scripts were developed to fetch data directly from the *InspireHEP API*, which:

- Automates retrieval of author details
- Stores extracted data in CSV format for consistency

**Problem 3: Manual page creation infeasibility** With thousands of entries, manual page creation would be error-prone and time-consuming.

**Solution:** An automated bot system was implemented, capable of:

- Reading structured data (CSV) and generating formatted wiki pages
- Performing bulk operations with error handling to avoid overwrites

**Problem 4: Lack of a structured source for abbreviations** Acronyms and terminology form a significant part of CERN's institutional language, but no single structured source was available.

**Solution:** The official CERN abbreviations list was adopted as the primary source, with:

- Manual addition support for missing or experiment-specific terms

- Structured formatting for automated page generation

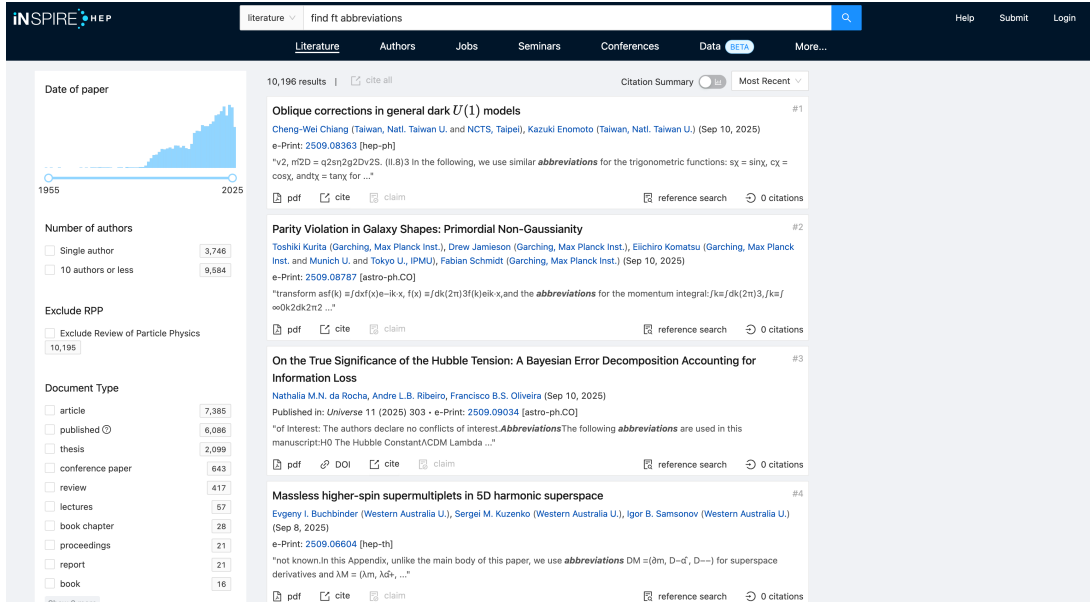


Figure 1: Unstructured Abbreviations present on InspireHEP

## 4 Implementation Workflow

1. **MediaWiki Setup:** The project began with a manual installation of MediaWiki on the server. This included configuring the database, setting up the PHP environment, and fine-tuning the `LocalSettings.php` file to meet project specifications.
2. **Bot Account Creation:** A dedicated bot account was created to perform automated operations. An authentication system was implemented to securely manage bot credentials, and appropriate permissions were assigned to allow bulk actions while maintaining security.
3. **Data Fetching Scripts:** Scripts were developed to fetch data from various APIs, handle errors, and store structured information suitable for automated page creation. Data validation processes were included to ensure accuracy and consistency.
4. **Page Creation Automation:** Using the bot account, automation for bulk page creation was implemented. Scripts were written to generate wiki pages programmatically, populate them with the fetched data, and maintain consistent formatting across all pages.
5. **Duplicate Detection:** A duplicate detection mechanism was developed to prevent redundant content creation. Enhanced error handling and reporting were integrated to monitor and resolve issues during the automation process.

