

Cathy Nguyen, Robin Chataut, Chetraj Pandey

<sup>1</sup>Texas Christian University, Department of Computer Science

## Background and Motivation

- Solar event records from the NOAA Space Weather Prediction Center (SWPC) contain valuable observations of solar phenomena of many forms (NOAA SWPC, n.d.). These events are important for studying solar activity and its effects on space weather.
- However, the archived records are distributed across multiple files and in different formats, which makes them difficult to search, compare, and analyze efficiently (Ahmadzadeh et al., 2020). In addition, prior research has supported data-driven applications in space weather research (Angryk et al., 2020).
- This project focuses on developing a computational framework to organize solar event data into a unified structure.

### Research Question:

How can a computational framework be developed to extract, standardize, and organize solar event records from NOAA SWPC archives to support easier retrieval and large-scale analysis?

## Data Source and Standardization Goal

- The dataset used is obtained from the NOAA Space Weather Prediction Center (SWPC) FTP archive, covering records from 2015-present. The archive contains daily text files documenting events and other solar activity.
- The raw archive data is semi-structured, formatted text entries with irregular spacing, abbreviations, and layout conventions that are designed for reporting rather than analysis (Fig. 1).
- The goal of the framework is to transform raw FTP archive records into a consistent tabular structure that supports easier retrieval and analysis.

```

#Product: 20150629events.txt
#Created: 2015 Jun 29 07:03:57 UT
#Date: 2015 06 29
# Prepared by the U.S. Dept. of Commerce, NOAA, Space Weather Prediction Center
# Please send comments and suggestions to SWPC.Website@noaa.gov
#
# Missing data: ///
# Updated every 5 minutes.
#
# Edited Events for 2015 Jun 29
#
#Event Begin Max End Obs Q Type Loc/Frq Particulars Reg#
7750 0014 /// 0014 CUL C RSP 040-090 III/1
7580 + 0027 0036 0042 G15 S XRA 1-8A C2.2 1.3E-03 2371
7740 + A0112 /// A1444 HOL 3 DSF N05E15 6
7590 0129 0153 0203 G15 S XRA 1-8A C1.2 2.1E-03 2373
7600 0422 /// 0423 LEA C RSP 026-074 III/1
7610 0423 /// 0423 SVI C RSP 027-055 III/1
7620 0523 0525 0528 SVI 3 FLA N08W03 5F 2374
7760 0548 /// 0548 CUL C RSP 020-050 III/1
7630 0732 /// 0759 SVI C RSP 025-079 VI/2
7630 + 0733 0738 0743 G15 S XRA 1-8A 09.6 4.1E-04
7640 + 0733 0738 0743 G15 S XRA 1-8A 09.6 4.1E-04
    
```

Fig 1: Example of a raw daily solar event report from the NOAA SWPC FTP archive. The figure shows the file header, including the product filename, file creation time, report date, missing data notation (///), and update information, followed by the fixed-width 80-column event table with standard headers. Each event row contains the SWPC event number, where a plus sign indicates that multiple reports were received and one was selected as the representative report; the begin, maximum, and end times in UTC; the reporting observatory (Obs); the quality field (Q); the event type (Type), such as X-ray events (XRA), optical flares (FLA); disappearing solar filaments (DSF), and sweep-frequency radio bursts (RSP); the location or frequency field (Loc/Frq); event-specific details in Particulars; and the associated solar region number (Reg#).

## Methodology: Data Extraction and Standardization Pipeline

- A Python-based pipeline was developed to automate the conversion of NOAA SWPC solar event archive files into a uniform tabular dataset.
- Daily FTP text reports are first collected and read as raw input. The pipeline then separates file metadata from event rows, parses the fixed-width event fields, and extracts key attributes.
- Cleaning and normalization are applied to handle missing values, spacing irregularities, and formatting inconsistencies. The processed records are CSV files that support easier filtering, comparison, and future integration into a retrieval tool for querying solar events by user-selected criteria (Fig. 2)

