

Quantitative Textanalyse

Sitzung 5: Datenerschließung – Scraping I

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

06. November 2024

Was ist Web-Scraping?



Web-Scraping beschreibt den Prozess des systematischen Sammelns von (oftmals unstrukturierten) Daten aus dem Internet, um sie in einem strukturierten Datenformat zu speichern.

Und was bringt das? I

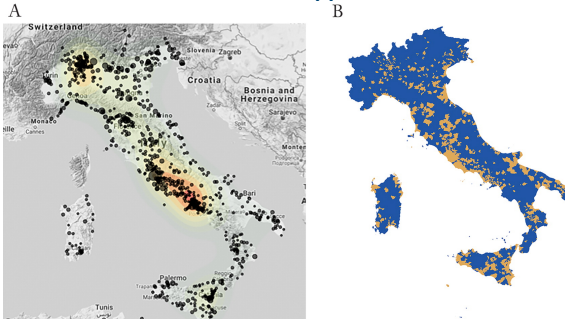


Figure 1.1 Location of UNESCO World Heritage Sites in danger (as of March 2014). Cultural sites are marked with triangles, natural sites with dots

Figure: Überblick über das Weltkulturerbe in Gefahr (Munzert 2015, p. 5)

Aus einer [Wikipedia-Tabelle](#) wird eine übersichtliche Visualisierung

Und was bringt das? II



Bischof and Kurer (2023) erschließen Daten von Kampagnenevents des italienischen Five Star Movements, um den Effekt der Mobilisierung auf ein Referendum zu messen.

Und was bringt das? III

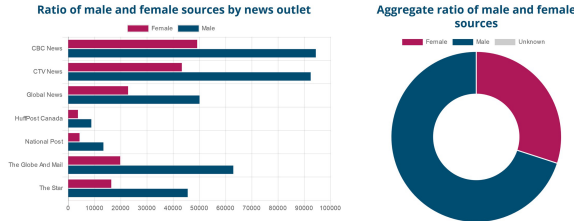


Fig 1. The Gender Gap Tracker online dashboard page. Reprinted from <https://gendergaptracker.informedopinions.org/> under a CC BY license, with permission from Informed Opinions, original copyright 2018.

Asr et al. (2021) laden Daten von kanadischen Zeitungsartikeln herunter, um zu analysieren, wie oft diese auf männliche/weibliche Expert*innen verweisen (nach wie vor online!)

Und was bringt das? IV

Web-Scraping oft als erster Schritt für eine anschließende Analyse,
nicht notwendigerweise einer Textanalyse.

Verschiedene Arten von Web-Scraping

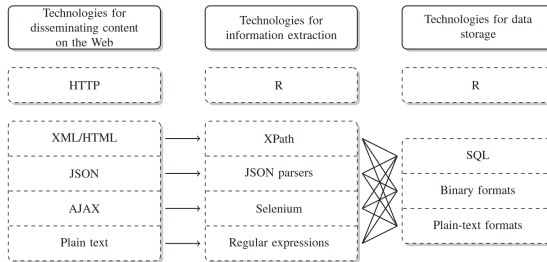


Figure 1.4 Technologies for disseminating, extracting, and storing web data

Munzert (2015, p. 10)

Unsere Ziele

1. **‘statische’ HTML-Strukturen** nutzbar machen (rvest)
 - Der Inhalt der uns interessiert ist im HTML-Quellcode einsehbar
2. **‘dynamische’ Webseiten** herunterladen (RSelenium)
 - Der Inhalt ist nicht (sofort) im HTML-Quellcode einsehbar, sondern bedarf Interaktion (bspw. über Klicken mit der Maus) von User*innen

Bevor ihr Daten scraped...

1. Benötigt ihr die Daten überhaupt?
 - Könnt ihr damit etwas messen, was andernfalls nicht ginge?
 - Entsprechen die Daten gewissen qualitativen Standards?
 - Ist es die Arbeit wirklich wert?
2. Sind die Daten schon andernorts verfügbar? Beispielsweise über direkten Datendownload oder eine API?
3. Gibt es rechtliche Bedenken?

Disclaimer: Rechtliche Bedenken I

Ist Web-Scraping erlaubt?

Kommt drauf an...

- Verletzt ihr damit die Nutzungsbedingungen von Webseiten-Anbieter*innen?

Disclaimer: Rechtliche Bedenken II

Without limitation, you shall not use (and you shall also not facilitate, authorise or permit the use of) the Guardian Site and/or any Guardian Content (including, any caption information, keywords or other associated metadata) for any other purpose without our prior written approval - this includes, without limitation, that you shall not use, copy, scrape, reproduce, alter, modify, collect, mine and/or extract the Guardian Content: (a) for any machine learning, machine learning language models and/or artificial intelligence-related purposes (including the training or development of such technologies); (b) for any text and data aggregation, analysis or mining purposes (including to generate any patterns, trends or correlations); or (c) with any machine learning and/or artificial intelligence technologies to generate any data or content or to synthesise or combine with any other data or content; or (d) for any commercial use.

