

R4SE0: Automatisierte Reports mit R generieren (Teil 1)

R-Leuchtungen! Teil 9

WEBSITE BOOSTING

#70

inkl. Ask Google!



QUALITÄT WISSEN FÜR BESSERE WEBSITE

ONLINE #FAKES!

Warum wir darauf reinfallen
und wie wir uns und andere
besser schützen können

PERFORMANCE-MARKETING

LANDINGPAGE- OPTIMIERUNG

Wer Geld für Klicks bezahlt, muss optimal willkommen heißen

ERFOLGSHEBEL

SEO FÜR ONLINE- SHOPS

Start einer neuen Serie speziell für E-Commerce-Websites .

MOMENT OF TRUTH

OPTIMIERUNG DES SNIPPETS

Wenn Google die Description nicht anzeigt, besteht Handlungsbedarf

R-Leuchtungen! Teil 9: Title-Geddon – war

da was bei Ihnen? – websiteboosting.com

In den Ausgaben bis 54 bis 58 der Website Boosting konnten Sie in der Serie „R4SEO“ von Patrick Lürwer nachvollziehen, wie man die kostenlose Software R verwendet, was sie leistet und wie man sie nutzbringend für die eigene Arbeit für SEO bzw. die Aufklärung im Online-Marketing einsetzen kann. R...

In der letzten Ausgabe der Website Boosting gab Tobias Aubele einen sehr guten Einblick in die Potenziale von R, um auf einfache und reproduzierbare Weise Daten über die Google-Analytics-API abzufragen. Diese neue Artikelserie von Patrick Lürwer baut darauf auf und zeigt, wie mittels R Reports erstellt werden können, die sich zu einem bestimmten Zeitpunkt automatisch aktualisieren und per Mail verschicken. Im ersten Teil geht es nun zunächst um das Abfragen verschiedener APIs für die Datenbeschaffung. In den folgenden Teilen werden die Daten aggregiert, visualisiert und in einem Report zusammengefasst.

Tipp

Um Ihnen das Abtippen des Codes zu ersparen, können Sie sich das Skript zu diesem Artikel unter folgendem Link auf GitHub herunterladen. Dennoch empfehle ich Ihnen, den Code selbst zu schreiben, um ein Gefühl für die Syntax zu bekommen. einfach.st/r4seocode

Das Google Data Studio (GDS) bietet eine einfache Möglichkeit, Daten aus dem Google-Kosmos – bspw. aus Google Analytics (GA) und der Google Search Console (GSC) – abzufragen, zu visualisieren und als Reports aufzubereiten. Problematisch wird es allerdings, wenn andere Datenquellen angebunden werden sollen. Denkbar sind hier APIs anderer Anbieter wie Sistrix. In solchen Fällen muss der „Umweg“ über Google Spreadsheets und Google Apps Scripts gegangen werden, um die Abfragen selbst zu programmieren. Dies ist durchaus ein valides Vorgehen, das jedoch schnell an seine Grenzen stoßen kann. Denn meist sind Reports nur der erste Schritt. Sobald in

diesen Auffälligkeiten erkennbar werden, schließt eine Analyse an, die eine umfassende Beschäftigung mit den zugrunde liegenden Daten erfordert. Derartige Analysen verlangen zumeist komplexe Datentransformationen und -aggregationen, für die Dashboard-Softwares, wie das GSA, mit ihren begrenzten Filtern- und Manipulationsoptionen naturgemäß nicht ausgelegt sind. Ganz abgesehen davon, dass der direkte Zugriff auf die Daten zur Weiterverarbeitung häufig gar nicht gegeben ist. Von großem Vorteil ist es daher, wenn Report- und Analysesoftware in einer Anwendung vereint sind. Diese Anforderungen kann R insofern erfüllen, als die Programmiersprache durch Packages so erweitert werden kann, dass mit ihr sowohl sehr individualisierte Reports erstellt als auch umfangreiche Analysen durchgeführt werden können.

Da Datenanalysen eine sehr komplexe Thematik sind, soll in diesem Artikel jedoch zunächst das Potenzial von R zur Report-Generierung vorgestellt werden. Denn im Grunde sind beide Aufgaben bis zu einem gewissen Grade sehr ähnlich. Zunächst werden Daten von verschiedenen Quellen abgefragt, dann aufbereitet und schließlich visualisiert. Bei Analysen ist dieser Prozess jedoch iterativ, ohne dass im Vorhinein klar ist, wie das Ergebnis resp. die Erkenntnis genau aussieht. Bei der Generierung von Reports hingegen ist der Ablauf des Datenprozesses vorgegeben. Sie sind daher kleine, standardisierte Analysen, die an einem bestimmten Punkt der Visualisierung aufhören. Das Ergebnisdokument ist somit immer gleich und jederzeit reproduzierbar. Daher eignen sie sich sehr gut, um exemplarisch die Möglichkeiten von R aufzuzeigen, und sind dabei aber gleichzeitig so vereinheitlicht, um durch den Leser selbst nachgebaut zu werden. Ich möchte Sie daher gerne dazu auffordern, das folgende Skript selbst in RStudio umzusetzen.