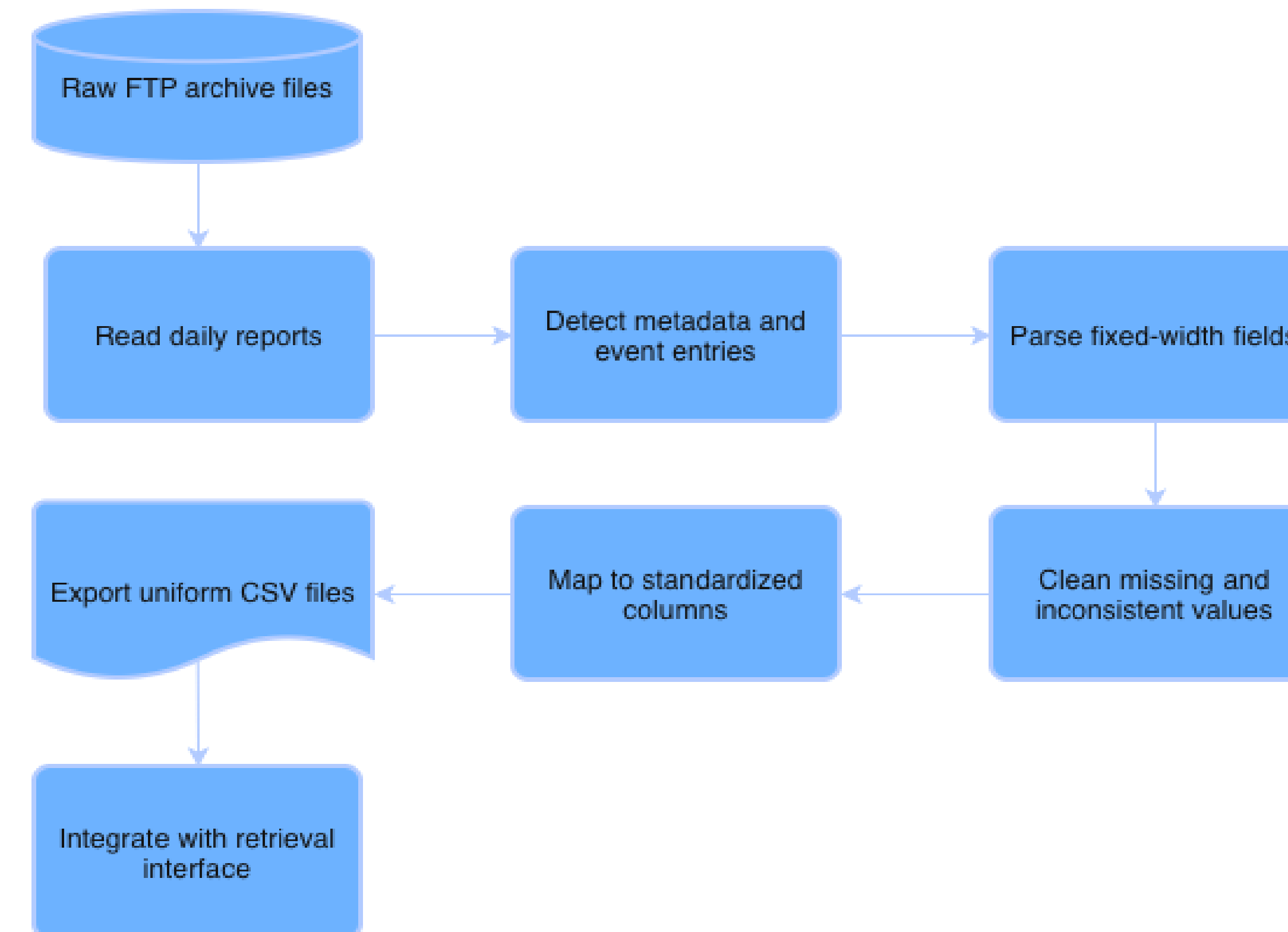


Fig 2: Flowchart of the proposed Python-based data extraction and standardization pipeline. The workflow begins with raw solar event files from the NOAA SWPC FTP archive, followed by reading daily reports, detecting metadata and event entries, parsing fixed-width fields, handling missing and inconsistent values, mapping extracted data to standardized columns, exporting uniform CSV files, and integrating the processed outputs with a retrieval interface.

## Planned Evaluation and Expected Outputs

- Evaluation focuses on whether raw SWPC archive reports can be reliably transformed into a uniform structured dataset. Success is measured by the pipeline's ability to extract key event attributes from daily text reports, normalize inconsistent formatting, and export standardized CSV files (Fig. 3) with consistent columns across records. Preliminary outputs suggest that the workflow can preserve essential event information while improving readability, comparability, and usability (Ahmadzadeh et al., 2020). The expected final outcome is a retrieval tool that allows users to query solar event records more efficiently by criteria such as event type and time range.

