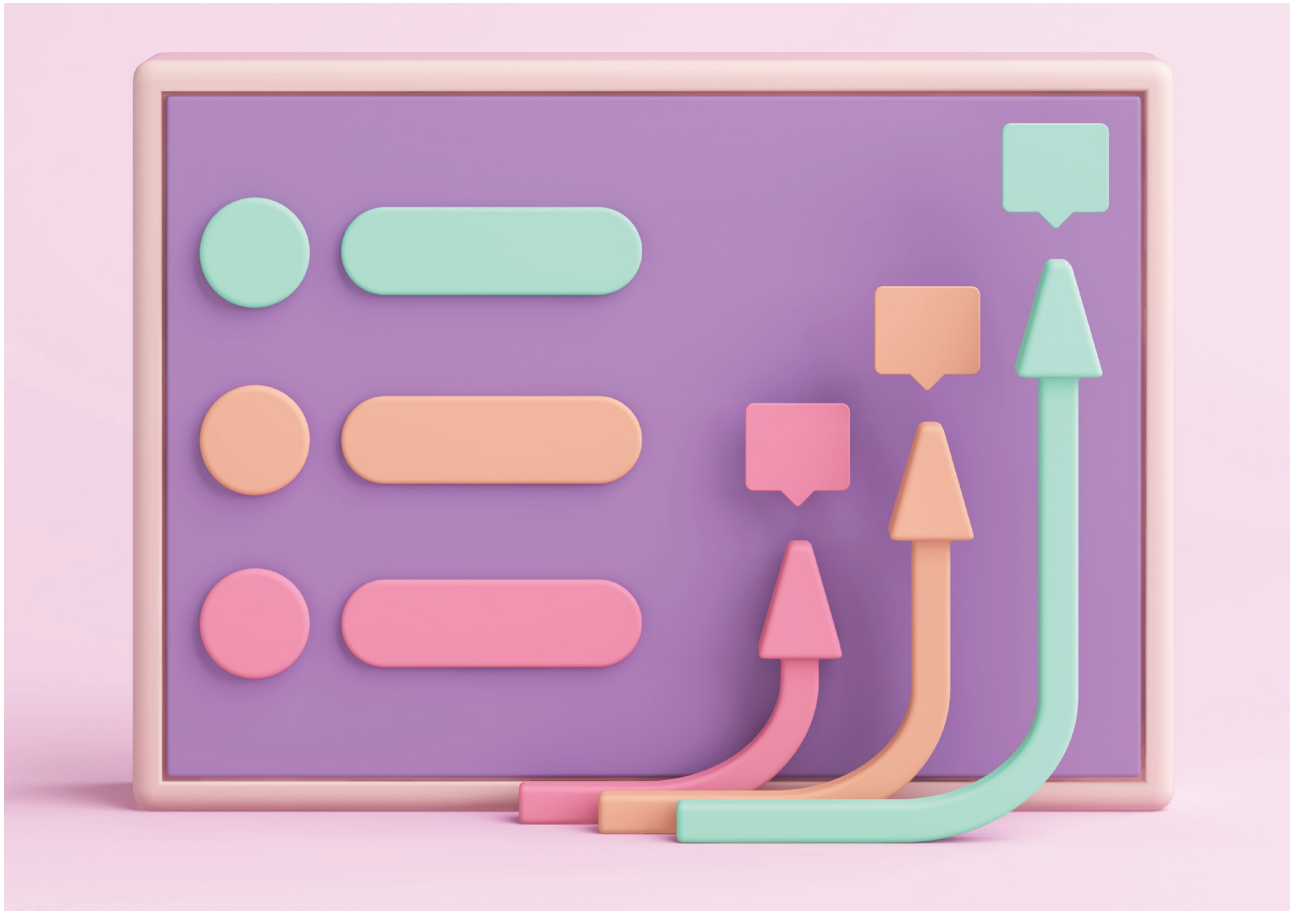

Sarah Zeus

KNIME FÜR EINSTEIGER: SVERWEIS MIT JOINER-NODE

In der letzten Ausgabe haben Sie erfahren, wie Sie mithilfe der „GroupBy"-Node Daten ganz einfach mit wenigen Klicks gruppieren und so mit KNIME deutlich Zeit sparen können. Wenn Sie Erfahrung mit Excel haben und über einen Wechsel nachdenken, fragen Sie sich sicherlich, wie Sie gewohnte Funktionen auch in KNIME nutzen können. Im vorliegenden Beitrag widmet sich Sarah Zeus daher einer häufig verwendeten Funktion: dem SVERWEIS und der Umsetzung mit KNIME.

