GESIS Leibniz Institute for the Social Sciences

Leibniz Association

Working with Web Tracking Data

Felix Schmidt Maximilian Haag

2026-02-19

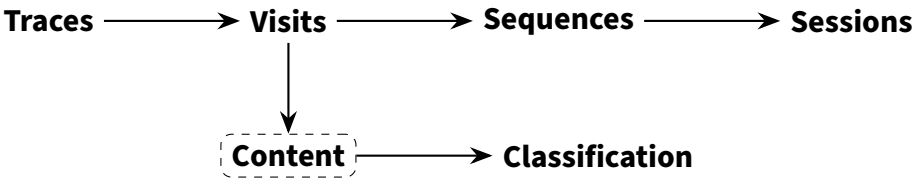
What is web tracking data?

- Records of individual-level browsing behavior (traces) collected via browser extensions or metering software
- Each record: **who** visited **what URL** at **what time**
- Typically linked with survey data (demographics, attitudes, etc.)

Minimum data structure:

panelist_id	url	timestamp
P001	https://www.spiegel.de/politik/…	2024-04-15 09:23:45
P001	https://www.google.com/search?q=…	2024-04-15 09:24:12

Goal


Traces
Visits
Sequences
Sessions
Content
Classification

Our data today

GESIS Panel Campusfile (ZA5670)

- ~1.15 million URL visits from ~600 panelists
- Tracking period: April–June 2024
- Enhanced with **synthetic full URLs** for teaching purposes

Data quality heterogeneity

- **Full URLs** with paths and query parameters
- **Domain-only** entries, no path information
- **Privacy placeholders** (“blacked_out”, “full_deny”)

→ Handling this heterogeneity is a core preprocessing challenge.

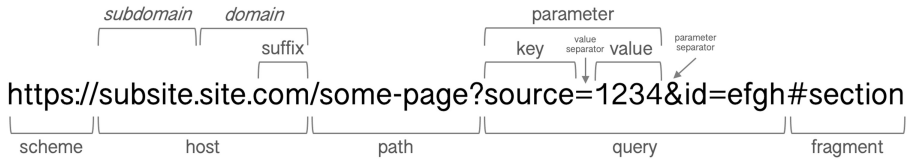
What about HTML content?

In addition to URLs, we also collect the HTML content of visited pages.

→ Working with HTMLs (parsing, text extraction, content analysis) will be covered in tomorrow's session.

Today's session focuses on **URL-level preprocessing** as the foundation.

URL anatomy



Source: Clemm von Hohenberg et al. (2024), Fig. 2

Each component carries different information for analysis:

- **Domain/Host**: Which website (e.g., spiegel.de)
- **Path**: Which page or section (e.g., /politik/article-name)
- **Query parameters**: Search terms, filters, tracking codes

The preprocessing workflow

1. **Parse URLs** – Extract components (host, domain, path, query)
2. **Clean and deduplicate** – Remove tracking artifacts, consecutive duplicates
3. **Classify visits** – Assign categories (news, social, e-commerce, …)
4. **Aggregate** – Summarize per panelist, per category, per time period

Each step involves **decisions** that affect downstream analysis.

Tools we'll use

R packages:

- `webtrackR` – purpose-built for web tracking data preprocessing (extract, deduplicate, classify, aggregate)
- `adaR` – robust URL parsing (alternative to manual regex)
- `dplyr` – general data manipulation

Let’s begin!

Leibniz Association