Editing Source Code Programmatically and Tokenizing Blaise Syntax via Regular Expressions

By

Jason Ostergren

The Health and Retirement Study (HRS) is a national longitudinal survey on a variety of topics associated with aging and retirement. The HRS utilizes a Blaise CAI instrument for biennial interviews of one to two hours in length given to around twenty thousand participants. The source code for the HRS instrument is lengthy (about nine megabytes of text). Every two years, HRS makes extensive changes to the instrument. Some of these changes are repetitive, such as formatting changes to field properties. Such changes may follow a particular pattern, which makes any individual change simple, but greatly multiplies the chance of mistakes due to sheer volume. Particularly due to the size of the HRS code, these changes often consume large quantities of programmer time, so HRS has made a number of attempts over the years to streamline or automate the process of making certain kinds of source code changes where possible, including one attempt which was the subject of a previous IBUC paper in 2007.

This paper will explore a new HRS effort to automate source code changes. The main drawback of the previous method was that it became rather complicated to run and maintain due to lengthy code and dependencies on processes involving external databases and programs. Much of the complication was a result of the tokenizing process, so HRS developed regular expressions to tokenize the Blaise source code more cleanly. This also had the result of vastly speeding up the process of tokenizing, so that the entire operation could be run in less than a minute, which was a great aid to the iterative development of the solution. Subsequently, HRS has constructed a program around this simple method of tokenization which makes it possible to perform operations such as formatting the code, removing comments, and adding new languages programmatically.