CORRELAID

GOOD CAUSES. BETTER EFFECTS.

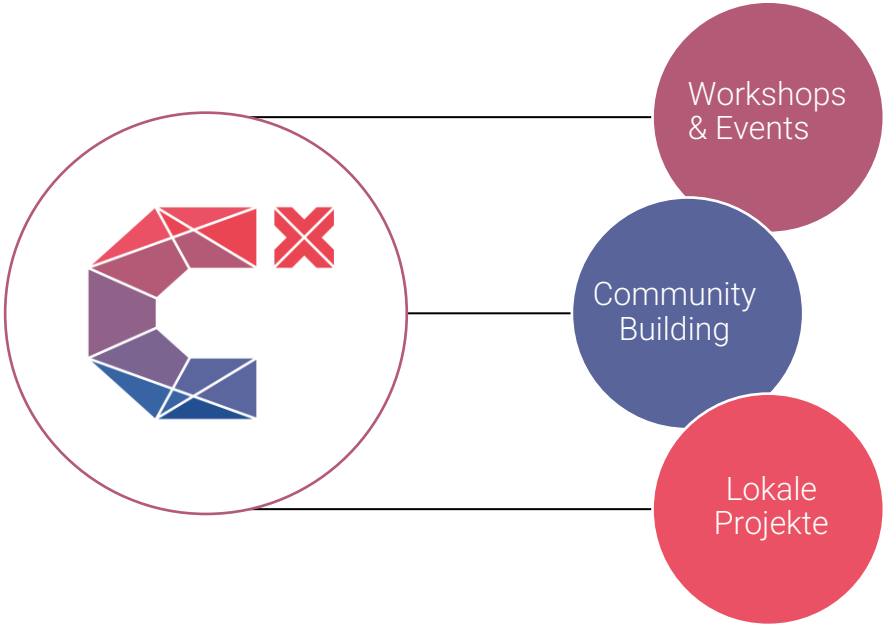
Web Scraping mit R

Web Data Collection: Scraping the Web & Using APIs

Agenda

- Was ist CorrelAid?
- Was ist CorrelAid in Konstanz?
- Web Scraping
  - Wofür brauchen wir das eigentlich?
  - HTML
  - XPath
  - APIs

CorrelAid & CorrelAidXKonstanz



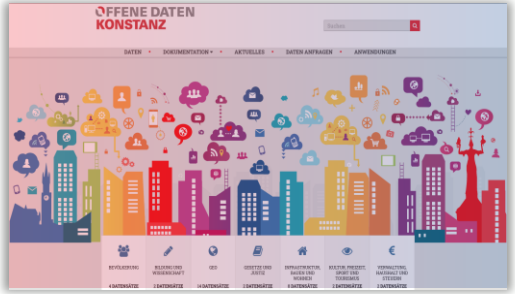
CorrelAid

- Netzwerk von über 1000 Datenanalyst:innen
- Studierende und Berufstätige
- Data4Good-Projekte

Local Chapter Konstanz

- Gründungsort!
- Workshops
- Lokale Projekte mit NPOs etc.

Unser Local Chapter

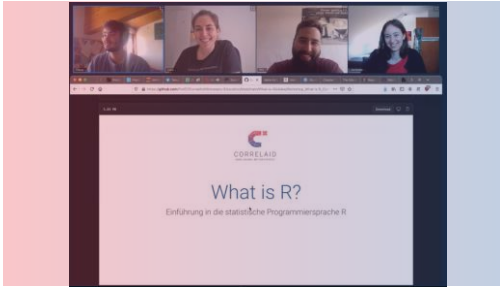


LOKALE PROJEKTE

In Data4Good-Projekten unterstützen wir andere Non-Profits wie Stolpersteine, das Meiste aus ihren Daten herauszuholen.

LOKALE KOOPERATIONEN

In Kooperationen mit der Stadt Konstanz, dem ADILT oder auch dem SWR werden Packages geschrieben, Vorlesungen gehalten und Hackathons veranstaltet.

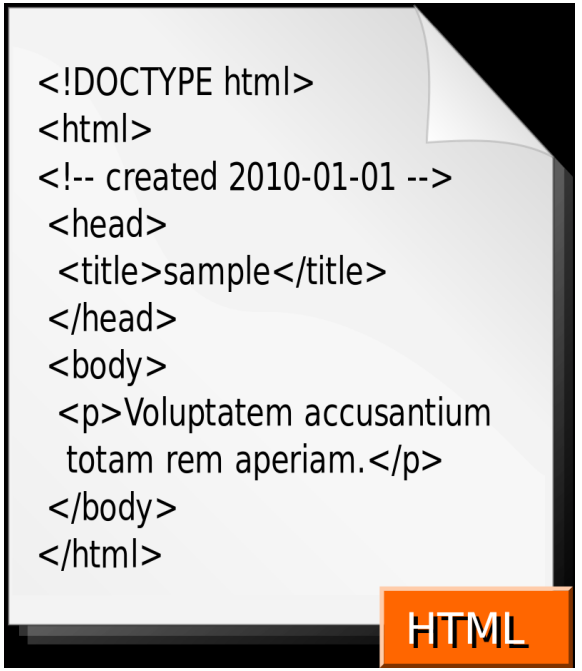


WORKSHOPS

An der Uni gibt es regelmäßig Workshops (bisher) zur verschiedenen Bereichen in R.