Figure: Nutzungsbedingungen der Zeitung 'The Guardian'

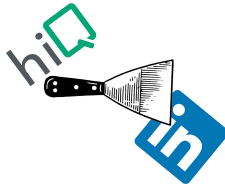
Disclaimer: Rechtliche Bedenken III

Ist Web-Scraping erlaubt?

Es kommt drauf an...

- Verstoß gegen die Nutzungsbedingungen?
- Das Sammeln personenbezogener Daten kann gegen die EU-[Datenschutz-Grundverordnung \(DSGVO\)](#) verstoßen
- Auch wenn keine personenbezogenen Daten erfasst werden, könnten Urheberrechtsvorschriften verletzt werden
- Ihr solltet immer die [robots.txt](#) überprüfen

Disclaimer: Rechtliche Bedenken IV



- hiQ sammelte Daten von öffentlichen LinkedIn-Profilen
- LinkedIn versuchte, dies zu verhindern
- Erste Instanz: Scraping erlaubt; spätere Instanzen: hiQ verletzte Nutzervereinbarung → außergerichtliche Einigung

Disclaimer: Rechtliche Bedenken V

Letztlich: “[I]egalities depend a lot on where you live. However, as a general principle, if the data is public, non-personal, and factual, you’re likely to be ok” (Wickham et al. 2023, June)

HTML-Grundlagen

Webseiten basieren auf der **H**yper**T**ext **M**arkup-**S**prache (HTML)

- HTML enthält Informationen über die Struktur einer Webseite
- HTML ist dafür verantwortlich, wie Inhalte grafisch dargestellt werden

Ein Beispiel

HTML-Elemente und Attribute

- HTML besteht aus Elementen, Tags und Attributen
 - Elemente sind die verschiedenen Komponenten einer Webseite (z.B. Überschriften, Text, Bilder)
 - Elemente sind meist in Tags eingebettet (`<element>Inhalt</element>`), einige jedoch ohne Anfangs- und End-Tags
 - Attribute sind zusätzliche Informationen zu einem Element (z.B. Bildgröße, Schriftart usw.)

Wir behandeln HTML nur sehr oberflächlich; falls ihr euch etwas vertiefen wollt, probiert dieses **Tutorial** aus.

HTML: head vs. body

```
1  <!DOCTYPE html>
2  <html lang="de" class="no-js">
3  <head>
4  <title>Universitaet Muenster</title>
5  </head>
6  <body>
7  ...
8  </body>
9  </html>
```

HTML-Dokumente bestehen aus einem **header** (die *Kopfnote*, mit Meta-Informationen zur Webseite) und einem **body** (mit Inhalten)
→ wir sind hauptsächlich am *body* interessiert!

Wiederkehrende Elemente

- *h1*, *h2*, *h3* usw.: Überschriften
- *p*: Absätze
- *a*: Hyperlinks
- *img*: Bilder

Wiederkehrende Attribute

- *href*: Weblink, kommt immer mit dem *a*-Element
- *src*: Quelle eines Bildes

HTML und Scraping

Wir müssen den CSS-Selektor des gewünschten Elements identifizieren.

Hierfür gibt es **zwei** Optionen:

- Manuelle Methode: Maus auf gewünschtes Element > Rechtsklick > Inspizieren
- (semi-)automatische Methode: **SelectorGadget** herunterladen oder als Lesezeichen speichern

Grundpipeline

Einrichtung

- SelectorGadget installieren
- R-Bibliothek: `rvest`

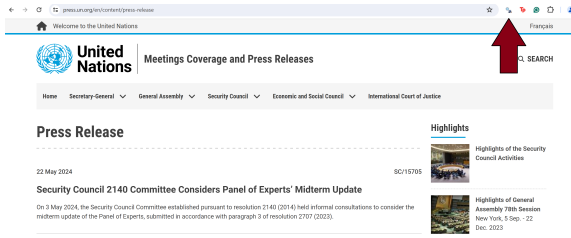
Schritt 1

HTML-Quelle mit rvest herunterladen

```
1 > library(rvest)
2 > url <-
  "https://press.un.org/en/content/press-release"
3 > html <- read_html(url)
4 > html
5 {html_document}
6 <html lang="en" dir="ltr">
7 [1] <head>\n<meta http-equiv="Content-Type"
  content="text/html; charset=UTF-8">\n<meta
  charset="utf-8">\n<link rel="canonical"
  href="https: ...
8 [2] <body class="layout-one-sidebar
  layout-sidebar-first page-view-home-press
  path-content">\n      <div
  class="visually-hidden-focusable bg- ..
```