Hinweis

Der Autor ist sich bewusst, dass gerade für Einsteiger der Anfang etwas holprig sein kann. Aber auch bei erfahreneren Lesern können durchaus Fragen aufkommen. Um hier Hilfestellung zu bieten, die Antworten der Community zugänglich zu machen und den Austausch anzuregen, möchte er Sie ermutigen, in der extra hierfür gegründeten Facebook-Gruppe nachzufragen. Es gibt dabei keine „dummen“ Fragen. Vielmehr hätte der Autor sich in seiner Anfangszeit selbst ein solches Forum mit dem Fokus auf analytische Programmierung und Online-Marketing sehr gewünscht. www.facebook.com/groups/ompyr/

Aber genug der langen Vorrede! In diesem Teil der Serie ist das Ziel, die Sessions aus GA sowie die Performance-Daten aus der GSC abzufragen, um die Packages für diese APIs besser kennenzulernen. Darüber hinaus werden Sie am Beispiel von Sistrix eine eigene API-Abfrage schreiben, um den Sichtbarkeitsindex für eine Domain zu ermitteln. Als Beispiel-Domain dient hier die Website `dein-trueffel.de`, ein Studentenprojekt an der Hochschule Darmstadt im Modul „Fortgeschrittene Suchmaschinenoptimierung“, das durch Jens Fauldrath und Stefan Keil betreut wird. Voraussetzung zum Nachvollziehen des Skripts sind eine aktuelle R-Installation (www.r-project.org) und RStudio (www.rstudio.com).

Anlegen eines neuen Projekts in RStudio

Zunächst wird ein neues Projekt angelegt (File → New Project ... → New Directory → New Project; Abb. 1). Benennung und Speicherort können Sie frei wählen. Der Einfachheit halber wird für diesen Beitrag das Projekt „report“ (Directory name) unter `C:\Users\{USER}\Documents\` (im Eingabedialog durch ~ automatisch abgekürzt; Abb. 2) erzeugt.