DIE AUTORIN

Sarah Zeus ist Online Marketing Consultant bei The Boutique Agency, einer Digital-Marketing-Agentur mit Sitz in München. Seit drei Jahren beschäftigt sie sich mit KNIME als Tool zur Automatisierung und Dokumentation wiederkehrender Analysen.

In der Regel gibt es bei der Erstellung von Workflows in KNIME mehr als einen Ansatz, der zum Ziel führt. So beim SVERWEIS, denn auch dafür bietet das Tool mehrere Möglichkeiten. Vorgestellt werden soll hier die gängigste: die Joiner-Node.

Zunächst ist es sinnvoll, sich noch einmal die Funktionsweise und Anwendungsbereiche der Funktion zu vergegenwärtigen. Wer damit vertraut ist, springt am besten gleich zum nächsten Abschnitt.

Funktionsweise und Anwendungsbereiche für den SVERWEIS

In Excel ermöglicht der SVERWEIS (eng. VLOOKUP – „vertical lookup"), in einer angegebenen Matrix nach einem bestimmten Wert zu suchen und einen zugehörigen Wert aus einer anderen Spalte abzurufen. Die Syntax lautet wie folgt:

=SVERWEIS(Suchkriterium; Matrix; Spaltenindex; [Bereich_Verweis])

» **Suchkritierium:** Wert, nach dem gesucht wird

Foto: Chawalit Banpot / gettyimages.de

A	B
URL	**Top-Keyword**
https://www.beispiel.de/beispiel-a/	Keyword 1
https://www.beispiel.de/beispiel-b/	Keyword 2
https://www.beispiel.de/beispiel-c/	Keyword 3
https://www.beispiel.de/beispiel-d/	Keyword 4
https://www.beispiel.de/beispiel-e/	Keyword 5
https://www.beispiel.de/beispiel-f/	Keyword 6
https://www.beispiel.de/beispiel-g/	Keyword 7
https://www.beispiel.de/beispiel-h/	Keyword 8
https://www.beispiel.de/beispiel-i/	Keyword 9
https://www.beispiel.de/beispiel-j/	Keyword 10

Abb. 1: Tabelle mit URL und Top-Keyword

A	B	C
Keyword	**Cluster**	**Suchvolumen**
Keyword 2	A	2900
Keyword 5	A	1600
Keyword 6	B	260
Keyword 9	C	590
Keyword 1	C	720
Keyword 4	C	1300
Keyword 3	B	880
Keyword 7	B	170

Abb. 2 : Tabelle mit Keyword, Cluster und Suchvolumen

Abb. 3: Für jede Datei öffnet sich eine geeignete Reader-Node.

- **Matrix:** Zellbereich, in dem die Daten gesucht werden. Die erste Spalte in der Matrix muss das Suchkriterium enthalten.
- **Spaltenindex:** die Nummer der Spalte, in der sich der Rückgabewert befindet
- **[Bereich-Verweis]:** Wahrheitswert, mit dem man angibt, ob nach einer ungefähren oder exakten Übereinstimmung gesucht werden soll

Einfach gesagt gibt man an, was man in welchem Bereich sucht und aus welcher Spalte man den zugehörigen Wert abrufen möchte.

Der SVERWEIS und damit auch die Joiner-Node in KNIME sind immer dann hilfreich, wenn Sie Daten aus zwei Tabellen zusammenführen möchten, überprüfen wollen, ob Daten aus einer Tabelle auch in einer anderen Tabelle vorhanden sind, oder zwei Tabellen miteinander vergleichen möchten.