- 
6. **CERN SSO Integration:** Single Sign-On (SSO) integration with CERN's authentication system was implemented to ensure secure access control, allowing only authorized users to edit or view specific sections of the wiki.
  7. **Extension Installation:** Necessary extensions were installed to enhance the wiki's functionality for structured content management.

## 5 Technical Challenges and Solutions

During implementation, several technical issues were encountered. These challenges and their corresponding solutions are summarized below:

**Issue 1: Missing Author Fields** Some author objects retrieved from the InspireHEP API lacked the `preferred_name` field, which caused the data-fetching script to fail.

**Fix:** Implemented a fallback mechanism that uses the `value` field when `preferred_name` is missing. Additionally, programmatic formatting was applied to standardize author names.

**Issue 2: Content Overwriting** When scripts were rerun, existing pages in MediaWiki were unintentionally overwritten, leading to potential data loss.

**Fix:** Introduced an existence check before page creation. The script now verifies whether a page already exists and skips it if found, ensuring previously created content remains intact.

**Issue 3: Duplicate Title Handling** In some cases, multiple CSV rows contained identical titles, which caused conflicts and overwriting of page content.

**Fix:** Implemented a duplicate detection and reporting system. The script flags duplicate entries and generates a list at the end of page creation for manual review.

**Issue 4: Bot Login Failure after SSO Integration** After integrating CERN SSO, the bot was unable to log in using the normal MediaWiki account password, which prevented automated page creation.

**Fix:** Every user who needs to run the script and create pages on CERNipedia must generate a bot password via the `Special:BotPasswords` page in CERNipedia. These credentials are then inserted into the page creation script, enabling the bot to authenticate and perform automated tasks securely.