Schritt 2

Beispiel: UN-Pressemitteilungen



The screenshot shows the United Nations website's 'Meetings Coverage and Press Releases' section. A red arrow points to the 'Français' link in the top right corner. The page features a search bar, a navigation menu with links to Home, Secretary-General, General Assembly, Security Council, Economic and Social Council, and International Court of Justice, and a 'Press Release' section dated 22 May 2024. The main headline is 'Security Council 2140 Committee Considers Panel of Experts' Midterm Update'. To the right, there is a 'Highlights' section with two items: 'Highlights of the Security Council Activities' and 'Highlights of General Assembly 78th Session'.

press.un.org/en/content/press-release

Welcome to the United Nations

United Nations

Meetings Coverage and Press Releases

Home Secretary-General General Assembly Security Council Economic and Social Council International Court of Justice

Press Release

22 May 2024

SC/15705

Security Council 2140 Committee Considers Panel of Experts' Midterm Update

On 3 May 2024, the Security Council Committee established pursuant to resolution 2140 (2014) held informal consultations to consider the midterm update of the Panel of Experts, submitted in accordance with paragraph 3 of resolution 2707 (2023).

Highlights

Highlights of the Security Council Activities

Highlights of General Assembly 78th Session
New York, 5 Sep. - 22 Dec. 2023

Schritt 3



United Nations | Meetings Coverage and Press Releases

Home Secretary-General General Assembly Security Council Economic and Social Council International Court of Justice

Press Release

22 May 2024

Security Council 2140 Committee Considers Panel of Experts' Midterm Update

On 3 May 2024, the Security Council Committee established pursuant to resolution 2140 (2014) held informal mid-term update of the Panel of Experts, submitted in accordance with paragraph 3 of resolution 2707 (2023).

21 May

Activities of the Secretary-General in Bahrain, 15-17 May

United Nations Secretary-General António Guterres flew from Muscat, Oman, to Manama, Bahrain, to meet with the Bahraini leadership and the United Nations Secretary-General's Special Representative for the Middle East.

Highlights

Highlights of the Security Council Activities

SC/15705

Highlights of General Assembly 78th Session New York, 5 Sep. - 22 Dec. 2023

68th Session of the Commission on the Status of Women

Clear (1) Toggle Position XPath ? X

Probiert es selbst aus!

Und jetzt in R

Überschrift

Hier rufen wir jede Level-1-Überschrift der Webseite ab.

```
1 > library(rvest)
2 > (top_level_headline <- read_html(url)
3 +   %>% html_elements("h1")
4 +   %>% html_text())
5 [1] "Pressemitteilung"
```

Text

Hier rufen wir jeden Absatz auf der Webseite ab.

```
1 > library(rvest)
2 > (paragraphs <- read_html(url) %>%
3 +   html_elements("p") %>%
4 +   html_text())
5 [1] "On 3 May 2024, the Security Council Committee
    established pursuant to resolution 2140 (2014)
    held informal consultations to consider the
    midterm update of the Panel of Experts, submitted
    in accordance with paragraph 3 of resolution 2707
    (2023)."
```

```
6 [2] "United Nations Secretary-General Antonio
    Guterres flew from Muscat, Oman, to Manama,
    Bahrain, in the early evening of Wednesday, 15
    May."
```

Links

Falls wir auf Links zugreifen möchten, müssen wir zuerst das a-Element abrufen und dann dessen Attribut href aufrufen.

```
1 > (pr_urls <- read_html(url) %>%  
2 +   html_elements(".field__item a") %>%  
3 +   html_attr("href"))  
4 [1] "/en/2024/sc15705.doc.htm"  
    "/en/2024/sgt3388.doc.htm"  
    "/en/2024/sgt3387.doc.htm"  
    "/en/2024/3386.doc.htm"
```

Tabellen

rvest hat eine vordefinierte Funktion `html_table()`, um Informationen aus HTML-Tabellen zu extrahieren.

```
1 > html <- read_html(url2)
2 > table <- html %>%
3 +   html_element(".wikitable:nth-child(4)") %>%
4 +   html_table()
```

Bilder

Bei Bildern ist es etwas komplizierter.

1. eine Sitzung öffnen
2. Bildquelle abrufen
3. das Bild in Ihr Verzeichnis herunterladen

Bilder II

```
1      > session <- session(url)
2      >
3      > # save links to image source
4      > imgsrc <- session %>%
5      +   read_html() %>%
6      +   html_nodes("img") %>%
7      +   html_attr("src")
8      >
9      > # open image source
10     > img <- session_jump_to(session,
11                               paste0(root_url, imgsrc[[1]]))
12     >
13     > # write into project directory
14     > writeBin(img$response$content,
15               basename(imgsrc[1]))
```


Loops I

Häufig möchten wir diese Schritte für mehrere Seiten automatisieren.