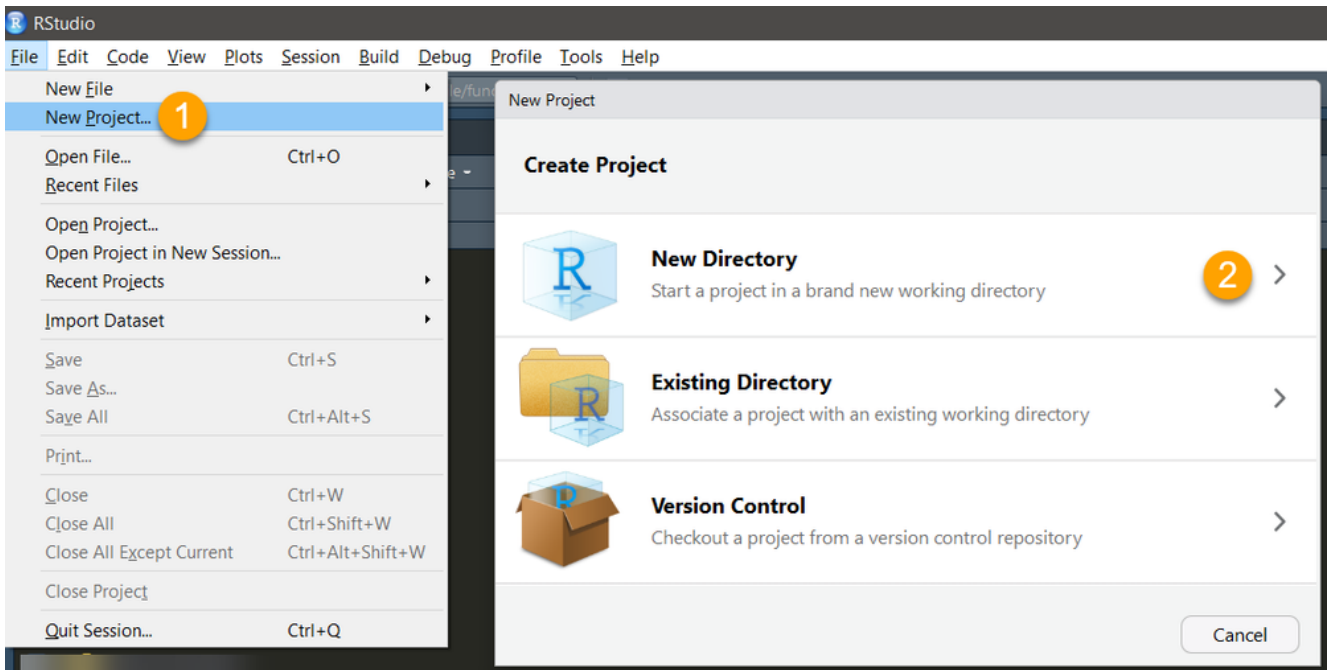


Abbildung 1: Anlegen eines neuen Projekts

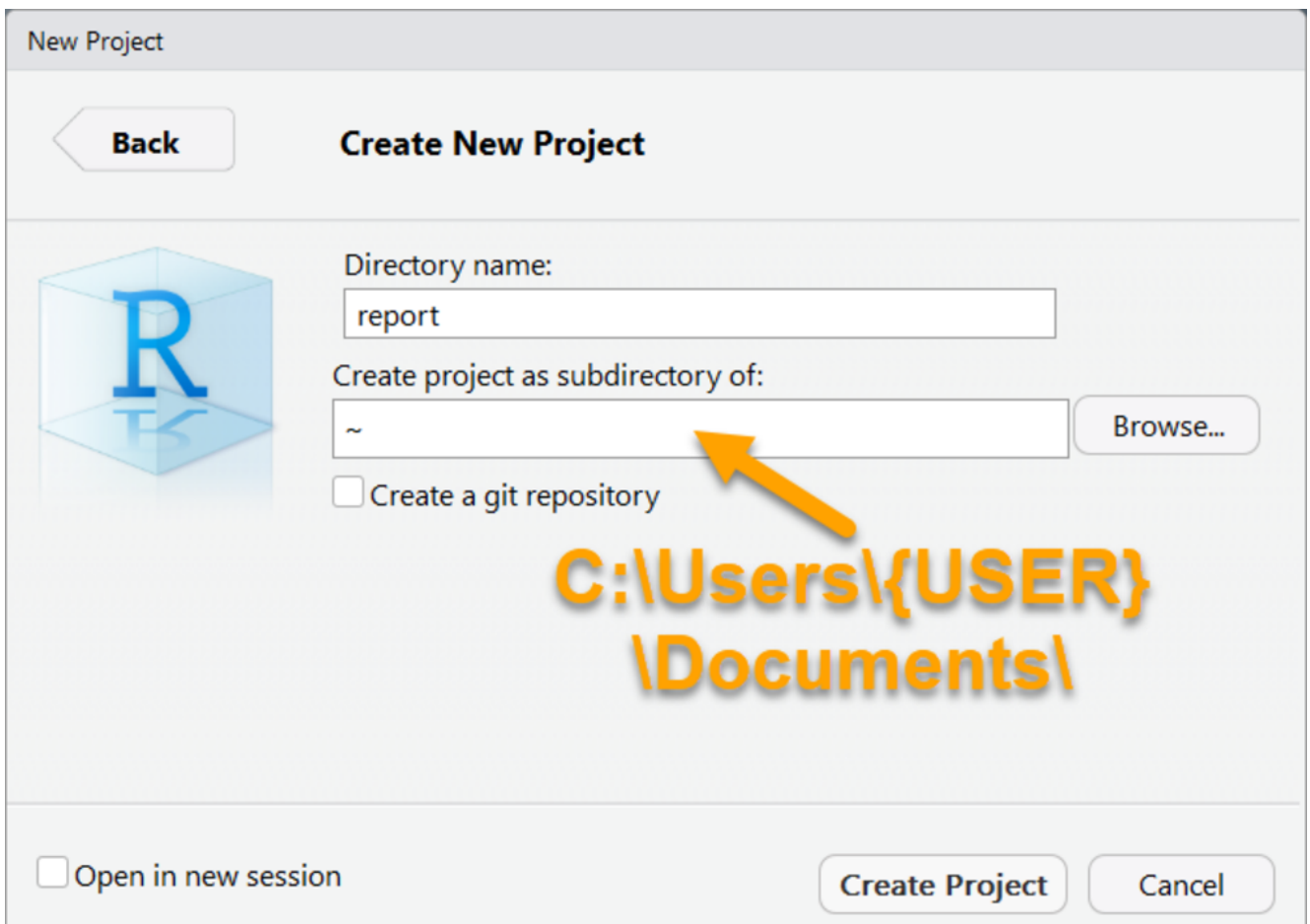


Abbildung 2: Speicherort und Benennung des Projekts

Das Projekt ist zunächst leer, im File-Explorer finden Sie nur eine RStudio-Projektdatei (report.Rproj), die ignoriert werden kann. Daher wird als erstes über File → New File → R Script (oder den darunter befindlichen Button) eine neue Datei

erzeugt und unter der Benennung `api_calls.R` gespeichert.