Ein simples Beispiel: Sie haben eine Tabelle mit einer Liste an URLs in Spalte A. In Spalte B befindet sich zu jeder URL das dazugehörige Top-Keyword. Über ein beliebiges Tool haben Sie das Suchvolumen zu den Keywords aus der Liste recherchiert, was Ihnen in einer separaten Tabelle vorliegt. Angenommen, Sie möchten die beiden Tabellen zusammenführen, indem Sie das Suchvolumen in der ersten Tabelle ergänzen. In Excel funktioniert das mit dem SVERWEIS. In KNIME verwendet man hierfür die Joiner-Node. Mithilfe dieser Node lassen sich zwei Tabellen über gemeinsame Werte einfach zusammenführen. Während die SVERWEIS-Funktion vier Argumente benötigt, sind für die Joiner-Node in KNIME nur zwei notwendig: die Spalten mit gemeinsamen Werten, die zusammengeführt werden sollen, sowie die Spalten mit den Werten, die in der Ergebnistabelle erscheinen sollen. Die Auswahl dieser Spalten benötigt nur wenige Klicks.

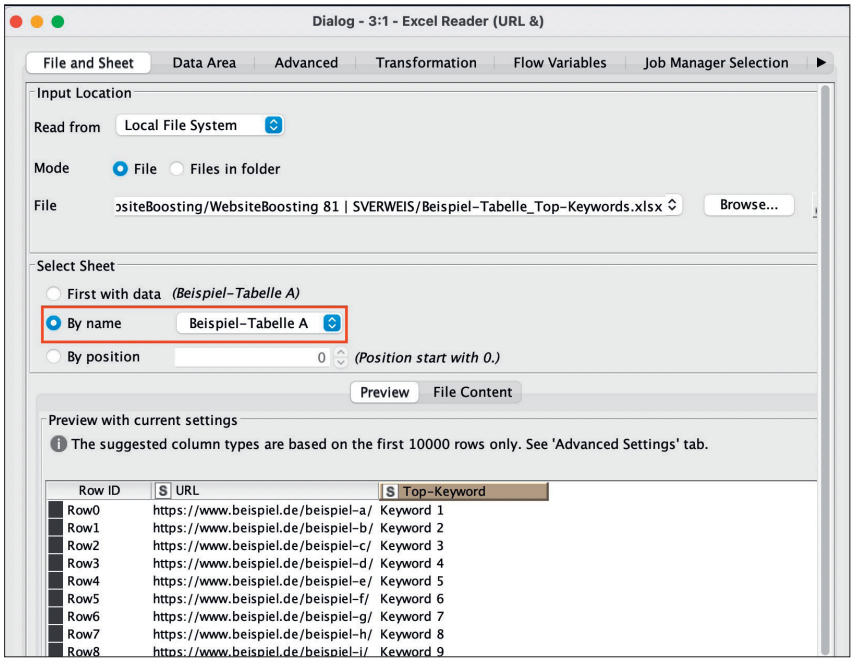

Abb. 4: Gegebenenfalls Tabellenblatt in der Excel-Reader-Node auswählen

SVERWEIS in mit der Joiner-Node

Für das genannte Beispiel werden zwei einfache Tabellen verwendet (Abb. 1 und Abb. 2). Die Keyword-Spalten sind die Spalten mit den gemeinsamen Werten, auf Basis derer die Tabellen zusammengeführt werden. Ein Vorteil von KNIME ist hierbei, dass es egal ist, ob sich hinzuzufügende Spalten rechts oder links von den Keyword-Spalten befinden. Bei einem SVERWEIS hingegen können nur Werte abgerufen werden, die rechts von der Suchspalte stehen. In Tabelle 2 fehlen für das Beispiel Keyword 8 und Keyword 10.

Schritt 1: Zunächst legen Sie in KNIME über File › New einen neuen Workflow an. Ziehen Sie dann die beiden Dateien, die Sie zum Zusammenführen der Daten benötigen, rechts in die Arbeitsfläche von KNIME.

Für die Datei, die dem Workflow hinzugefügt wird, öffnet sich eine geeignete Reader-Node – im vorliegenden Beispiel jeweils die Excel-Reader-Node, genauso würde dies aber beispielsweise mit einer CSV-Datei funktionieren (Abbildung 3).

Mit Doppelklick auf eine Node kann man diese konfigurieren. Beim erstmaligen Hinzufügen der Dateien öffnet sich der Konfigurationsdialog automatisch. Wenn eine Datei über mehrere Tabellenblätter verfügt, kann man über Select Sheet › By name das Tabellenblatt auswählen, das verwendet werden soll (vgl. Abbildung 4). In der Regel muss in der Excel-Reader-Node ansonsten nichts angepasst werden. Mit Klick auf „OK“ schließt sich der Dialog, über Rechtsklick auf die Node kann sie mit der Auswahl von „Execute“ ausgeführt werden, die Ampel der Node springt auf Grün.