---

## 6 Visual Diagrams

### 6.1 Data Flow Diagram

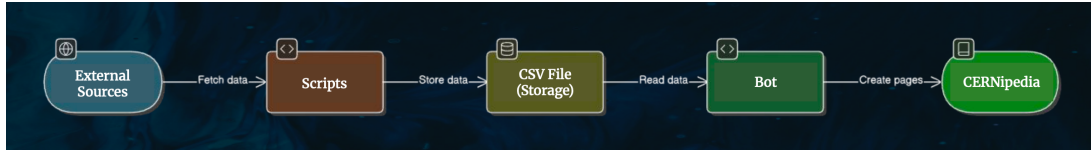


Figure 2: Data flow diagram for CERNipedia implementation

### 6.2 Detailed Flow

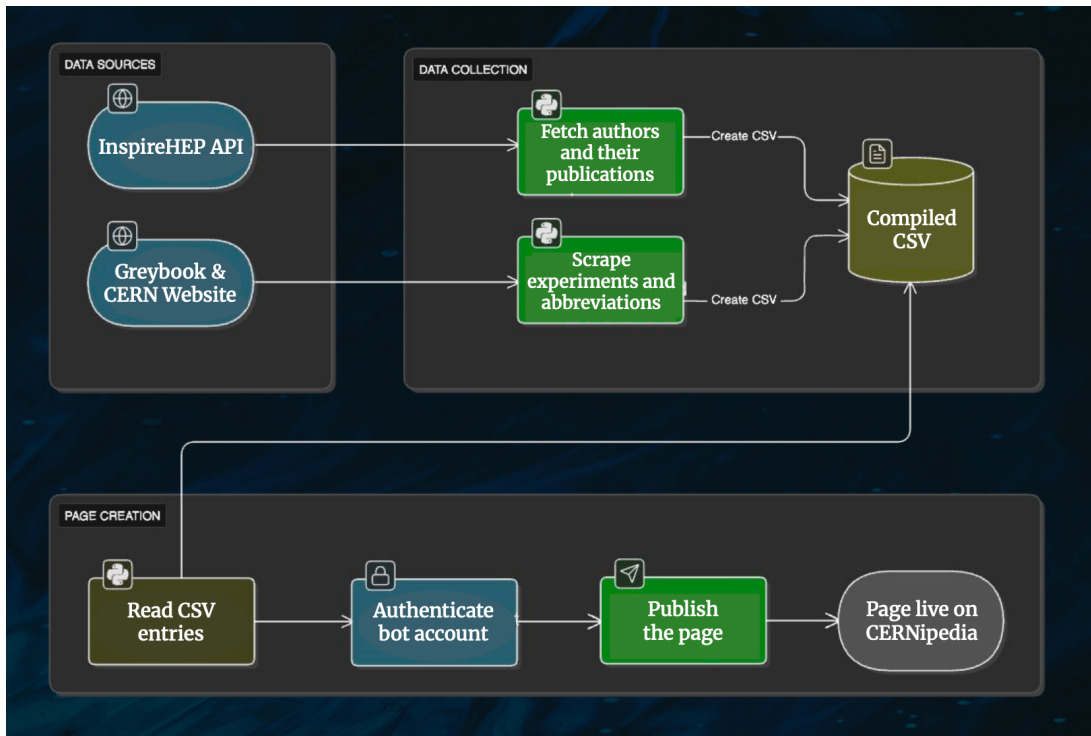


Figure 3: Detailed flow for CERNipedia implementation

---

## 7 Results and Achievements

The CERNipedia project successfully delivered a robust, scalable, and secure internal knowledge platform. Its results and accomplishments can be summarized as follows:

### 7.1 Quantitative Results

- **MediaWiki Local Instance:** A fully functional local installation of MediaWiki was established and configured with CERN-specific settings. This provided the foundation for centralized knowledge management without relying on external public platforms.
- **Author Profiles:** More than 6,400 author entries were imported directly from the InspireHEP API. Each profile included essential metadata and cross-links to InspireHEP, ensuring consistency and ease of reference.
- **Experiments and Terminologies:** Over 1,000 structured pages were created covering experiments, acronyms, and CERN-specific terminologies. This significantly improved accessibility to institutional information that was previously fragmented.
- **Automation Level:** The workflow achieved complete automation for data ingestion, processing, and page creation. Automated duplicate detection and filtering minimized errors and reduced the need for manual intervention.

### 7.2 Technical Accomplishments

- **Data Fetching and Processing Pipeline:** A robust pipeline was developed to fetch data from multiple external sources, validate it, and convert it into structured formats suitable for wiki integration.
- **Automated Page Creation:** The bot-driven page generation system allowed thousands of wiki entries to be created with consistent formatting, error handling, and cross-linking mechanisms.
- **Scalable Architecture:** The system was designed with modular scripts, enabling future expansion. New datasets, experiments, or abbreviations can be integrated without altering the existing infrastructure.
- **CERN Authentication Integration:** By implementing CERN Single Sign-On (SSO), the platform achieved secure access control. This ensured that only authorized CERN members could contribute while keeping the system open for collaborative internal editing.