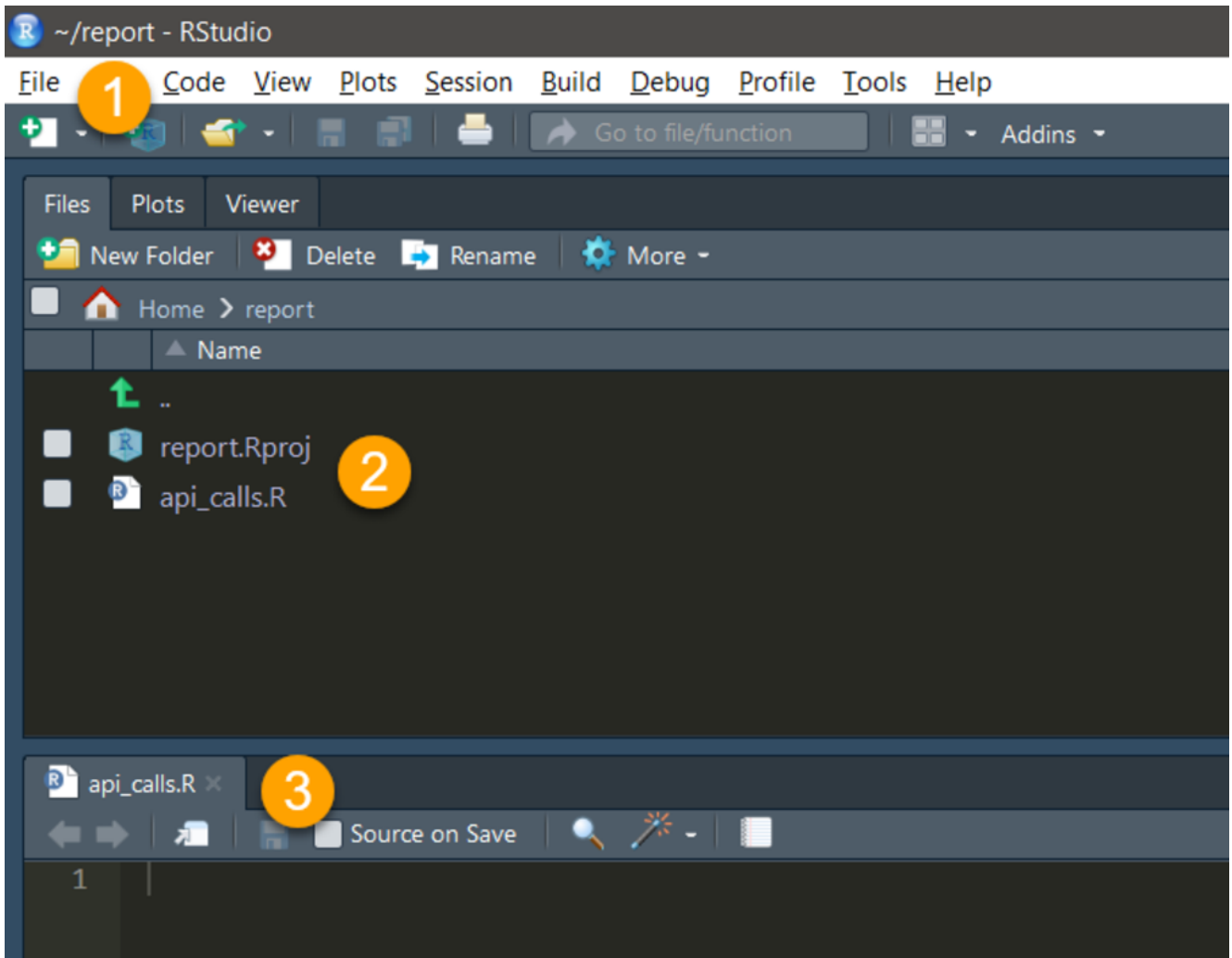
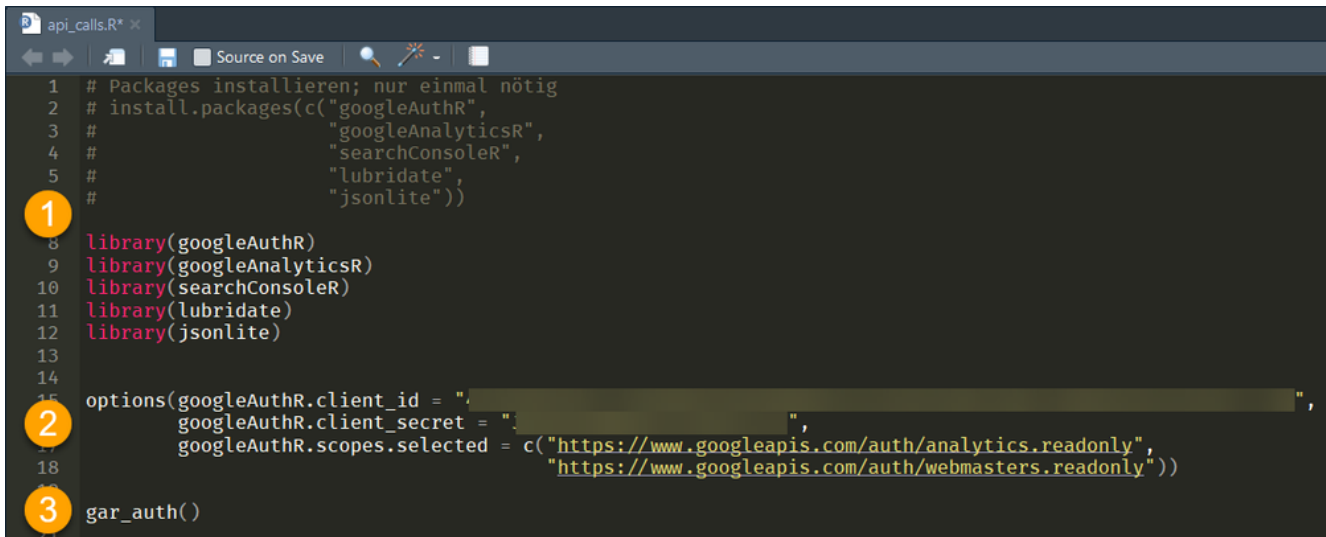


