Rebecca Schwarz

KNIME: Die SEO-Qualitätssicherung mit COLUMN-COMPARE spart Zeit und Nerven

In Ausgabe #81 konnten Sie erfahren, wie in KNIME der JOINER verwendet wird, um Daten miteinander zu kombinieren. Als Erweiterung dazu widmet sich Rebecca Schwarz nun dem COLUMN-COMPARATOR. Damit können die Inhalte verschiedener Datensätze sehr einfach und schnell verglichen werden. Sie stellt Ihnen vor, wie das in der SEO-Qualitätssicherung Zeit und Nerven spart und wie KNIME den Abgleich als wiederholbaren Task sehr effizient abbildet.
In einem konkreten Anwendungsfall wird Schritt für Schritt gezeigt, wie angeforderte Title-Tags mit der tatsächlichen Umsetzung auf einer Website verglichen werden. Hierbei werden zwei Möglichkeiten veranschaulicht: der Vergleich mithilfe eines Crawl-Exports und der Vergleich mit einer Echtzeitabfrage der Website direkt in KNIME.

DIE AUTORIN

Rebecca Schwarz ist SEO-Consultant bei der get traction GmbH. Ihr Schwerpunkt liegt in der Entwicklung von SEO-Strategien und der Durchführung von Content-Audits. Um größere Datenmengen effizient zu verarbeiten und um bei wiederkehrenden SEO-Tasks Zeit zu sparen, nutzt sie die Open-Source-Software KNIME.

Die Qualitätssicherung (kurz: QS) zählt in der SEO zu den wichtigsten Tasks, um sicherzustellen, dass Anforderungen mit der Umsetzung tatsächlich übereinstimmen. Jedoch kann diese Aufgabe sehr aufwendig sein, da sie Genauigkeit erfordert und häufig nicht im ersten Anlauf abgeschlossen ist. Denn: Fallen in der QS-Phase Fehler auf, müssen Anpassungen durchgeführt werden, die wiederum zu einer QS-Phase führen. Für diese Art der Wiederholbarkeit ist die Software KNIME prädestiniert. Ist ein Workflow (= Aneinanderreihung von Knoten in der Anwendungsoberfläche) einmal angelegt und konfiguriert, können jederzeit neue Daten der gleichen Struktur durch den Workflow fließen. Wie hilfreich KNIME in der QS-Phase sein kann, soll am COMPARE-Knoten mit folgendem Anwendungsfall veranschaulicht werden: Title-Tags von ausgewählten URLs wurden anhand eines Übergabedokuments optimiert. Nun soll überprüft werden, ob die Anpassungen der Title mit den angeforderten Title-Optimierungen aus dem Übergabedokument übereinstimmen oder ob die Title davon abweichen.

Vorbereitung zum Start in KNIME

Für den Aufbau eines Workflows in KNIME sollte immer folgende Überlegung vorangestellt werden: Welche Daten stehen zur Verarbeitung in welcher Form zur Verfügung? Was soll durch den Workflow herausgefunden werden? Wie sollen die Daten aufbereitet werden, um schlussendlich einen sinnvollen Export aus der Software zu erhalten? Dabei hilft eine möglichst genaue Formulierung des Anwendungsfalls: Es soll verglichen werden, ob ein angeforderter Title mit dem aktuellen Title-Tag einer URL in einem Livesystem übereinstimmt. Als Export soll dabei eine Tabelle entstehen, in der gekennzeichnet ist, welche Title korrekt im Livesystem umgesetzt sind und welche nicht. Für den Vergleich in KNIME stehen folgende Datensätze zur Verfügung:

» Datensatz eins: eine Tabelle mit definierten Werten für den Title-Tag und der zugehörigen URL (hier im .xlsx-Format)
» Datensatz zwei: ein aktueller Crawl-Export der Website (hier im .csv-Format)

Schritt eins: Wie bei allen Workflows in KNIME wird immer damit begonnen, die oben beschriebenen Daten in die KNIME-Umgebung einzulesen. Hierfür wird jeweils ein READER-Knoten benötigt. Für .csv-Dateien wird ein CSV-READER verwendet, für Excel-Dateien ein EXCEL-READER. Die Datensätze werden in die passenden Knoten importiert und anschließend ausgeführt. Der Import der Datei funktioniert entweder über die „Browse“-Funktion innerhalb des Knotens oder kann per Drag-and-drop erfolgen. Hierzu wird einfach die Datei auf die grafische Oberfläche des Knotens gezogen.