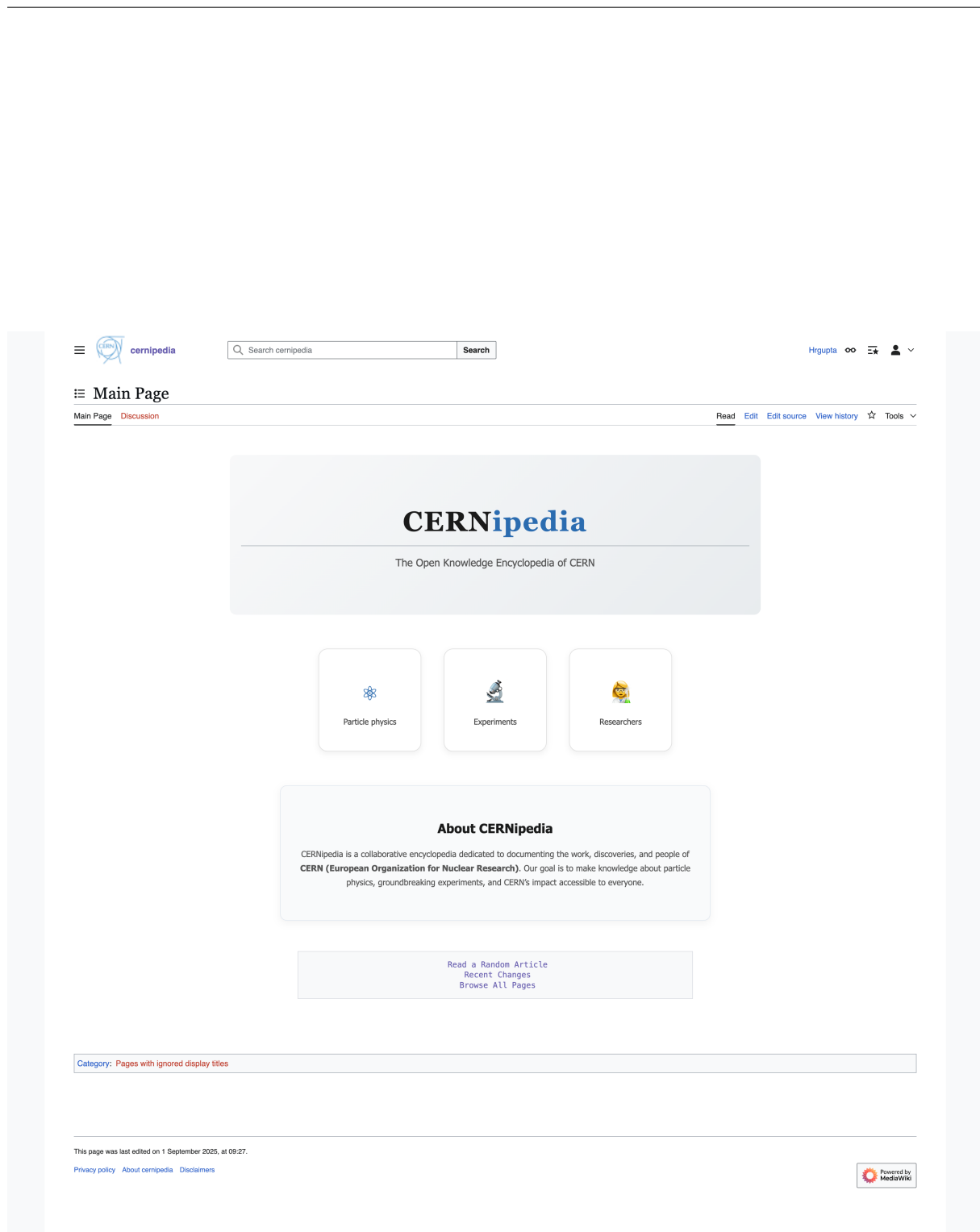


Figure 4: CERNpedia Home Page

---

## Bibliography

- [1] *CERN SSO Guide*. URL: <https://auth.docs.cern.ch/applications/role-based-permissions/>.
- [2] *CERN Website Abbreviations*. URL: <https://scientific-info.cern/archives/acronyms>.
- [3] *MediaWiki API*. URL: [https://www.mediawiki.org/wiki/API:Main\\_page](https://www.mediawiki.org/wiki/API:Main_page).
- [4] *MediaWiki Extension: Cite*. URL: <https://www.mediawiki.org/wiki/Extension:Cite>.
- [5] *MediaWiki Extension: OpenID Connect*. URL: [https://www.mediawiki.org/wiki/Extension:OpenID\\_Connect](https://www.mediawiki.org/wiki/Extension:OpenID_Connect).
- [6] *MediaWiki Extension: PluggableAuth*. URL: <https://www.mediawiki.org/wiki/Extension:PluggableAuth>.
- [7] *MediaWiki Extension: VisualEditor*. URL: <https://www.mediawiki.org/wiki/Extension:VisualEditor>.
- [8] *MediaWiki Installation*. URL: <https://www.mediawiki.org/wiki/Manual:Contents>.
- [9] *MediaWiki Parsoid*. URL: <https://www.mediawiki.org/wiki/Parsoid>.
- [10] *SWAN Documentation*. URL: <https://swan.docs.cern.ch/>.