Schritt 2: Als Nächstes müssen beide Reader-Nodes mit der Joiner-Node verbunden werden. Diese findet man am schnellsten, indem man links

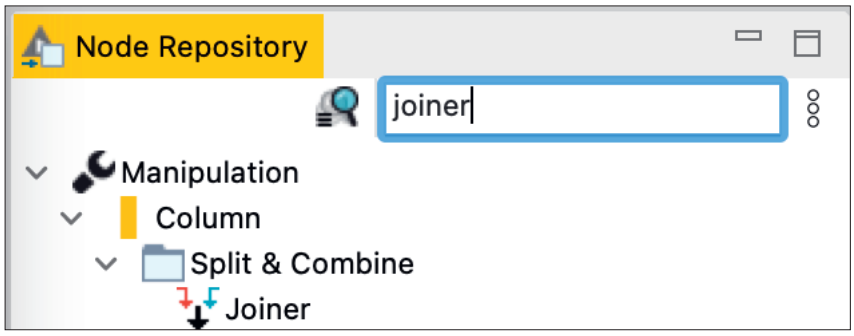

Abb. 5: Im Node-Repository nach Joiner suchen

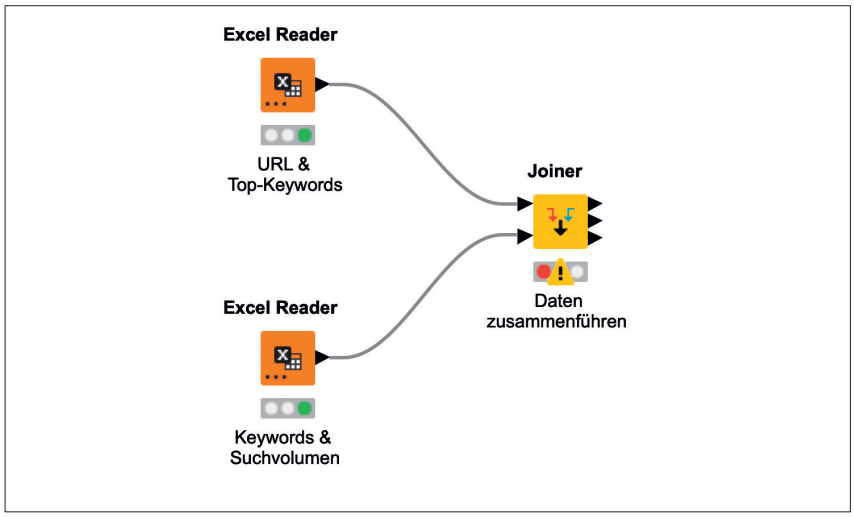

Abb. 6: Die Nodes sind einfach miteinander zu verbinden.