Datensatz 1: Excel-Datei

	A	B
1	URL	definierter Title
2	example.org/beispiel1.html	neuer Title 1
3	example.org/beispiel2.html	neuer Title 2
4	example.org/beispiel3.html	neuer Title 3
5	example.org/beispiel4.html	neuer Title 4
6	example.org/beispiel5.html	neuer Title 5

Datensatz 2: CSV-Datei

```
1  Address;Title
2  example.org/beispiel1.html;neuer Title 1
3  example.org/beispiel2.html;neuer Title 2
4  example.org/beispiel3.html;neuer Title 3
5  example.org/beispiel4.html;alter Title 1
6  example.org/beispiel5.html;alter Title 2
```

Abb. 1: Datensätze für die Arbeit in KNIME

Anschließend wird der Knoten über › Rechtsklick › „Execute“ ausgeführt. Ist die Ampel (drei Punkte unterhalb jedes Knotens) grün, sind die Datensätze korrekt eingelesen und es kann mit dem Workflow weitergehen (Abbildung 2). Hinweis: Sollten in Titles Umlaute, Sonderzeichen oder Emojis enthalten sein, kann es vorkommen, dass diese in der Import-Vorschau nicht korrekt angezeigt werden. Ist dies der Fall, muss im Reiter „Encoding“ wahrscheinlich der Typ auf „UTF-8“ oder „UTF-16“ umgestellt werden.

Schritt zwei: Als Nächstes müssen die beiden Datensätze miteinander verknüpft werden. In Ausgabe #81 wurde dazu bereits der JOINER-Knoten vorgestellt. Dieser kommt hier zum Einsatz, um Daten der beiden Datensätze miteinander zu kombinieren und in einem Datensatz zu vereinen. Dadurch können die Daten zu einem späteren Zeitpunkt verglichen werden. Wie auch im SVERWEIS() in Excel wird eine Referenzspalte benötigt, um die Datensätze korrekt zu verknüpfen. Im vorliegenden Fall wird die URL beider Datensätze verglichen, weshalb diese Spalten zu Hilfe genommen werden. Wie in Abbildung 2 zu sehen werden im Workflow deshalb jeweils die READER-Knoten mit dem JOINER verbunden.

In der Excel-Datei heißt die Spalte „URL“, im Crawl-Export ist die Spalte mit „Address“ benannt. In Abbildung 3 ist zu sehen, wie die Konfiguration des JOINER-Knotens aussieht. Neben der Auswahl der Spalten werden für die Weiterverarbeitung Daten verwendet, die in beiden Datensätzen vorkommen, und zusätzlich Daten, die in der Excel-Datei (Datensatz eins) und nicht im Crawl-Export (Datensatz zwei) vorhanden sind. Das ist der sogenannte LEFT JOIN. Diese Art der Verknüpfung stellt sicher, dass keine angeforderten Title-Optimierungen verloren gehen, wenn zum Zeitpunkt des Crawls keine passende URL existiert hat.

