



CORRELAID

GOOD CAUSES. BETTER EFFECTS.

Web Scraping mit R

- **INSIDE DATA BODENSEE** -

Web Data Collection: Scraping & APIs

Web Scraping mit R

- Agenda -

Data4Good bei CorrelAid

Zoé Wolter | Was ist CorrelAid und warum sind wir hier?

Web Data Collection

Zoé Wolter | Was ist Web Scraping und wie geht das mit R?

Application Programming Interfaces

Philipp Bosch | Was sind APIs und wie nutzt man diese?

21. Oktober

28. Oktober

Web Scraping mit R

- Agenda -

Data4Good bei CorrelAid

Zoé Wolter | Was ist CorrelAid und warum sind wir hier?

Web Data Collection

Zoé Wolter | Was ist Web Scraping und wie geht das mit R?

Application Programming Interfaces

Philipp Bosch | Was sind APIs und wie nutzt man diese?

Data4Good bei CorrelAid

16:00 - 16:10

Web Data Collection - Theorie

16:10 - 16:30

Web Scraping mit R - Grundlagen

16:30 - 17:00

Hands-On Session

17:10 - 17:30

Web Scraping mit R - At Scale

17:30 - 18:00

Hands-On Session

Bis nächste Woche :)

Wer wir sind



Philipp Bosch

Ehemaliger Head of Community Management bei CorrelAid e.V.

Aktuell: Data Scientist beim Statistischen Kantonsamt Zürich



Zoé Wolter

Bildung & Data Literacy bei CorrelAid e.V.

M.Sc. Social & Economic Data Science
Konstanz



Frie Preu

Chief Operating Officer bei CorrelAid e.V.

Berlin



Nicolas Fröhlich

Volunteer bei CorrelAid e.V.:
Mentoringprogramm

M.Sc. Social & Economic Data Science
Konstanz

Wer wir sind



Wer wir sind





CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Data4Good bei CorrelAid



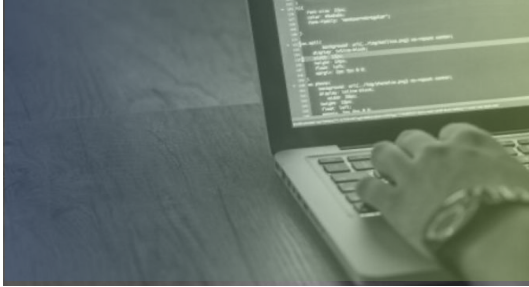
Wer wir sind



Wir sind ein deutschlandweites Netzwerk von über 2,200 Data Scientists, die die Welt durch Data Science verbessern wollen.

#MetaWeltretter

Unsere Mission



DATA4GOOD PROJEKTE

Wir führen pro-bono Datenanalyseprojekte für gemeinnützige Organisationen durch.



BILDUNG

Wir vernetzen engagierte sozial denkende Datenanalytist:innen und bieten ihnen Möglichkeiten ihr Wissen anzuwenden und zu erweitern. Außerdem vermitteln wir engagierten Menschen von gemeinnützigen Organisationen grundlegende Data Literacy Skills.



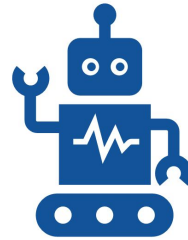
COMMUNITY

Wir treten in den Dialog über den Wert und Nutzen von Daten und Datenanalysen für das Gemeinwohl.



CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Setup





CORRELAID

GOOD CAUSES. BETTER EFFECTS.

<https://github.com/CorrelAid/workshop-webscraping>

The screenshot shows a GitHub repository page for 'CorrelAid/workshop-webscraping'. At the top, there's a navigation bar with 'README.md' selected. Below this is a header section with the CorrelAid logo on the left and the text 'CORRELAID GOOD CAUSES. BETTER EFFECTS.' on the right. The main content area has a title 'Inside Data Bodensee - Webscraping mit R' followed by a welcome message: 'Herzlich Willkommen zu unserem Workshop zu Webscraping mit R!'. Below this is a 'Download' section with a bulleted list of instructions. The word 'hier' in the first bullet point is highlighted with a red box.

README.md



CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Inside Data Bodensee - Webscraping mit R

Herzlich Willkommen zu unserem Workshop zu Webscraping mit R!

Download

- Die Materialien zum Workshop können **hier** heruntergeladen werden.
- Danach die Datei entzippen und in dem Ordner speichern, wo das R-Projekt abgelegt werden soll.
- Öffnen des R-Projekts über die Datei `web-scraping.Rproj`



CORRELAID

GOOD CAUSES. BETTER EFFECTS.

