

## Executive Summary

The FCDS has developed two templates on using the R package [fastLink](#) for probabilistic record linkage (PRL). This 2023 monograph, *A Feasibility Study of the Python Package [spLink](#) for Probabilistic Record Linkage*, showed that [spLink](#) is about 50 times faster and more accurate but more difficult to use than [fastLink](#). The monograph consists of the main text (5 pages), and of a technical showcase (25 pages).

The data cleaning pipeline has four steps. Compared with the [fastLink](#) templates, these are the [spLink](#) results for the four steps:

- 1) pre-processing: [spLink](#) is more difficult to use by having no template.
- 2) blocking: [spLink](#) enables faster and more accurate PRL by offering OR (disjunctive) blocking.
- 3) PRL: [spLink](#) is much faster by using the SQL back-end database [DuckDB](#).
- 4) post-processing: [spLink](#) is more difficult to use by having no template.

The following are three feasible uses for [spLink](#) at the FCDS, in order of suggested priority:

- Replace [fastLink](#) for linkage data requests. It can save about 33% or 1 week of 3 each time.
- Reduce the amount of manual review for Match\*Pro de-duplication.
- Replace real-time deterministic record linkages with real-time PRL.

For the 2024 monograph, the FCDS recommends to continue improving PRL. Specifically, these are the recommendations:

- Improve the accuracy and speed of [fastLink](#) by using the expected new release.
- Improve the user-friendliness of [fastLink](#) by updating the FCDS templates from R Markdown to Quarto Markdown. It will enable new useful features such as multi-format support and code annotation.
- Improve the user-friendliness of [fastLink](#) and [spLink](#) by reducing the need for data cleaning. Decide if and how “standardized” variables in the FCDS database can be used or created for the linkage data requests.
- Work with Abraham Flaxman to improve and make public the artificial test data. The test data need to have more noise to test how accurate PRL is in terms of wrong matches (FP).