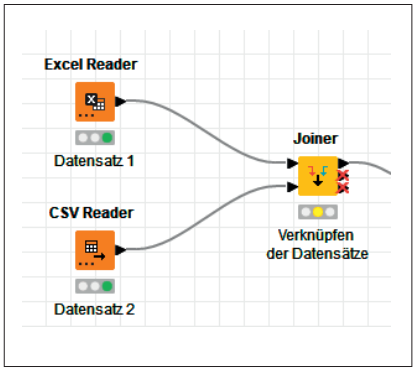


Abb. 2: Verknüpfung der Datensätze im JOINER-Knoten

Die Vereinigung der beiden Datensätze ist in Abbildung 4 zu sehen. Alle Spalten der beiden Datensätze sind nun innerhalb einer neuen Tabelle vorhanden. Tipp: Mit Rechtsklick auf den JOINER-Knoten und die Auswahl „Join result“ kann überprüft werden, ob es Zeilen innerhalb der Datensätze gibt, die nicht verknüpft werden konnten. Ist das der Fall, befinden sich leere Zellen in der Vorschau, die durch rote Fragezeichen gekennzeichnet werden.

Schritt drei: Nach dieser Vorbereitung geht es nun in den eigentlichen Vergleich. An den JOINER-Knoten wird nun ein weiterer Knoten angebunden, der COLUMN-COMPARATOR (dieser Knoten ist, wie alle hier vorgestellten Knoten, Teil der Standardinstallation in

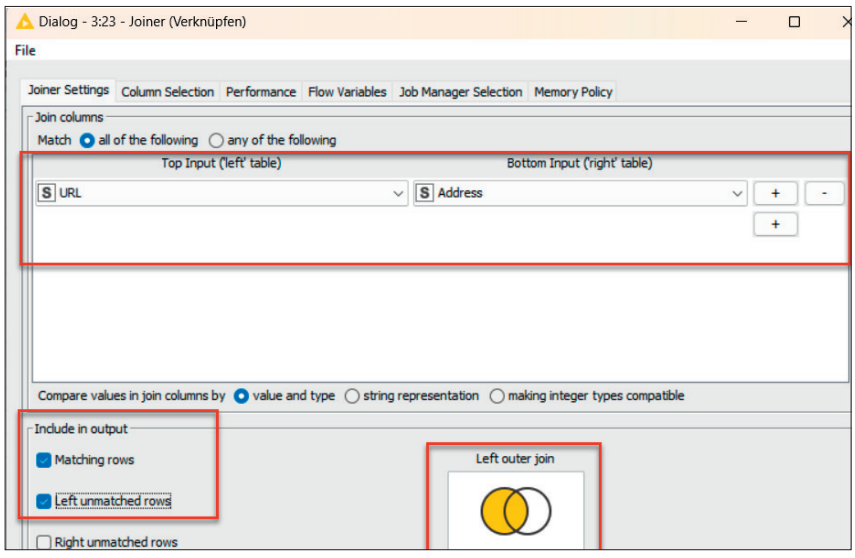


Abb. 3: Konfiguration des JOINER-Knotens

KNIME). Mithilfe dieses Knotens können nun verschiedene Spalten einer Tabelle miteinander verglichen werden. Der oberste Pfeil des JOINER-Knotens wird dazu mit dem COLUMN-COMPARATOR-Knoten verbunden. Sobald die Ampel am Knoten gelb wird, kann mit der Konfiguration gestartet werden (Abbildung 5).

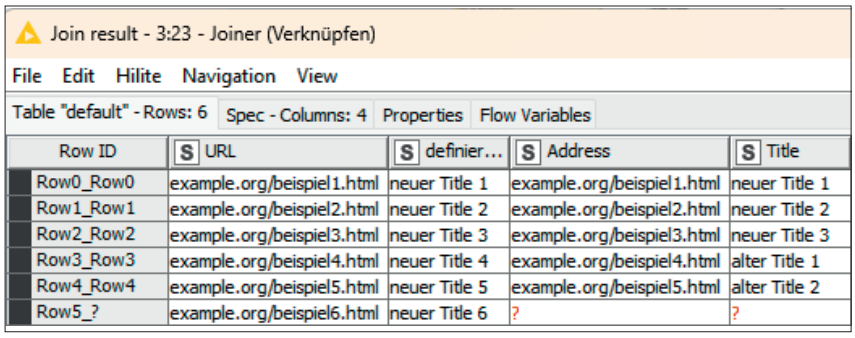
Innerhalb der Konfiguration werden jetzt die Spalten ausgewählt, die miteinander verglichen werden sollen. Hier wird deshalb die Spalte „definierter Title“, die ursprünglich aus Datensatz eins importiert wurde, mit der Spalte „Title“ verglichen, die aus dem Crawl-Export stammt (Abbildung 6).