<https://github.com/CorrelAid/workshop-webscraping>

web-scraping.Rproj > 01session.Rmd

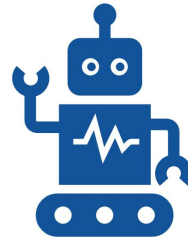
Zeile 34:

```
```\nsource(knitr::purl("packages.Rmd", quiet = TRUE))\n```
```



CORRELAID  
GOOD CAUSES. BETTER EFFECTS.

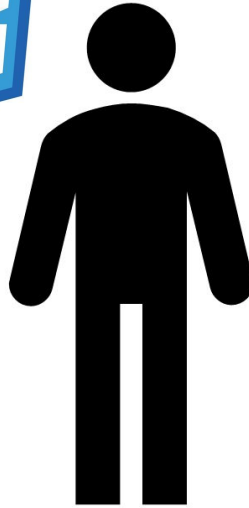
# Web Data Collection



# Aus was bestehen Websites?



Seitenstruktur und Inhalt



Layout und Design



Interaktion und Verhalten

# HTML

Hypertext Markup Language

Anweisungen an den Browser, was wann und wo **angezeigt** werden soll

Für Web Scraping: wir müssen HTML nicht schreiben, aber verstehen hilft!

Baumstruktur: hierarchisch

Tags mit Attributen