Abbildung 3: (1) Neues Skript erstellen, (2) Projekt-Datei und Skript im File-Explorer, (3) Skript-Editor zeigt die erstellte Datei an

Installation und Laden der benötigten Packages

In der soeben erstellen `api_calls.R`-Datei werden nun die benötigten Packages installiert bzw. – falls bereits vorhanden – aktualisiert und geladen (Abb. 4; 1). Die Google-Packages kennen Sie bereits aus dem Beitrag von Tobias Aubele. Das Package `lubridate` erweitert den Funktionsumfang von R für die komfortable Arbeit mit Datums- und Zeitwerten, `jsonlite` dient zum Abfragen und Verarbeiten von JSON-Dateien. Sobald Funktionen aus diesen Packages verwendet werden, wird explizit

darauf hingewiesen. Die vorangestellten Rauten kommentieren den Code aus. Beim ersten Ausführen müssen Sie diese also entfernen. Markieren Sie dann den gesamten Code mit der Maus und führen Sie ihn durch Drücken der STRG- und ENTER-Tasten aus.



```
1 # Packages installieren; nur einmal nötig
2 # install.packages(c("googleAuthR",
3 #                   "googleAnalyticsR",
4 #                   "searchConsoleR",
5 #                   "lubridate",
6 #                   "jsonlite"))
7
8 library(googleAuthR)
9 library(googleAnalyticsR)
10 library(searchConsoleR)
11 library(lubridate)
12 library(jsonlite)
13
14
15 options(googleAuthR.client_id = " ",
16         googleAuthR.client_secret = " ",
17         googleAuthR.scopes.selected = c("https://www.googleapis.com/auth/analytics.readonly",
18                                         "https://www.googleapis.com/auth/webmasters.readonly"))
19
20 gar_auth()
```

Abbildung 4: Installieren bzw. Aktualisieren und Laden der benötigten Packages

In der anschließenden `options()`-Funktion definieren Sie Ihre Google-Client-ID sowie Ihr -Secret (2). Diese können Sie sich unter `console.developers.google.com` generieren. Das Argument `scopes` definiert die Berechtigungen für die GA- bzw. GSC-API. `gar_auth()` dient schließlich der Authentifizierung (3). Führen Sie den Code das erste Mal aus, öffnet sich Ihr Browser, in dem Sie den Zugriff des Skripts bestätigen müssen. Ist dies erfolgt, finden Sie in Ihrem File-Explorer eine Authentifizierungsdatei (`sc.oauth`). Bei einem erneuten Ausführen des Skripts verwendet `gar_auth()` diese Datei, sodass Sie nicht erneut eine manuelle Authentifizierung durchführen müssen – praktisch, denn zukünftig soll das Skript ja automatisch laufen.

Konfiguration wichtiger Variablen

Nun folgt ein Code-Abschnitt, in dem einige Variablen definiert werden (Abb. 5), die an verschiedenen Stellen des Skripts zum Einsatz kommen. Dass die Variablen hier

großgeschrieben werden, ist eine Konvention des Autors, die anzeigt, dass die Variablen erst zu einem späteren Zeitpunkt genutzt werden – grundsätzlich könnten Sie die Variablen auch kleinschreiben. Dies ist somit die Stelle im Skript, an der Sie Ihre eigenen Daten eintragen müssen.

```
22
23 # Konfiguration -----
24
25 # Google Analytics View-ID ermitteln und eintragen
26 # ga_account_list() %>% View()
27 GA_VIEW_ID ← " "
28
29 # GSC Property ermitteln und eintragen
30 # list_websites() %>% View()
31 GSC_PROP ← "https://dein-trueffel.de/"
32
33 # Sisitrix
34 SISTRIX_DOMAIN ← "dein-trueffel.de"
35 SISTRIX_API_KEY ← " "
36
37 # Start- und Endzeitpunkt der Datenabfragen definieren
38 # Hier: Sechs Monate rückwirkend vom Beginn des aktuellen Monats bis zum
39 # aktuellen Datum
40 START_DATE ← floor_date(today(), "month") %m-% months(6)
41 END_DATE ← today()
42
```

Abbildung 5: Definition globaler Variablen, für die die Daten von den APIs abgefragt werden sollen

Im Detail:

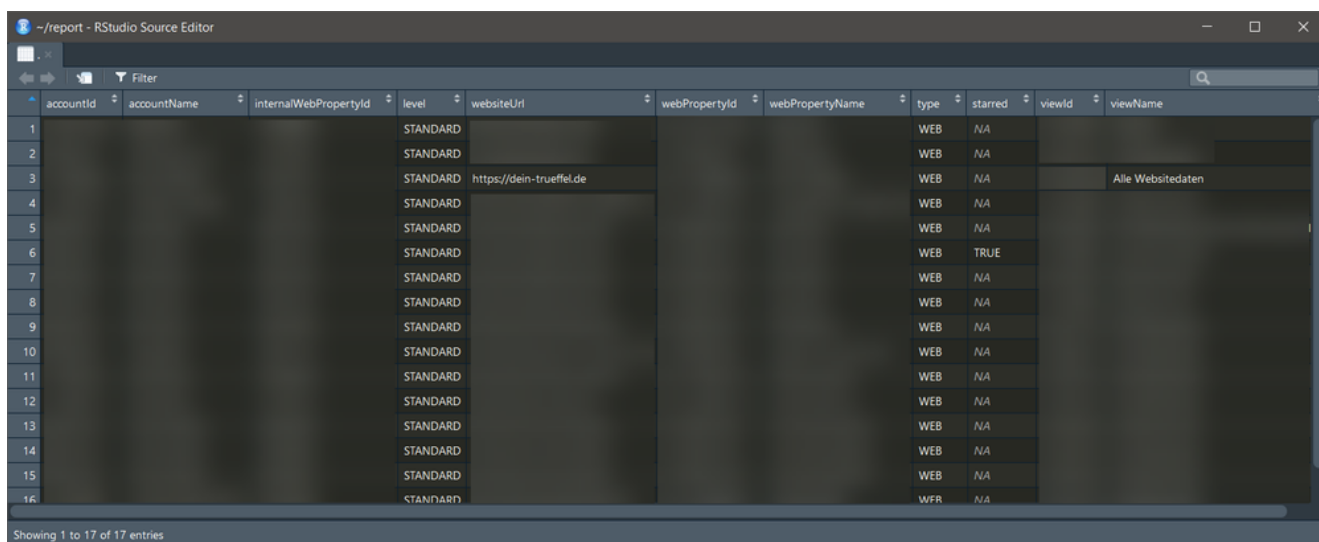
(1) Um die passende View-ID Ihrer GA-Datensicht zu ermitteln, können Sie die auskommentierte Funktion `ga_account_list() %>% View()` verwenden. Der Pipe-Operator `%>%` übergibt den von der API zurückgegebenen DataFrame mit den Account-Details an die Funktion `View()`, welche die Tabelle in einem Explorer statt auf der Console öffnet (Abb. 6). Falls der letzte Satz nur ein böhmisches Dorf für Sie war, seien Sie beruhigt, eine ausführlichere Erklärung folgt im nächsten Artikel. Wichtig ist erst einmal nur, dass Sie Ihre View-ID definieren. Entnehmen Sie dazu die gewünschte View-ID aus der Spalte `viewId` und tragen Sie sie bei der Variable `GA_VIEW_ID` ein.

(2) Analog dazu verfahren Sie mit der `siteUrl` für die Variable `GSC_PROP`.

(3) Die Sistrix-Daten sollen später auf Domain-Ebene abgefragt werden, entsprechend tragen Sie bei SISTRIX_DOMAIN diese ein. Ihren Sistrix-API-Key für die Variable SISTRIX_APY_KEY finden Sie unter app.sistrix.com/account/api.

(4) START_DATE und END_DATE definieren den abzufragenden Berichtszeitraum. In der aktuellen Konfiguration umfasst dieser sechs Monate rückwirkend vom Beginn des aktuellen Monats bis zum heutigen Datum. `floor_date(today(), „month“) %m-% months(6)` bedient sich dreier Funktionen aus dem Package `lubridate`. `today()` gibt das heutige Datum zurück („2018-01-11“). `floor_date(„2018-01-11“), „month“)` rundet dann zum Beginn des aktuellen Monats („2018-01-01“). `%m-% months(6)` subtrahiert von diesem Datum sechs Monate („2018-07-01“). `%m-%` ist ein besonderer Operator aus dem Package, der das Subtrahieren von Monaten unter Berücksichtigung unterschiedlicher Tageszahlen im Monat und Schaltjahren unterstützt. Sie landen mit dieser Funktion also immer am ersten eines Monats.

Die definierten Variablen können Sie nun im Environment sehen (Abb. 7).



accountId	accountName	internalWebPropertyId	level	websiteUrl	webPropertyId	webPropertyName	type	starred	viewId	viewName
1			STANDARD				WEB	NA		
2			STANDARD				WEB	NA		
3			STANDARD	https://dein-trueffel.de			WEB	NA		Alle Websitedaten
4			STANDARD				WEB	NA		
5			STANDARD				WEB	NA		
6			STANDARD				WEB	TRUE		
7			STANDARD				WEB	NA		
8			STANDARD				WEB	NA		
9			STANDARD				WEB	NA		
10			STANDARD				WEB	NA		
11			STANDARD				WEB	NA		
12			STANDARD				WEB	NA		
13			STANDARD				WEB	NA		
14			STANDARD				WEB	NA		
15			STANDARD				WEB	NA		
16			STANDARD				WFR	NA		

Abbildung 6: Data-Explorer mit den GA-Accounts

Values	
END_DATE	2019-01-11
GA_VIEW_ID	"[REDACTED]"
GSC_PROP	"https://dein-trueffel.de/"
SISTRIX_API_KEY	"[REDACTED]"
SISTRIX_DOMAIN	"dein-trueffel.de"
START_DATE	2018-07-01