Die Konfiguration des COLUMN-COMPARATOR-Knotens im Einzelnen:

» Column Left: „definierter Title“
» Column Right: „Title“
» Operator: == (bedeutet in der Abfrage „stimmt überein“)
» Operator result „true“: USER_DEFINED; Tag: TRUE
» Operator result „false“: USER_DEFINED; Tag: FALSE
» New Column: „Title-Tag gleich?“

Das Ergebnis des Knotens ist, dass eine neue Spalte ergänzt wird, die anzeigt, ob die Werte übereinstimmen oder nicht. Bei Übereinstimmung wird in die neue Spalte „TRUE“ geschrieben, bei Unterschieden steht in der neuen Spalte „FALSE“. Der Name, der im Bereich „New Column“ eingetragen wurde, ist die Benennung der neuen Spalte. Sollen noch weitere Spalten verglichen werden, können im Workflow entsprechend weitere COLUMN-COMPARATOR-Knoten angehangen werden. Eine naheliegende Ergänzung wäre hier, nicht nur den Title, sondern auch die Meta-Descriptions oder die H1-Überschrift zu vergleichen, wenn diese optimiert wurden.

Schritt vier: Im letzten Schritt wird nun ein EXCEL-WRITER angehängt, um die Erkenntnisse des Vergleichs expor-



Row ID	URL	definier...	Address	Title
Row0_Row0	example.org/beispiel1.html	neuer Title 1	example.org/beispiel1.html	neuer Title 1
Row1_Row1	example.org/beispiel2.html	neuer Title 2	example.org/beispiel2.html	neuer Title 2
Row2_Row2	example.org/beispiel3.html	neuer Title 3	example.org/beispiel3.html	neuer Title 3
Row3_Row3	example.org/beispiel4.html	neuer Title 4	example.org/beispiel4.html	alter Title 1
Row4_Row4	example.org/beispiel5.html	neuer Title 5	example.org/beispiel5.html	alter Title 2
Row5_?	example.org/beispiel6.html	neuer Title 6	?	?