Event#	EName	Start	Stop	Peak	GOES Class	Derived Position
1	gev_20150629_0027	2015/06/29 00:27:00	2015/06/29 00:42:00	2015/06/29 00:36:00	C2.2	1-8A (2371)
2	gev_20150629_0129	2015/06/29 01:29:00	2015/06/29 02:03:00	2015/06/29 01:53:00	C1.2	1-8A (2373)
3	gev_20150629_0733	2015/06/29 07:33:00	2015/06/29 07:43:00	2015/06/29 07:38:00	B9.6	1-8A
4	gev_20150629_0959	2015/06/29 09:59:00	2015/06/29 10:19:00	2015/06/29 10:08:00	C1.5	1-8A (2373)
5	gev_20150629_1342	2015/06/29 13:42:00	2015/06/29 13:53:00	2015/06/29 13:48:00	C1.1	1-8A (2376)
6	gev_20150629_1443	2015/06/29 14:43:00	2015/06/29 15:05:00	2015/06/29 14:55:00	C2.3	1-8A (2376)
7	gev_20150629_1758	2015/06/29 17:58:00	2015/06/29 18:14:00	2015/06/29 18:07:00	C2.4	1-8A (2373)
8	gev_20150630_0211	2015/06/30 02:11:00	2015/06/30 02:18:00	2015/06/30 02:15:00	B6.6	1-8A (2373)
9	gev_20150630_0522	2015/06/30 05:22:00	2015/06/30 05:34:00	2015/06/30 05:28:00	B7.5	1-8A (2373)
10	gev_20150630_1222	2015/06/30 12:22:00	2015/06/30 12:59:00	2015/06/30 12:43:00	C1.4	1-8A (2376)
11	gev_20150630_2140	2015/06/30 21:40:00	2015/06/30 22:24:00	2015/06/30 22:11:00	C1.0	1-8A (2376)
12	gev_20150701_0540	2015/07/01 05:40:00	2015/07/01 06:29:00	2015/07/01 06:13:00	B8.8	1-8A (2376)
13	gev_20150701_2242	2015/07/01 22:42:00	2015/07/01 23:01:00	2015/07/01 22:51:00	C1.4	1-8A (2376)
14	gev_20150702_0113	2015/07/02 01:13:00	2015/07/02 02:05:00	2015/07/02 01:43:00	C1.0	1-8A (2376)

Fig 3: Example of standardized solar event records produced by the proposed pipeline and exported into a uniform tabular format. The output organizes extracted event information into consistent fields, including event identifier, event name, start time, stop time, peak time, GOES class, and derived position.

- As shown in Fig. 4 and Fig. 5, the expected outcome of this work extends beyond data standardization to the development of a user-facing retrieval framework. The proposed system is designed to connect parsed and standardized solar event records to a central query layer, allowing users to retrieve events efficiently through command-line, API-based, or dashboard-style interaction using criteria such as event type, date range, peak time, and GOES class.