Abbildung 7: Environment oder Variable-Explorer listet alle zugewiesenen Variablen

GA-Sessions abfragen

Wenn die Konfiguration abgeschlossen ist, kann mit der ersten API-Abfrage begonnen werden. Als Erstes sollen die Sessions, die aus den Channels Organic Search resp. Direct resultieren, aus GA abgerufen werden (Abb. 8). Um die Abfrage auf diese Dimensionen einzuschränken, werden mittels `dim_filter()` zwei Filter-Komponenten konfiguriert und in der Variablen `channel_filter_organic` bzw. `channel_filter_direct` gespeichert (1). Diese werden an eine weitere Funktion `filter_clause_ga4()` gegeben, welche intern die Filter in eine für die API verständliche Syntax übersetzt (2). Die Funktion `google_analytics()` führt schließlich die eigentliche Abfrage aus (3). Hier sehen Sie auch die Variablen für die GA-Datensicht sowie das Start- und Enddatum, die Sie zuvor konfiguriert haben. Führen Sie den Code aus, erscheint in Ihrem Environment die Variable `ga_sessions` (Abb. 9; 1), die Sie mit der Maus anklicken können und die einen DataFrame mit der API-Antwort enthält (2).

```

43
44 # Google Analytics -----
45
46 # Organic Search / Direct Sessions abfragen
47
48 channel_filter_organic ← dim_filter(dimension = "channelGrouping",
49                                     operator = "EXACT",
50                                     expressions = "Organic Search")
51
52 channel_filter_direct ← dim_filter(dimension = "channelGrouping",
53                                    operator = "EXACT",
54                                    expressions = "Direct")
55
56 filter_clause ← filter_clause_ga4(list(channel_filter_organic,
57                                       channel_filter_direct),
58                                   operator = "OR")
59
60 ga_sessions ← google_analytics(viewId = GA_VIEW_ID,
61                                date_range = c(START_DATE, END_DATE),
62                                metrics = c("sessions"),
63                                dimensions = c("date", "channelGrouping"),
64                                dim_filters = filter_clause,
65                                anti_sample = TRUE)
66

```

Abbildung 8: API-Abfrage der GA-Sessions

The screenshot shows the RStudio interface. In the Environment pane, the variable 'ga_sessions' is highlighted with a blue circle containing the number '1'. Below it, the 'Values' pane shows a preview of the data table with columns 'date', 'channelGrouping', and 'sessions'. A blue circle with the number '2' is placed over the 'ga_sessions' variable name in the Values pane.

	date	channelGrouping	sessions
1	2018-07-01	Direct	1
2	2018-07-02	Organic Search	2
3	2018-07-03	Organic Search	1
4	2018-07-04	Direct	1
5	2018-07-05	Direct	1
6	2018-07-05	Organic Search	1
7	2018-07-06	Direct	1
8	2018-07-06	Organic Search	1

Abbildung 9: DataFrame mit den GA-Sessions

GSC-Performance-Daten abfragen

Ähnlich funktioniert die anschließende Abfrage der GSC-API (Abb. 10). Von dieser werden verschiedenen Dimensions-Kombinationen abgefragt. (1) Die erste Abfrage ruft die Clicks, Impressions, CTR und Position für den searchType „web“ auf Tagesbasis ab. Auch hier sehen Sie die Verwendung der anfangs definierten Variablen. (2) Dann erfolgt die Abfrage der Metriken für die Kombination aus Datum und Suchanfrage sowie schließlich (3) für die Kombination aus Datum und URL. Auch hier finden Sie nach dem Ausführen der API-Calls die Antwort-Tabellen in Ihrem Environment.

```
67
68 # GSC -----
69
70 # GSA: Date
71
72 1 gsa_dim_date ← search_analytics(siteURL = GSC_PROP,
73                                 startDate = START_DATE,
74                                 endDate = END_DATE,
75                                 searchType = "web",
76                                 dimensions = "date")
77
78 # GSA: Date ~ Query
79
80 2 gsa_dim_date_query ← search_analytics(siteURL = GSC_PROP,
81                                       startDate = START_DATE,
82                                       endDate = END_DATE,
83                                       searchType = "web",
84                                       dimensions = c("date", "query"))
85
86 # GSA: Date ~ Page
87
88 3 gsa_dim_date_page ← search_analytics(siteURL = GSC_PROP,
89                                       startDate = START_DATE,
90                                       endDate = END_DATE,
91                                       searchType = "web",
92                                       dimensions = c("date", "page"))
93
```

Abbildung 10: API-Abfrage der GSC-Metriken für verschiedene Dimensionskombinationen

Sistrix-Sichtbarkeitsindex abfragen

Zu guter Letzt müssen Sie noch selbst eine API-Abfrage schreiben, denn es gibt natürlich nicht für alle Datenquellen fertige Packages. Sistrix ist so ein Fall, der aber aufgrund

der sehr einfachen Abfrage-Syntax ohne großen Aufwand umgesetzt werden kann. Die Abfrage ist nämlich nur eine URL, deren Query-Parameter die gewünschten Abfragewerte aufnehmen. Wird die URL aufgerufen, erhalten Sie als Antwort eine JSON-Datei. Diese URL bauen Sie mit der Funktion `paste0()` zusammen (Abb. 11; 1). Konkret verbindet die Funktion die in Klammern angegebenen Argumente. Diese bestehen aus den Bestandteilen der Basis-URL für die Abfrage des historischen Sichtbarkeitsindex (SI), in die die zuvor definierten Variablen für Ihren API-Key sowie die abzufragende Domain eingefügt werden.