Abb. 4: Ergebnistabelle des JOINER-Knotens

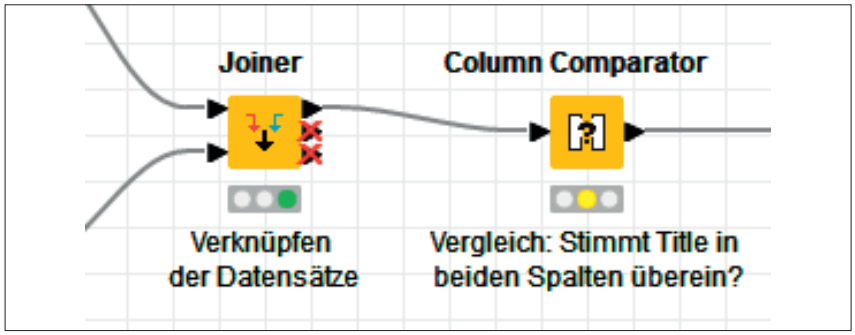


Abb. 5: Verbindung zum COLUMN-COMPARATOR

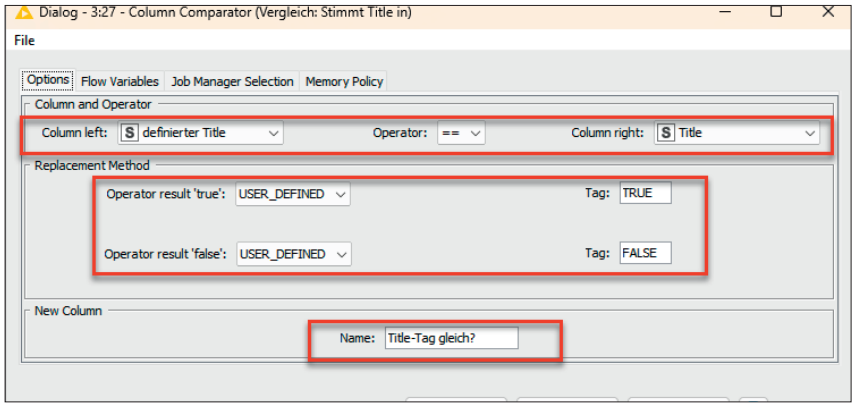


Abb. 6: Konfiguration des COLUMN-COMPARATOR-Knotens

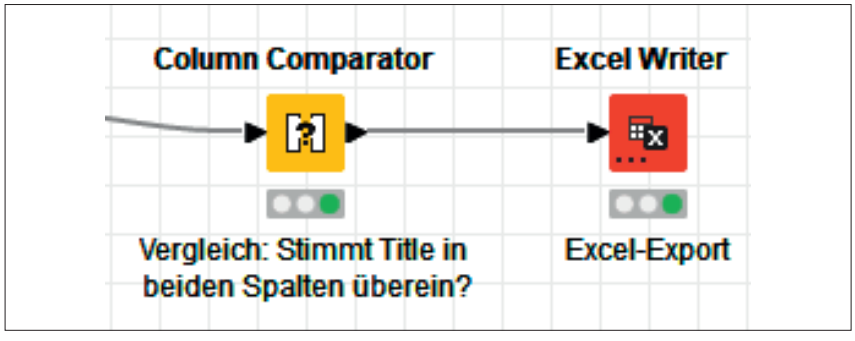


Abb. 7: Verbindung zum EXCEL-WRITER

A	B	D	E
URL	definierter Title	Title	Title-Tag gleich?
example.org/beispiel1.html	neuer Title 1	neuer Title 1	TRUE
example.org/beispiel2.html	neuer Title 2	neuer Title 2	TRUE
example.org/beispiel3.html	neuer Title 3	neuer Title 3	TRUE
example.org/beispiel4.html	neuer Title 4	alter Title 1	FALSE