HTML

- Hypertext Markup Language
- Text mit Markup als Standard für Webseiten
- Anweisungen an den Browser, was wann und wo angezeigt werden soll
- Für Web Scraping: wir müssen HTML nicht schreiben, aber verstehen hilft!
- Baumstruktur
- Systematisch, hierarchisch, nested
- Tags mit Attributes



https://de.wikipedia.org/wiki/Hypertext_Markup_Language

HTML

- Hypertext Markup Language
- Text mit Markup als Standard für Webseiten
- Anweisungen an den Browser, was wann und wo angezeigt werden soll
- Für Web Scraping: wir müssen HTML nicht schreiben, aber verstehen hilft!
- Baumstruktur
- Systematisch, hierarchisch, nested
- Tags mit Attributes

<a> anchor tag

 and organizational tags

<p> paragraphs tag

<h1>, <h2>, ... headline tags

, <ol>, <dl> list tags

 list item tag

 line break tag

<b>, <i>, emphasis/layout tags

<table>, <tr>, <td>, <th> table tags

<form> server interaction tag

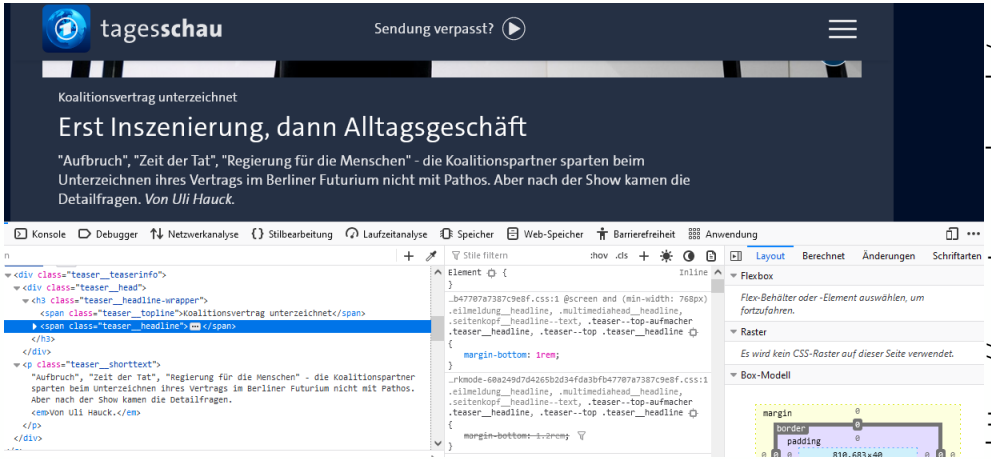
<script> script container tag

XPath

- Anfragesprache, um Teile von HTML/XML-Dateien zu extrahieren
- Ausnutzen von Bezeichnungen, Attributen und Beziehungen von Nodes/Tags
- Basiert auf der hierarchischen Anordnung von Nodes
- Absolute Pfade: “/html/body/div/p/i"
- Relative Pfade: “//p/i“

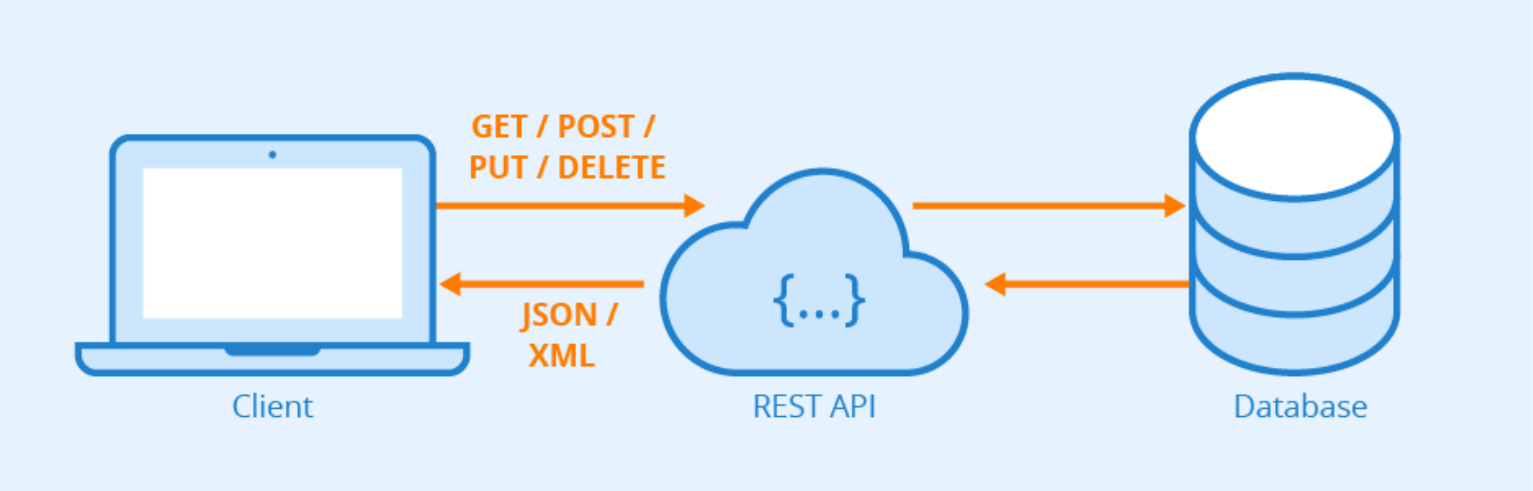
SelectorGadget:
point and click CSS selectors

https://selectorgadget.com/



https://www.tagesschau.de/

APIs



https://www.astera.com/wp-content/uploads/2020/01/rest.png

CORRELAID

KONSTANZ

Contact us at:

konstanz@correlaid.org

@CorrelaidxKN

or join our Slack-Channel:

#lc-konstanz