Anschließend fragen Sie die zusammengebauten URLs gegen die API an (2). Dazu verwenden Sie die Funktion `fromJSON()` aus dem Package `jsonlite`. Die Funktion führt einen GET-Request aus, nimmt die JSON-Antwort der API entgegen und überführt sie automatisch in eine Liste – ein R-Datentyp zum Speichern hierarchischer Daten. Die Antwort wird zunächst in die jeweiligen Variablen geschrieben. Klicken Sie eine der beiden Variablen im Environment an, sehen Sie, dass das JSON in eine Listen-Struktur überführt wurde (Abb. 12). Hier sehen Sie auch schon den „Weg“, den Sie innerhalb der Liste gehen müssen, um an die SI-Daten heranzukommen. Denn neben diesen erhalten Sie noch weitere Meta-Informationen, wie bspw. die verbrauchten Credits, zurück. Um die Tabelle mit den SI-Daten zu extrahieren, müssen Sie sich an den übergeordneten Listenpunkten entlanghangeln. Dies macht der Ausdruck `api_response_desktop$answer$sichtbarkeitsindex[[1]]` für den Desktop-SI und überführt den DataFrame in die Variable `si_desktop` (Abb. 11; 3).

```

94
95 # Sistrix -----
96
97 # API-URLs zusammenbauen
98 si_url_desktop ← paste0("https://api.sistrix.com/domain.sichtbarkeitsindex?history=true&format=json&api_key=",
99                          SISTRIX_API_KEY,
100                          "&domain=",
101                          SISTRIX_DOMAIN)
102
103 si_url_mobile ← paste0("https://api.sistrix.com/domain.sichtbarkeitsindex?history=true&format=json&api_key=",
104                        SISTRIX_API_KEY,
105                        "&domain=",
106                        SISTRIX_DOMAIN,
107                        "&mobile=true")
108
109 # API abfragen
110 api_response_desktop ← fromJSON(si_url_desktop)
111 api_response_mobile ← fromJSON(si_url_mobile)
112
113 # JSON-Antwort in DataFrame überführen
114 si_desktop ← api_response_desktop$answer$sichtbarkeitsindex[[1]]
115 si_mobile ← api_response_mobile$answer$sichtbarkeitsindex[[1]]
116

```

Abbildung 11: Konstruktion der Sistrix-API-Abfrage

The screenshot shows the RStudio environment pane for the object 'api_response_desktop'. The object is a list of length 3. The first element is a character vector of length 1 containing 'domain.sichtbarkeitsindex'. The second element is a data frame with 1 row and 1 column, containing a list of length 1. This list contains a data frame with 566 rows and 3 columns. The columns are 'domain', 'date', and 'value'. The 'domain' column contains the string 'dein-trueffel.de' repeated 566 times. The 'date' column contains ISO timestamps for 2019-01-14 and 2019-01-07. The 'value' column contains numerical values ranging from 0.0015 to 0.0002. The third element of the list is a data frame with 1 row and 1 column containing the integer value 566.

Name	Type	Value
api_response_desktop	list [3]	List of length 3
method	character [1 x 1]	'domain.sichtbarkeitsindex'
answer	list [1 x 1] (S3: data.frame)	A data.frame with 1 rows and 1 columns
sichtbarkeitsindex	list [1]	List of length 1
[[1]]	list [566 x 3] (S3: data.frame)	A data.frame with 566 rows and 3 columns
domain	character [566]	'dein-trueffel.de' 'dein-trueffel.de' 'dein-truef...
date	character [566]	'2019-01-14T00:00:00+01:00' '2019-01-07...
value	double [566]	0.0015 0.0002 0.0002 0.0002 0.0000 0.0002 ...
credits	list [1 x 1] (S3: data.frame)	A data.frame with 1 rows and 1 columns
used	integer [1]	566

Abbildung 12: Liste mit der Antwort der Sistrix-API

Fazit und Ausblick

Damit haben Sie das Ziel dieses Beitrags erreicht. Sie haben ein reproduzierbares Skript geschrieben, dass Sie jederzeit erneut ausführen können. Der Berichtszeitraum der abzufragenden Daten passt sich bei jeder erneuten Ausführung automatisch an. Als Resultat liegen Ihnen die GA-, GSC- und Sistrix-Daten in tabellarischer Form vor. Im nächsten Artikel dieser Reihe wird es dann um die Aufbereitung dieser Daten gehen. Aktuell liegen die Daten auf Tagesbasis vor. Für den

Report sollen bspw. die GSC-Daten auf Query- resp. URL-Basis aggregiert werden, um somit Tabellen der Top-10-Suchphrasen bzw. URLs der Vorwoche zu erstellen.