Abb. 8: Export-Datei mit Abgleich des Title-Tags

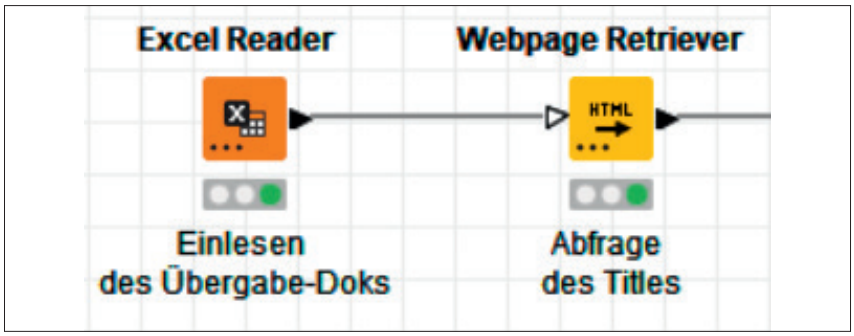


Abb. 9: Verbindung zum WEBPAGE-RETRIVER

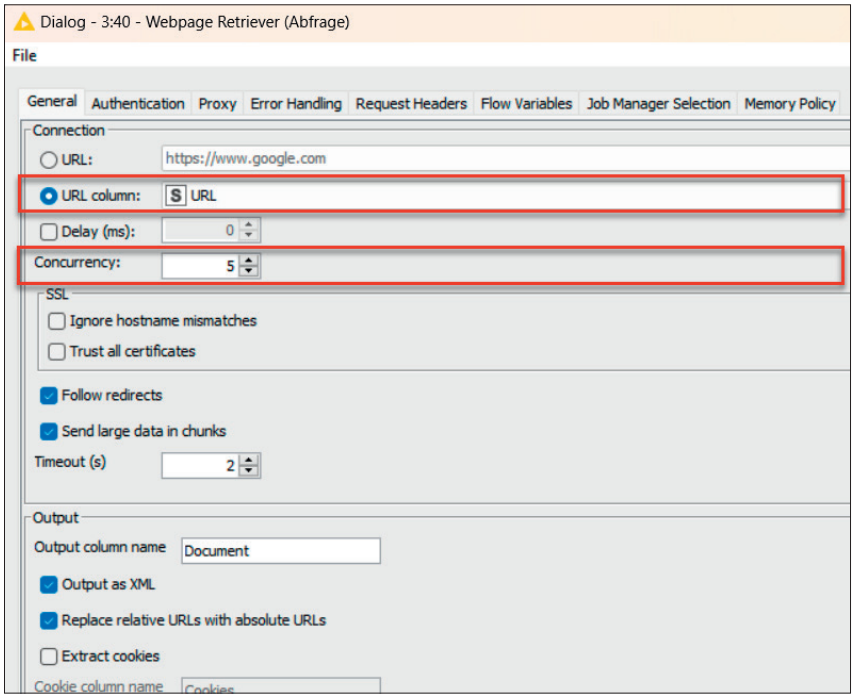


Abb. 10: Konfiguration des WEBPAGE-RETRIEVERS

tieren und teilen zu können (Abbildung 7). Die resultierende Export-Datei (Abbildung 8) kann später übergeben werden, um die Title, die nicht mit den Zielwerten übereinstimmen, anzupassen.

An diesem Punkt wird der Stellenwert der Wiederholbarkeit deutlich. Denn sobald weitere Anpassungen an den Titles vorgenommen wurden, kann der Workflow einfach noch einmal durchgeführt werden! Dazu muss nur der Crawl-Export aktualisiert und wieder per Drag-and-drop in den Workflow eingelesen werden. Anschließend wird der Workflow einfach erneut ausgeführt. Die neue Export-Datei zeigt anschließend an, ob die Werte nun tatsächlich alle mit den Zielwerten übereinstimmen. Tipp: Es kann bereits mit Rechtsklick auf den COLUMN-COMPARATOR › „Input with comparison column“ überprüft werden, ob im zweiten Anlauf noch Unterschiede der Titles vorhanden sind. Wenn nicht, ist eine Export-Datei natürlich unnötig.

Ein weiterer Knoten für noch mehr Automatisierung in KNIME

Nun wird, wie im Teaser angekündigt, noch eine weitere Möglichkeit aufgezeigt, um Website-Inhalte mit angeforderten Optimierungen zu vergleichen. Die zuvor benötigte Crawl-Datei wird hierbei durch einen neuen Knoten, den WEBPAGE-RETRIEVER, ersetzt. Das Besondere: Der Knoten kann HTML-Websites abrufen und ermöglicht es so, das gesamte HTML-Dokument einer Webpage analysieren zu können.

Schritt eins: Zum Start dieses Workflows wird deshalb nur eine Datei mit URLs und definierten Titles benötigt. Dafür kann der bestehende Workflow wiederverwendet werden, indem der EXCEL-READER-Knoten mit Datensatz eins mit „Strg + C“ kopiert wird und mit dem WEBPAGE-RETRIVER verbunden wird (Abbildung 9).

Schritt zwei: Zur Konfiguration des WEBPAGE-RETRIEVERS muss ausgewählt werden, in welcher Spalte sich die URLs befinden, die abgefragt werden sollen. Zusätzlich sollte der Wert für „Concurrency“ erhöht werden, da ansonsten nur eine URL nach der anderen abgefragt wird. Als Empfehlung ist als Wert „5“ zu setzen. Alle anderen Einstellungen können, wie in Abbildung 10 gezeigt, beibehalten werden. Mit der Ausführung des Knotens werden für alle enthaltenen URLs die HTML-Dokumente abgerufen und in Form eines XML-Dokuments in einer neuen Spalte namens „Document“ an den Datensatz angefügt.

