

1. Overview

The **PGN Filter** is a standalone, browser-based tool (single HTML file) for filtering, cleaning, and deduplicating chess games stored in PGN format. It runs entirely offline in your browser — no installation or internet connection is required after downloading the file.

Main goals:

- Extract games of a specific player (or all games)
- Filter by color and date range
- Remove duplicates intelligently using main-line moves
- Produce clean, uncommented output with the best available metadata
- Handle very large PGN files (tested up to approximately 1.4 GB)

2. Features

2.1 File Handling

- Supports one or multiple PGN files at once
- Custom file selection button (“Select file(s)”)
- Shows selected file name(s) and size in MB
- Robust encoding detection: prefers UTF-8, falls back to Windows-1252 (common for older ChessBase / German files)
- Large file support via chunked streaming (8 MB chunks)
- Progress indication in the log while scanning and processing large files
- Output is written as UTF-8 with BOM (correct display of umlauts etc.)

2.2 Player Filter

- Optional — leave empty to process all games
- Case-sensitive matching
- Extra whitespace is ignored
- Intelligent abbreviation support: “Heyl, T” matches “Heyl, Thomas”; single-letter initials are recognized
- Smart suggestions: only names whose last name starts with the typed text are shown (prevents matches like “Heyl” → “Sheyla”)

2.3 Color Filter

- Both colors
- White only
- Black only

2.4 Date Filter (From – To)

Accepted input formats:

- Full date: 18.04.2018 or 2018.04.18 (separators . - / are accepted)
- Year only: 2018
- The tool correctly handles incomplete dates in the source PGN (e.g. 2024.??.??)

2.5 Deduplication

Comparison is based **only on the main-line moves**. The following elements are ignored for duplicate detection:

- Comments {...}

- Variations (...)
- NAGs (\$1, \$2, ...)
- Move evaluations (! ? !! ?? !? ?!)
- Engine evaluations / %eval
- All headers

90 % similarity rule: Two games are considered duplicates if the common move prefix is at least 90 % of the length of the longer game. This is useful when one version is missing the last move(s).

When duplicates are found, the version with the longer move list is preferred and metadata is merged (see below).

2.6 Metadata Merging

When two games are recognized as the same:

- For each header tag the “better” value is kept
- Preference: non-empty value without “?” first, then longer value
- Values containing “?” (e.g. 2024.??.??) are kept if no better alternative exists
- Tag names are normalized to canonical form to prevent duplicates caused by different casing

2.7 Output

- Only clean, uncommented games
- Games sorted from newest to oldest
- UTF-8 with BOM
- Filename: filtered_<PlayerName>.pgn (or filtered_all.pgn)

3. How to Use

1. Open **pgn-filter.html** in a modern browser (Chrome, Firefox, Edge, Safari...).
2. Click **Select file(s)** and choose one or more PGN files.
3. Wait until the player list is ready (progress is shown in the log for large files).
4. (Optional) Type a player name — suggestions appear automatically.
5. (Optional) Choose color and/or date range.
6. Click **Filter and download**.
7. The filtered PGN is downloaded automatically.

4. Technical Details

Aspect	Implementation
Architecture	Single HTML file, pure vanilla JavaScript
File reading	Chunked (8 MB) using File.slice() + FileReader
Memory usage	Only matching games + small buffers are kept
Game separation	Looks for \n\n[or [Event
Move cleaning	Removes comments, variations, NAGs, annotations
Similarity	Tokenized SAN moves, common prefix length
Sorting	Numeric date (YYYYMMDD), newest first
Output encoding	UTF-8 with BOM

5. Limitations & Recommendations

File size	Expected behaviour
-----------	--------------------

≤ 400–600 MB	Comfortable on most machines
≈ 1–1.4 GB	Possible on machines with sufficient RAM (tested)
> 1.5–2 GB	Likely to fail due to browser memory limits
10 GB+	Not realistic in a pure browser tool

- Very large result sets (many thousands of unique games) may make the fuzzy deduplication phase slower. In that case the tool automatically falls back to exact matching only.
- The tool does not modify the original files.
- For multi-gigabyte databases a desktop tool or a Python script is recommended.

6. Tips

- For the best results with player names, type at least the beginning of the last name (case-sensitive).
- Leave the player field empty if you only want to filter by date or remove duplicates from a whole database.
- Partial dates (2024.??.??) are preserved when no complete date is available.
- The log window shows detailed progress and statistics — useful when processing large files.
- You can select multiple files; the player list and filtering work across all of them.

7. Quick Reference — Name Matching Examples

You type	Matches	Does not match
Heyl, T	Heyl, Thomas Heyl, T.	hey!, Thomas (case) Sheyla
Heyl	Heyl, Thomas Heyl, T	Sheyla hey!
Heyl, Thomas	Heyl, Thomas Heyl, T (abbrev.)	Heyl, Tim (different first name)

This handbook describes the current feature set of the browser-based PGN Filter. For multi-gigabyte databases consider a dedicated desktop tool or a Python script.