Zwei Optionen:

1. Leere Objekte erstellen und sie in einem `for`-Loop füllen
2. Eine Funktion definieren, sie anwenden und die gewünschten Objekte aus einer Liste abrufen

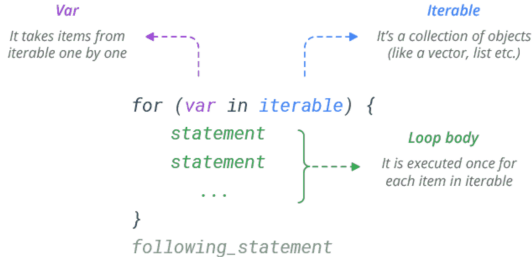
→ Meistens sind Funktionen vielseitiger und können leichter parallel ausgeführt werden.

Loops II

Vor der Erstellung von Loops/Funktionen

1. die Struktur der Seitenpaginierung prüfen (z. B. verwendet die [UN](#) "?page=#" zur Anzeige der Ergebnisse)
2. prüfen, welche Elemente abgerufen werden müssen (häufig benötigt man nur Links, aber manche Informationen, wie das Datum, sind möglicherweise nicht auf Unterseiten verfügbar und sollten daher ebenfalls gesammelt werden)
3. Pipeline an einem einzelnen Element testen, bevor wir sie in den Loop einbauen

For-Loops



Tutorial zu for-Loops

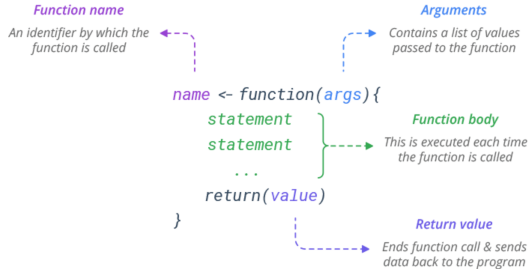
Grafik und Anleitung zu for-Schleifen

for-Loops für die Automatisierung

```
1
2 > urls <- c("https://www.uni-muenster.de/de/",
3             "https://www.uni-osnabrueck.de/startseite/")
4 > links <- c()
5 > for(i in 1:length(urls)){
6   +     html <- read_html(urls[[i]])
7   +     links[i] <- html %>%
8   +       html_node("h1") %>%
9   +       html_text()
10 > links
11 [1] "Universitaet Muenster" "Hauptinhalt"
```

Und jetzt in R

Funktionen in R



Tutorial zu Funktionen Graphik und Anleitung zu Funktionen

Funktionen für die Automatisierung

```
1 > h1_scrape <- function(url){  
2   +   html <- read_html(urls[[url]])  
3   +   links[url] <- html %>%  
4   +     html_node("h1") %>%  
5   +     html_text()  
6   + }  
7  
8 >  
9 > (links <- sapply(1:length(urls), h1_scrape))  
[1] "Universitaet Muenster" "Hauptinhalt"
```

Übungsfile in R (scraping_exercises_empty.Rmd)

Ausblick

- heute haben wir einfach strukturierte Webseiten heruntergeladen
- nächste Woche schauen wir uns Webseiten wie diese [hier](#) an
- außerdem: Einblick in die Datensammlung über APIs

Bis nächste Woche

- tragt euch im Learnweb in das Aufgabentool ein (max. 2 pro Sitzung)
- ggf. ladet eure eigene Webscraping-Anwendung in das Abgabetool hoch; bereitet euch für die nächste Woche vor, sodass ihr es der Gruppe vorstellen könnt
- ladet euch [Java](#) herunter und speichert die Installation als [Umgebungsvariable](#) ab
- ladet euch [RTools](#) herunter
- meldet euch für die [NYT-API](#) und die [Manifesto-Project-API](#) an

Literatur I

- Asr, F. T., Mazraeh, M., Lopes, A., Gautam, V., Gonzales, J., Rao, P., & Taboada, M. (2021). The gender gap tracker: Using natural language processing to measure gender bias in media. *PloS one*, 16(1), e0245533.
- Bischof, D., & Kurer, T. (2023). Place-Based Campaigning: The Political Impact of Real Grassroots Mobilization. *The Journal of Politics*, 85(3), 984–1002.
<https://doi.org/10.1086/723985>
- Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley.
- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, June). *R for Data Science*. O'Reilly Media, Inc.
<https://r4ds.hadley.nz/>