Schritt drei: Nun müssen die entsprechenden Daten, hier die Title-Tags, aus dem XML-Dokument extrahiert werden. Dafür wird die Abfragesprache für XML-Dokumente, XPath, benötigt. Auch hierfür gibt es in KNIME einen eigenen Knoten, den XPATH-Knoten. Dieser wird an den Workflow angehangen und kann anschließend sehr einfach konfiguriert werden (Abbildung 11). Tipp: Zur Konfiguration sind nicht zwingend Kenntnisse in XPath notwendig. Wenn in der XML-Preview ein Doppelklick auf das zu öffnende Title-Tag ‹title› gemacht wird, erscheint automatisch der Befehl zum Ansteuern des Elements als XPath-Abfrage. Hier: /html/head/title. Für die Abfragen von Standardelementen helfen aber auch Cheat-Sheets für XPath. Für den Inhalt des Title-Tags ist das in

XPath: //title. Beide Abfragemöglichkeiten führen zum gleichen Ergebnis.

Durch Ausführung des Knotens wird der Title aus allen XML-Dokumenten extrahiert und in einer neuen Spalte des Datensatzes ergänzt. Die Information, die im vorherigen Workflow über einen Crawl-Export in KNIME integriert werden musste, konnte nun direkt in der Anwendung abgefragt werden. Dies ist vor allem bei einem kleinen URL-Set ein schneller Weg, um an Informationen zu kommen.

Schritt vier: Zur Reduzierung des Datensatzes wird jetzt die Spalte „Document" mit den XML-Dokumenten wieder aus dem Datensatz entfernt, da der Datensatz sonst zu groß ist, um einen Excel-Export zu erstellen. Zum Herausfiltern wird hinter den XPATH-Knoten ein COLUMN-FILTER-Knoten angefügt. Mit Rechtsklick auf „Configure" wird die Spalte „Document" mit der Pfeilauswahl auf die linke Seite „Excludes" innerhalb des Knotens verschoben und ist somit im weiteren Verlauf des Workflows nicht mehr vorhanden.

Alle weiteren Knoten und deren Konfiguration unterscheiden sich jetzt nicht mehr vom vorherigen Workflow. Denn nachfolgend wird der Workflow durch den COLUMN-COMPARATOR ergänzt. In Abbildung 12 sind noch einmal die beiden Workflows visualisiert.

Fazit:

Für die QS-Phase eignet sich KNIME hervorragend und spart Zeit und Nerven. Nicht zu unterschätzen ist auch der übersichtliche Export, der nicht nur die Durchführung der QS, sondern auch die Anpassung von kleinen Fehlern erleichtert.

In diesem Sinne: viel Spaß mit KNIME in der SEO-Qualitätssicherung!