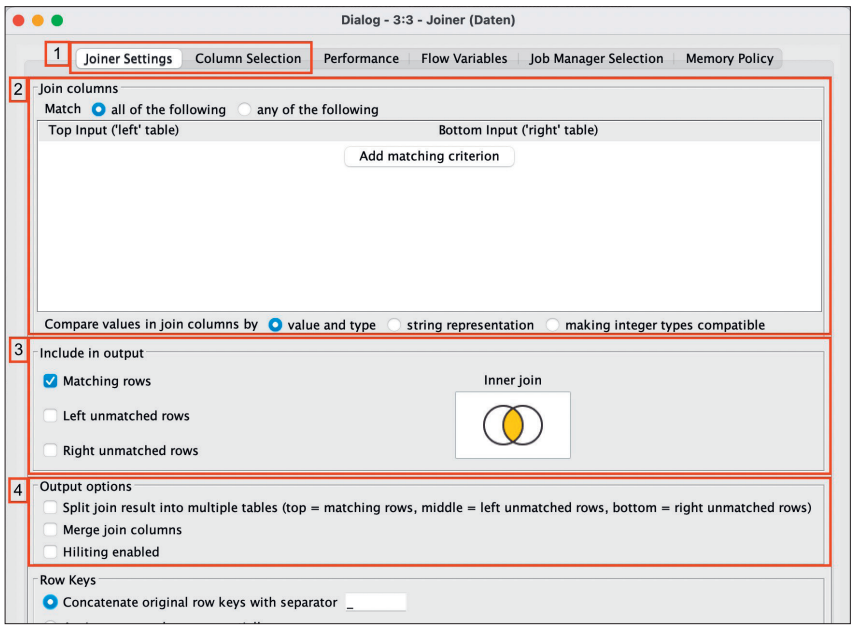

Abb. 7: Alle Bereiche, die in den „Joiner Settings“ angepasst werden sollten

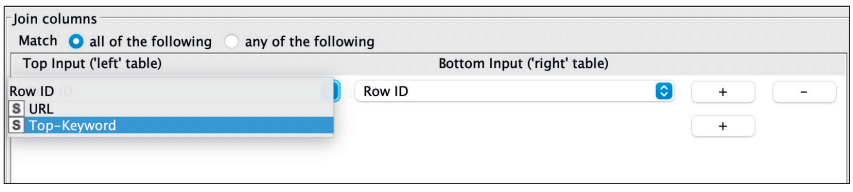

Abb. 8: Matching-Spalten auswählen

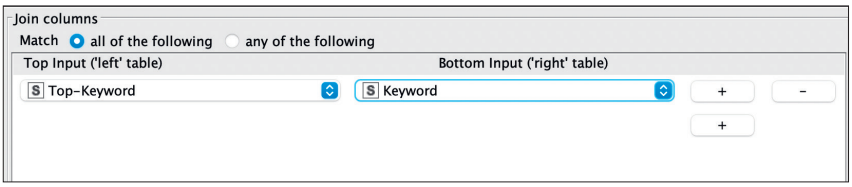

Abb. 9: Zusammengeführt werden die Keyword-Spalten aus beiden Tabellen.

Abb. 10: Left outer join wählen

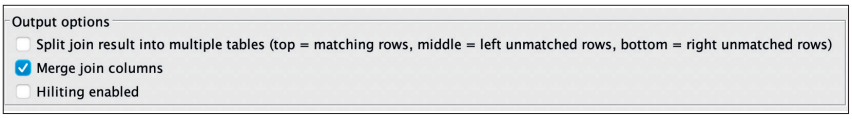

Abb. 11: „Merge join columns“ auswählen

unten im Node-Repository nach „Joiner“ sucht (Abbildung 5). Ziehen Sie die Node in die Arbeitsfläche.

Schritt 3: Die Reader-Nodes verfügen über je einen Ausgang, die Joiner-Node über zwei Eingänge. Verbinden Sie die Nodes miteinander, indem Sie mit der Maus eine Verbindung vom Ausgang der Reader-Node zu einem der Eingänge der Joiner-Node ziehen.

Schritt 4: Die Ampel der Joiner-Node steht nun noch auf Rot mit einem gelben Ausrufezeichen, sie muss also noch konfiguriert werden (Abbildung 6). Einstellungen müssen in den beiden ersten Reitern – „Joiner Settings“ und „Column Selection“ – vorgenommen werden (Abbildung 7, Ziffer 1).

Unter den „Joiner Settings“ gliedert sich der Dialog in mehrere Abschnitte.

Im Abschnitt „Joiner columns“ (Abbildung 7, Ziffer 2) muss zunächst ein Matching Criterion hinzugefügt werden, das heißt, die Spalten mit gemeinsamen Werten werden aus beiden Tabellen ausgewählt, die zusammengeführt werden sollen. In unserem Beispiel sind dies die Keyword-Spalten, die in beiden Tabellen existieren. Nach dem Klick auf „Add matching criterion“ kann die entsprechende Spalte in beiden Tabellen einfach über eine Drop-down-Liste ausgewählt werden.

Als „Top Input“ bzw. „left table“ wird dabei die Tabelle bezeichnet, die mit dem oberen Eingang der Joiner-Node verbunden ist. Die Tabelle, die mit dem unteren Eingang verbunden ist, ist der „Bottom Input“ bzw. „right table“ (Abbildungen 8 und 9).

Schritt 5: Im nächsten Abschnitt „Include in output“ kann man auswählen, welche Zeilen in die Output-Tabelle aufgenommen werden sollen (Abbildung 10):

Bei „Matching rows“ (Inner join) werden nur Zeilen aufgenommen, für die in den Keyword-Spalten ein identischer Wert in beiden Tabellen existiert, das heißt, Keywords, die nicht in beiden Tabellen auftauchen, werden ausgeschlossen.