```
<!DOCTYPE html>
```

```
<html>
```

```
<head>
```

```
<meta charset="UTF-8">
```

```
<title>TITEL DER SEITE</title>
```

```
</head>
```

```
<body>
```

```
<h1>ÜBERSCHRIFT</h1>
```

Tag

Element

```
<p>TEXT DES PARAGRAPHEN</p>
```

Element-  
name

```
KLICKE HIER
```

Attribut-

Attributwert

```
</body>
```

```
</html>
```

# HTML

<code>&lt;a href=""&gt;</code>	Link
<code>&lt;div&gt;</code> und <code>&lt;span&gt;</code>	Organisation der Seite
<code>&lt;p&gt;</code>	Paragraph
<code>&lt;h1&gt;</code> , <code>&lt;h2&gt;</code> ,...	Überschriften in unterschiedlichen Größen
<code>&lt;ul&gt;</code> , <code>&lt;ol&gt;</code> , <code>&lt;dl&gt;</code>	Listen (unordered, ordered...)
<code>&lt;li&gt;</code>	Einzelnes Listenelement
<code>&lt;br&gt;</code>	Zeilenumbruch
<code>&lt;b&gt;</code> , <code>&lt;i&gt;</code> , <code>&lt;strong&gt;</code>	Layout
<code>&lt;table&gt;</code> , <code>&lt;th&gt;</code> , <code>&lt;td&gt;</code> , <code>&lt;tr&gt;</code>	Tabellen
<code>&lt;script&gt;</code>	Script Container



# XPath

## XML Path Language

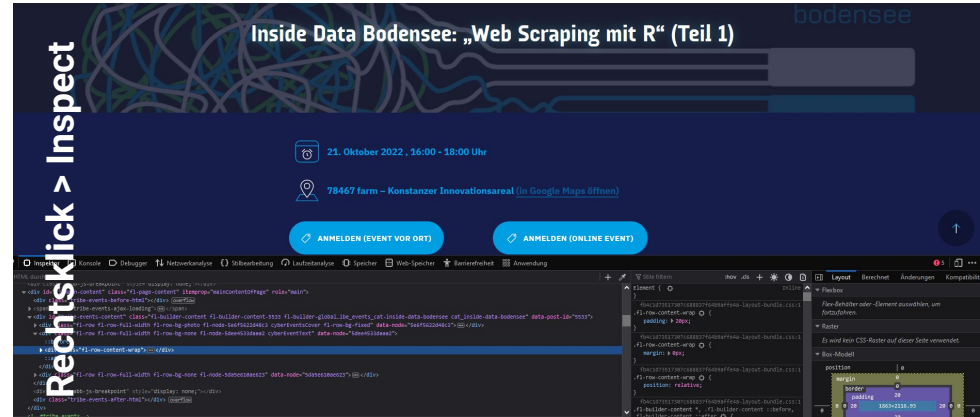
**Anfragesprache**, um Teile von HTML / XML-Dateien zu **extrahieren**

Ausnutzen von Bezeichnungen, Attributen und Beziehungen von **Nodes** / Tags

Basiert auf hierarchischer Anordnung von Nodes

**Absolute** Pfade: `"/html/body/div/p"`

**Relative** Pfade: `"//p"`



# robots.txt

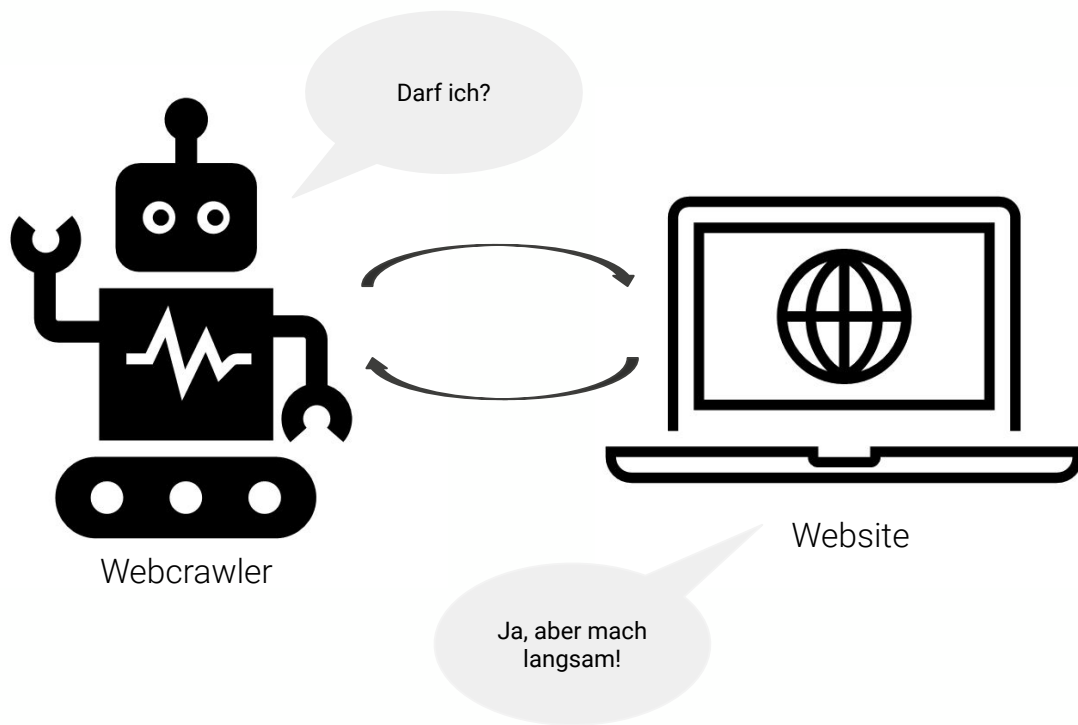
## Robots Exclusion Standard

**Mitteilung** an (Suchmaschinen-) **Crawler**, auf welche URLs einer Website sie **zugreifen dürfen**

Ziel: Vermeidung einer **Überlastung** der Website

Definition eines **Crawl-Delays**  
[cyberlago.net/robots.txt](http://cyberlago.net/robots.txt)

```
User-agent: *
Disallow: /wp-admin/
Allow: /wp-admin/admin-ajax.php
```



# Web Scraping in R



## “Web-Etikette”

**bow:** Client-Vorstellung beim Host, Abfrage des robots.txt

**nod:** neuer Pfad, wenn Session bereits erstellt

**Crawl-Delay:** automatisch berücksichtigt!



## Scraping

`html_element()`

`html_nodes()`

`html_text()`

`html_table()`



## Datenbereinigung

Daten in ein sauberes Format bringen, um damit weiterarbeiten zu können!



## Scaling

Was für einer Website klappt, funktioniert auch für ganz viele!

# Werdet Teil von CorrelAid!



## Open Onboarding Call am 07.11.!

<https://pretix.eu/correlaid/open-onboarding/>



## Interessiert an der Lokalgruppe?

[konstanz@correlaid.org](mailto:konstanz@correlaid.org) oder fragt direkt Zoé oder Phil




## Eigene Projektideen?

[info@correlaid.org](mailto:info@correlaid.org)



## R Lernen für NPOs!

[info@correlaid.org](mailto:info@correlaid.org)



CORRELAID  
GOOD CAUSES. BETTER EFFECTS.

What is  
Data4Good?

SEMESTER  
KICK-OFF!

WHEN?  
24th October 2022 • 6pm

WHERE?  
KNIME GmbH, Reichenaustraße 11  
78467 Konstanz