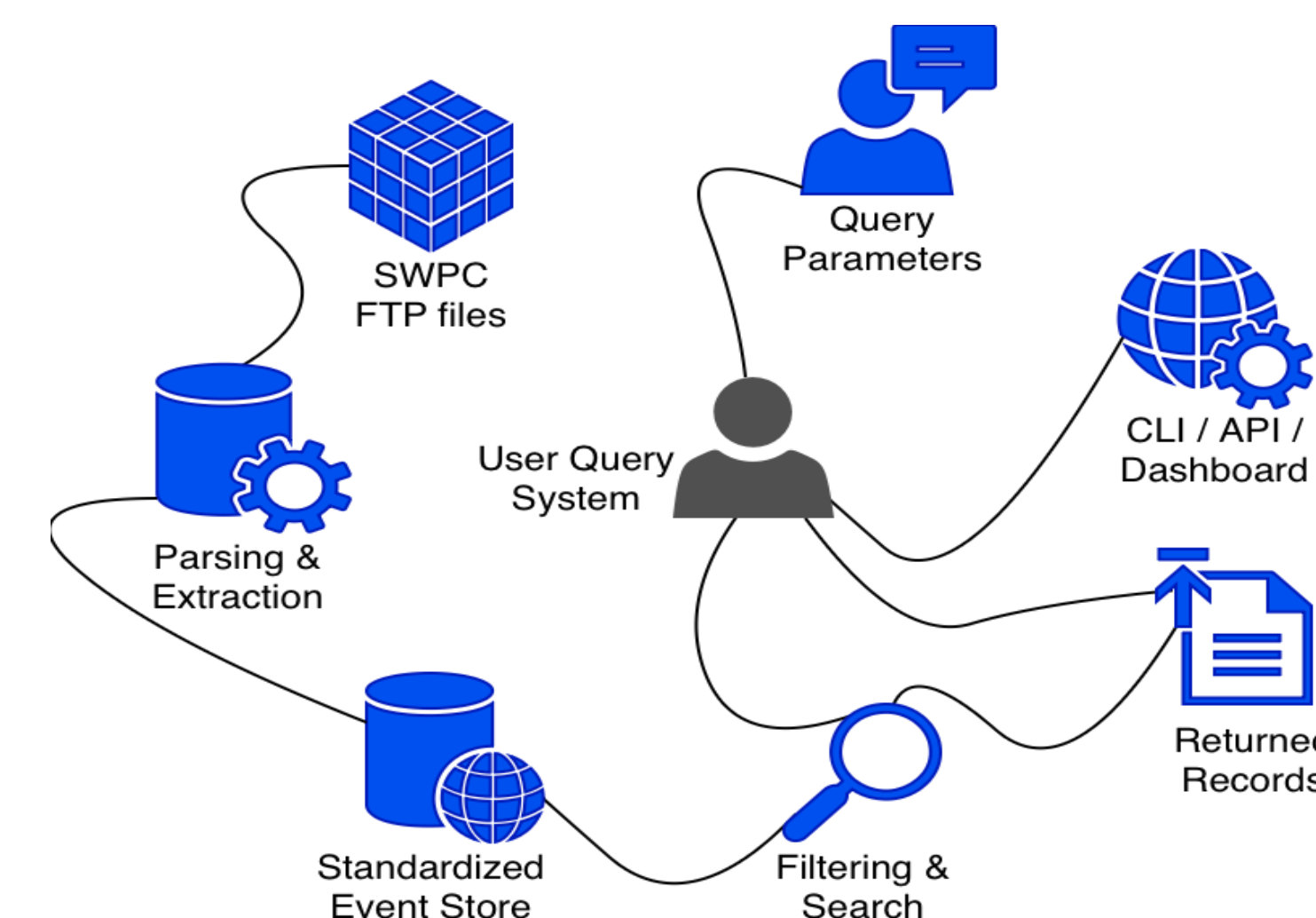


Fig 4: Proposed user-centered retrieval architecture for the solar event framework. Raw SWPC archive reports are parsed and standardized into structured event records, which are then queried through a central user-facing system that supports flexible search parameters and returns matching solar event records.