Bei „Left unmatched rows“ (Left outer join) werden auch Zeilen aus der linken Eingabetabelle aufgenommen, für die keine passende Zeile in der rechten Eingabetabelle gefunden wird. Wenn es in der Tabelle mit URL und Top-Keyword also ein Keyword

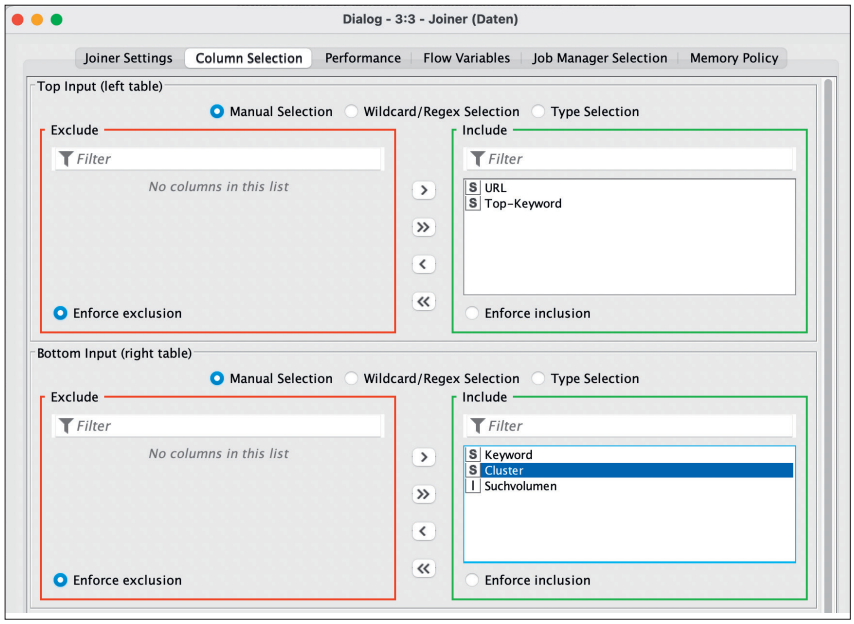

Abb. 12: Spalten auswählen, die nicht übernommen werden sollen

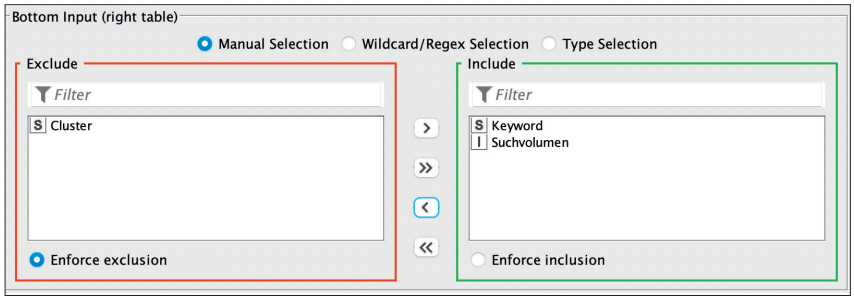

Abb. 13: Gewählte Spalten über Pfeil-Button ausschließen

Row ID	URL	Top-Keyword=Keyword	Suchvolumen
Row0	https://www.beispiel.de/beispiel-a/	Keyword 1	720
Row1	https://www.beispiel.de/beispiel-b/	Keyword 2	2900
Row2	https://www.beispiel.de/beispiel-c/	Keyword 3	880
Row3	https://www.beispiel.de/beispiel-d/	Keyword 4	1300
Row4	https://www.beispiel.de/beispiel-e/	Keyword 5	1600
Row5	https://www.beispiel.de/beispiel-f/	Keyword 6	260
Row6	https://www.beispiel.de/beispiel-g/	Keyword 7	170
Row7	https://www.beispiel.de/beispiel-h/	Keyword 8	?
Row8	https://www.beispiel.de/beispiel-i/	Keyword 9	590
Row9	https://www.beispiel.de/beispiel-j/	Keyword 10	?

