2025-09-15

fozziejoin: High-Performance String Distance Joins in R

Jon Downs
Goose Data Science and Engineering, LLC

Executive Summary

This proposal requests support for the continued development and publication of `fozziejoin`, a performant R package designed as an alternative to the widely used `fuzzyjoin` (Robinson 2020) for approximate string matching.

`fozziejoin` currently implements 9 of the 10 string distance algorithms supported by `fuzzyjoin`, with consistent improvements in runtime and memory usage across Windows and Linux. Benchmarks show speedups exceeding 100× for certain algorithms, such as Hamming distance.

These gains stem from a Rust-based backend that bypasses the `stringdist` (Mark van der Loo 2014) package, enabling efficient computation tailored to fuzzy dataframe joins.

This proposal aims to:

- Implement the `soundex` algorithm and a `semi` join type, completing core functionality
- Align function signatures and outputs with `fuzzyjoin` for easy migration
- Finalize documentation, including vignettes
- Submit the package to CRAN

If successful, `fozziejoin` will support scalable record linkage in large administrative datasets, with applications in public health, social science, and government analytics. The project embraces open development and welcomes input from the R community.

Signatories

Project Team

The project is led by Jon Downs, who will carry out all tasks described in this proposal. Jon brings extensive experience developing proprietary tools for public sector data science teams. Work will be conducted through a personal LLC. A Contributor Covenant–based Code of Conduct and other improvements will be developed to facilitate future collaboration and community contributions.

Consulted

During proposal preparation, Jon consulted Hadley Wickham, who responded that the approach "sounds like a good plan." The code repository has also been shared with the extendR Discord community, where it received positive feedback and interest.

The Problem

Approximate string matching is essential for data integration and entity resolution in administrative datasets. In R, the `fuzzyjoin` package is widely used for this purpose. Download metrics for `fuzzyjoin` show 9,103 CRAN downloads in August 2025 (Csárdi 2022).

Approximate string matching is facilitated in `fuzzyjoin` via the `stringdist` package. This approach is not tailored to the approximate string matching use case, causing substantial and avoidable memory allocation. Namely, all pairwise string distances are returned as an intermediate result. Most of these will be discarded by a user-defined threshold.

Other packages, such as `zoomerjoin` (Green 2025), offer performant probabilistic methods via Locality Sensitive Hashing (LSH). However, LSH limits algorithm choice and result accuracy, so `zoomerjoin` does not fully replace `fuzzyjoin`.

The proposal

Overview

To address performance limitations in `fuzzyjoin`'s string distance joins, we propose a new R package: `fozziejoin`. To date, `fozziejoin` has implemented 9 of 10 string distance algorithms available in `fuzzyjoin`. Benchmarks based on `fuzzyjoin`'s motivating example show consistent performance gains on Windows and Linux. Speedups exceed 100x for cases like Hamming distance.

These improvements enable scalable approximate joins in R, with applications in public health, social science, and government analytics.

The codebase is designed with CRAN compliance in mind. This grant aims to replicate `fuzzyjoin`'s user-friendly interface, develop documentation and supporting materials, and achieve CRAN publication. Development will span December 2025 to July 2026.

Detail

Minimum Viable Product

The minimum viable product will be publishing the `fozziejoin` package to CRAN. The initial release will include the `stringdist_join` family of functions, comprehensive documentation, an example dataset, a usage vignette, and a testing suite.

To align output structure and function definitions with `fuzzyjoin`, `fozziejoin` will replicate `by` argument behavior, add a case sensitivity option, return a `tibble` when available, and match `fuzzyjoin`'s output column naming conventions. This list is not exhaustive; additional differences will be addressed iteratively, guided by community feedback and practical use.

Architecture

The initial release will include string distance join functions modeled after their `fuzzyjoin` counterparts: `stringdist_join`, `stringdist_inner_join`, `stringdist_left_join`, `stringdist_right_join`, stringdist_full_join, stringdist_semi_join, and stringdist_anti_join. The package will follow a standard rextendr structure.

Assumptions

This project assumes that fozziejoin's performance gains will justify its installation requirements. If not, adoption may be limited. It also assumes that alternatives like zoomerjoin are not direct substitutes for the flexibility and functionality of fuzzyjoin. If this proves false, fozziejoin may offer redundant capabilities.

External dependencies

Installing fozziejoin requires R 4.2 or higher and the base stats package. Future versions will suggest tibble (Müller and Wickham 2023) to support tidy output. Additional packages are recommended for development, benchmarking, and testing. Source installs require cargo, rustc, and the xz decompression utility. Since CRAN provides binaries for Windows and macOS, these requirements mainly affect Linux users and developers.

Project plan

Start-up phase

The project uses the MIT license and is hosted on GitHub. The Contributor Covenant will be adopted and GitHub Discussions will support collaboration and feedback. Monthly updates and quarterly summaries will be provided for ISC reviewers.

Technical delivery

Development will span 8–9 months beginning Nov/Dec 2025, organized into four milestones:

- By February 15, 2026: Add the soundex algorithm and semi join type.
- By April 15, 2026: Align function signatures and output structures with fuzzyjoin.
- By May 15, 2026: Complete documentation, vignettes, and source install guidance.
- By July 31, 2026: Submit the package to CRAN and respond to feedback as needed.

Other aspects

To publicize the work, the proposer will:

- Share announcements and delivery updates on the R Consortium blog at least quarterly
- Promote the project via LinkedIn and the extendR Discord community
- Submit a talk proposal to the next UseR! conference
- Attend ISC meetings as requested to provide updates and receive feedback

Budget & funding plan

The budget covers labor costs for development, documentation, and CRAN prep at $100/hour — a standard rate for specialized R/Rust work. The scope is defined for high-impact delivery.

Milestone	Target Date	Estimated Hours	Funding Estimate
Finalize core functionality	Feb 15, 2026	40 hrs	$4,000
Match `fuzzyjoin` signatures	Apr 15, 2026	25 hrs	$2,500
Documentation & vignettes	May 15, 2026	25 hrs	$2,500
CRAN acceptance	Jul 31, 2026	20 hrs	$2,000
Total		**110 hrs**	**$11,000**

Success

Definition of done

Success is defined as the public release of the minimal viable product on CRAN with all milestones completed.

Measuring success

Progress will be tracked through milestone completion and public deliverables. Monthly updates and quarterly summaries will be shared with ISC reviewers. Community feedback via GitHub Discussions and social media will serve as informal indicators of adoption and interest.

Future work

`fozziejoin` could be extended in several directions:

- Additional join types from `fuzzyjoin`, such as numeric distance, regex, or geographic joins
- Supporting `arrow` and other output modalities
- Utilize `fuzzyjoin` to create a `RecordLinkage` alternative

The project is modular and extensible, enabling future community-centered development.

Csárdi, Gábor. 2022. "Cranlogs: Download Logs from the RStudio CRAN Mirror." https://cran.r-project.org/package=cranlogs.

Green, Beniamino. 2025. "Zoomerjoin: Superlatively Fast Fuzzy Joins." https://CRAN.R-project.org/package=zoomerjoin.

Mark van der Loo. 2014. "Stringdist: Approximate String Matching, Fuzzy Text Search, and String Distance Functions." https://cran.r-project.org/package=stringdist.

Müller, Kirill, and Hadley Wickham. 2023. "Tibble: Simple Data Frames." https://CRAN.R-project.org/package=tibble.

Robinson, David. 2020. "Fuzzyjoin: Join Tables Together on Inexact Matching." https://cran.r-project.org/package=fuzzyjoin.