Websites hold their information in the structure of an HTML document, and parsing that information is generally accomplished through the use of external libraries to scan the document for CSS selectors and specific HTML tags. Just as AWK handles the parsing of standard text documents, the motivation for HAWK—HTML is All We Know—was to build a language to allow the user to easily handle parsing websites without the need to import any external libraries.

HAWK is a programming language designed for web scraping. Our solution to this was inspired by AWK’s powerful text parsing capabilities and we use the same (pattern, action) model as the basis for our HAWK programs. HAWK matches against both regular expressions and CSS selectors to select elements of a website document. The HAWK compiler compiles code to Java code, which is compiled to Java Bytecode, and then run on a Java Virtual Machine.