Abb. 14: Vorschau auf die Ergebnistabelle

gibt, das in der anderen Tabelle nicht auftaucht, wird die Zeile mit URL und Keyword trotzdem in der Ergebnistabelle erscheinen, nur ohne Suchvolumen.

Bei „Right unmatched rows“ würden nach demselben Muster zusätzlich alle Zeilen aus der rechten Tabelle ausgegeben, für die es keinen Wert in der linken Tabelle gibt.

Für unser Beispiel ist es sinnvoll, die ersten beiden Varianten auszuwählen, so geht keine Zeile aus der Tabelle verloren, auch wenn für ein Keyword kein Suchvolumen enthalten ist.

Schritt 6: Unter „Output options“ sollte „Merge join columns“ ausgewählt werden, so werden die beiden Keyword-Spalten aus beiden Tabellen verschmolzen. Wäre hier kein Haken gesetzt, würden einfach beide Keyword-Spalten in der Ergebnistabelle erscheinen, wobei die Spalte natürlich nur einmal benötigt wird (Abbildung 11).

Unter dem Abschnitt „Row Keys“ erhält jede Zeile durch die Auswahl von „Assign row keys sequentially“ eine neue Row ID.

Schritt 7: Im Reiter „Column Selection“ kann man nun auswählen, welche Spalten in die Ergebnistabelle aufgenommen werden und welche nicht auftauchen sollen. Dafür wählt man einfach die betroffene Spalte und schiebt sie durch Klick auf den Pfeil-Button nach links in den Exclude-Bereich. In unserem Beispiel wird die Spalte „Cluster“ entfernt, die sich in der rechten Tabelle befindet (Abbildungen 12 und 13).

Schritt 8: Durch Klick auf „OK“ schließt sich der Konfigurationsdialog. Die Ampel der Node steht nun auf Gelb, das heißt, sie ist bereit, um ausgeführt zu werden (Rechtsklick auf die Joiner-Node › Execute). Mit Klick auf das Tabellen-Symbol mit Lupe in der Toolbar über dem Arbeitsbereich kann das Ergebnis betrachtet werden (Abbildung 14): Es enthält die URL-Spalte aus der linken Tabelle, die zusammengeführte Keyword-Spalte und die Suchvolumen-Spalte. Die Tabelle enthält Zeilen mit Keywords, für die kein Suchvolumen enthalten ist. Die Spalte „Cluster“ wurde nicht übernommen.

Das Ergebnis kann nun entweder exportiert werden, indem dem Workflow eine Writer-Node (zum Beispiel Excel Writer) in Anschluss an die Joiner-Node hinzugefügt wird, oder als Teil eines umfangreicheren Workflows verwendet werden.

Fazit

Zur Demonstration wurde hier ein sehr einfaches Beispiel gewählt – sicherlich fallen Ihnen viele Situationen ein, für die weitaus umfangreichere Tabellen erstellt werden müssen, bevor man mit einer Analyse starten kann. Arbeitet man mit dem SVERWEIS und sollen Daten aus mehreren verschiedenen Quellen hinzugefügt werden – beispielsweise aus einem Screaming-Frog-Crawl, der Google Search Console, Google Analytics oder Tools wie Ahrefs oder Sistrix –, muss der Prozess mehrfach wiederholt werden. Je nach Analysevorhaben kann das Zusammenführen der Daten aus mehreren Tabellen so sehr aufwendig werden. Vor allem bei regelmäßigen Analysen, für die die Zusammenstellung jedes Mal erneut erfolgen muss, geht hier unnötig Zeit verloren. In solchen Fällen lohnt sich ein Umstieg von Excel auf KNIME besonders. Denn für jede Analyse muss man diesen Prozess nur einmalig definieren. Im Anschluss lässt sich der Workflow dann immer wieder ohne zusätzlichen Aufwand anwenden und in komplexere Workflows integrieren.

Übrigens: Ab Version 5.0.0 erhält KNIME ein verbessertes UI (derzeit zum Test mit eingeschränkten Funktionen verfügbar) und der SVERWEIS kann dann mit noch weniger Klicks durchgeführt werden: Ab dieser Version steht die Node „Value Lookup“ zur Verfügung, die – mit etwas weniger Funktionen – nach demselben Muster wie hier beschrieben funktioniert. Sie benötigt nur noch die Angabe einer Lookup Column und einer Key Column und führt zum gleichen Ergebnis. ¶