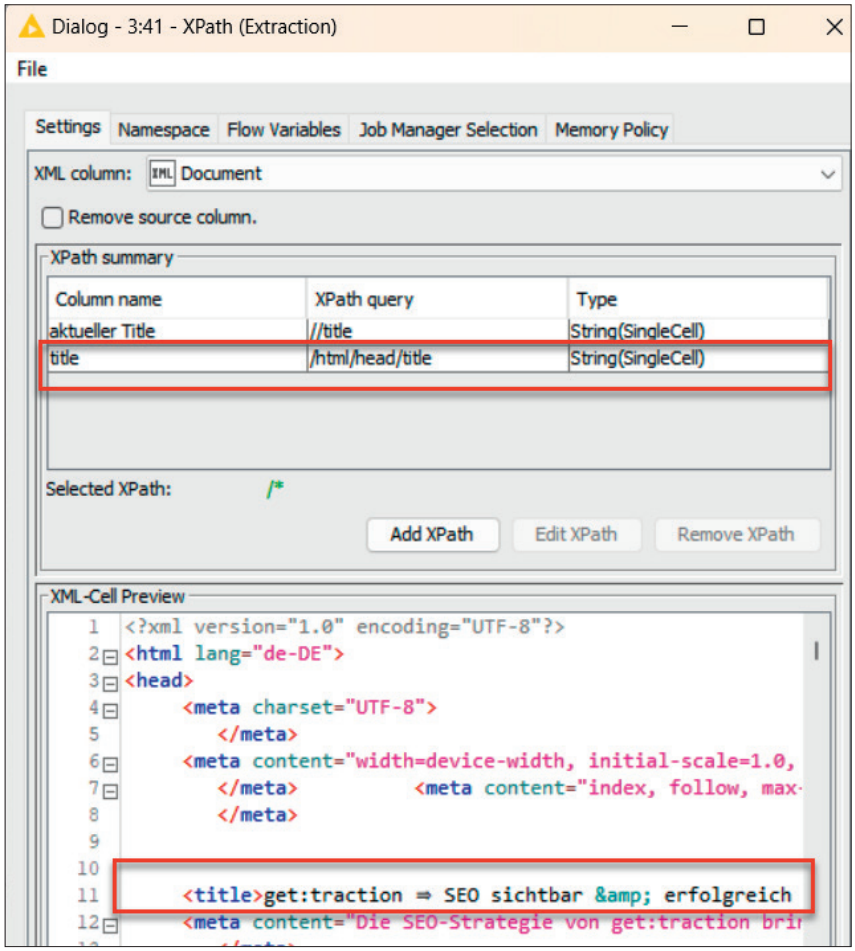


Abb. 11: Konfiguration des XPATH-Knotens zur Extraktion des Title-Tags

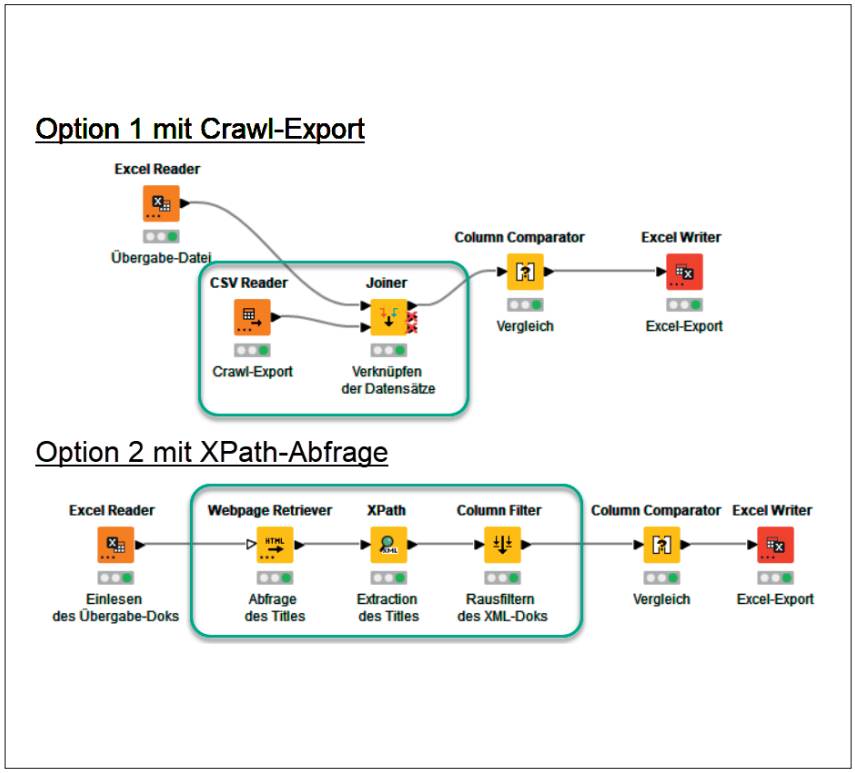


Abb. 12: Beide Workflows zum Vergleich der Title-Tags