

International Journal of Population Data Science

Journal Website: www.ijpds.org



Splink: recent developments and realtime linkage

Andrew Bond¹

¹Ministry of Justice, London, United Kingdom

Splink is the award-winning open-source data-linkage software developed by the UK's Ministry of Justice (MoJ). It implements the probabilistic Fellegi-Sunter model in SQL, allowing users to employ their execution backend of choice, whether that be with DuckDB, Spark, or another engine of preference. Splink's focus on performance, accuracy, scalability, and rich feature set have enabled it to accrue a wide user base across the public sector, academia, and beyond. Splink is in continuous development, motivated by the ever-developing needs for record linkage within the MoJ, as well as the challenges faced by our wide range of users across the globe. We will discuss some of the latest developments within Splink, as well as a look at features that are on the near horizon. Recent

focus includes work to allow scaling of linkages to ever-larger datasets, and to provide much greater flexibility for users to control the individual portions of their data-linking workflow. A particular focus has been our work on near-realtime record linkage. We have developed a system for operational record-linkage that keeps in sync with continuous updates to source data using incremental linkage, which brings unique challenges for performance and scaling compared to batch-processed linkage. We will discuss some of the issues around this, how it informed the design of our system in order to make its operation a reality, and how we have then folded these ideas back into Splink to make it more straightforward for other users to do likewise.