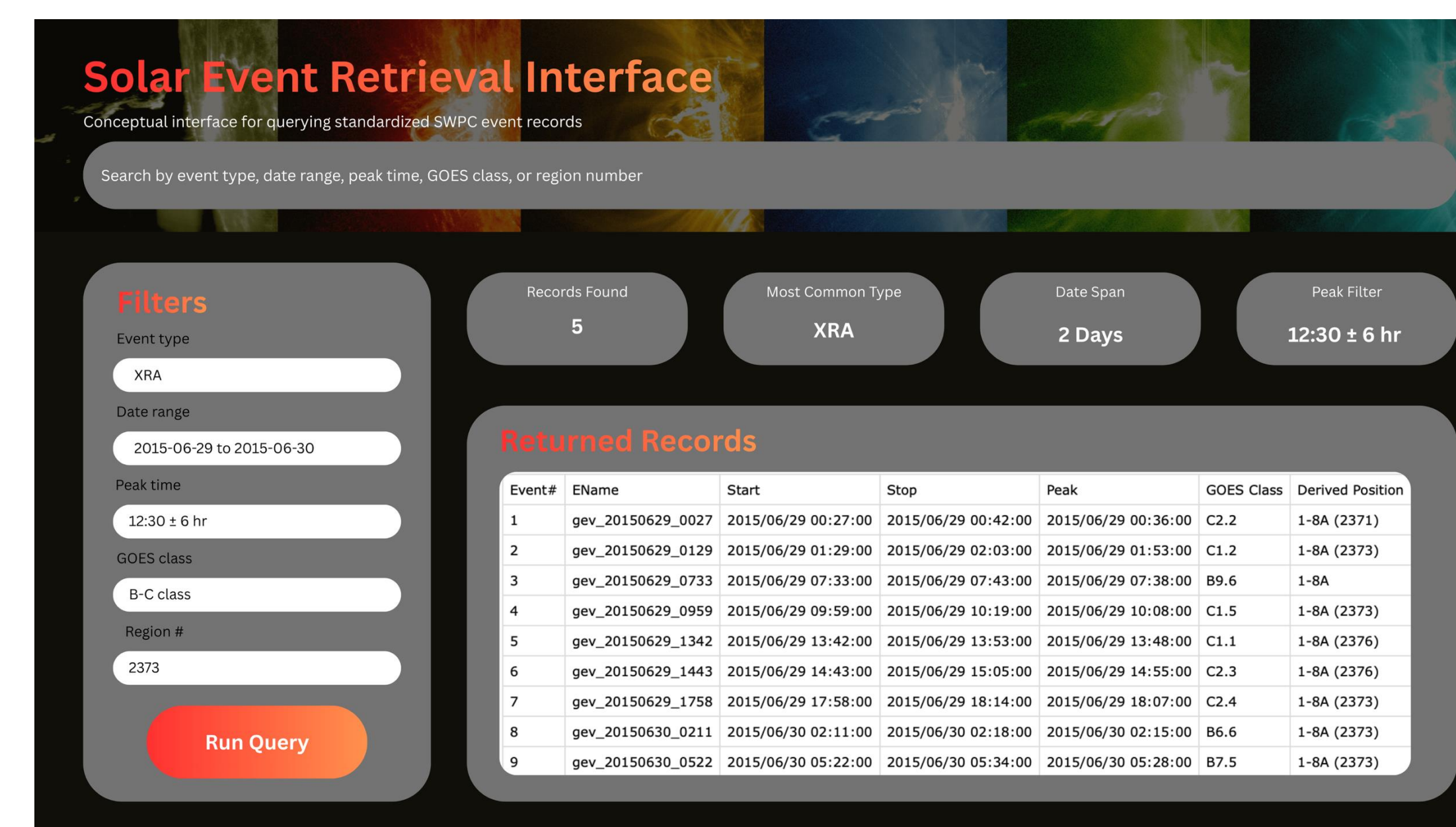


Fig 5: Conceptual interface for the proposed solar event retrieval tool, showing query filters, summary metrics, and structured returned records for user-selected search criteria.

## Remarks and Future Directions

- This work demonstrates the potential to convert raw NOAA SWPC solar event archives into a standardized and analysis-ready format.
- The proposed framework improves data usability by organizing semi-structured event reports into consistent records that are easier to search, compare, and interpret.
- Overall, this work establishes a foundation for faster and more accessible retrieval of archived solar event data.
- Future work includes extending the framework to additional event types, strengthening parsing and normalization across edge cases, and developing a user-facing retrieval tool that supports flexible event-based queries.
- In the longer term, this framework may also support broader integration with searchable databases or lightweight interfaces for more flexible access to archived solar event records.

## References

- Ahmadzadeh, A., Sinha, K., Aydin, B., & Angryk, R. A. (2020). *MVTS-Data Toolkit: A Python package for preprocessing multivariate time series data*. SoftwareX, 12, 100518.
- Angryk, R. A., Martens, P. C., Aydin, B., et al. (2020). *Multivariate time series dataset for space weather data analytics*. Scientific Data, 7, 227.
- National Oceanic and Atmospheric Administration (NOAA). *Space Weather Prediction Center (SWPC) data archive*. <ftp://ftp.swpc.noaa.gov/pub/indices/events>
- Bobra, M. G., Sun, X., Hoeksema, J. T., et al. (2014). *The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs*. Solar Physics, 289, 3549–3578.
- Hurlburt, N., et al. (2010). *HelioPhysics Event Knowledgebase for the Solar Dynamics Observatory*. Solar Physics, 275, 67–78.
- Schou, J., et al. (2011). *Design and calibration of the Helioseismic and Magnetic Imager (HMI)*. Solar Physics, 275, 229–259.

## Acknowledgements

This work is supported by the Department of Computer Science at Texas Christian University, Fort Worth, Texas. The data used in this study is a courtesy of NASA/SDO and the AIA,EVE, and HMI science teams, and the NOAA (NGDC), and the NOAA (SWPC).

## Contact Information

Cathy Nguyen  
Undergraduate Student  
Data Science Major  
Department of Computer Science  
Davis College of Science and Engineering  
Texas Christian University  
Email: [nguyen.nguyen@tcu.edu](mailto:nguyen.nguyen@tcu.edu)