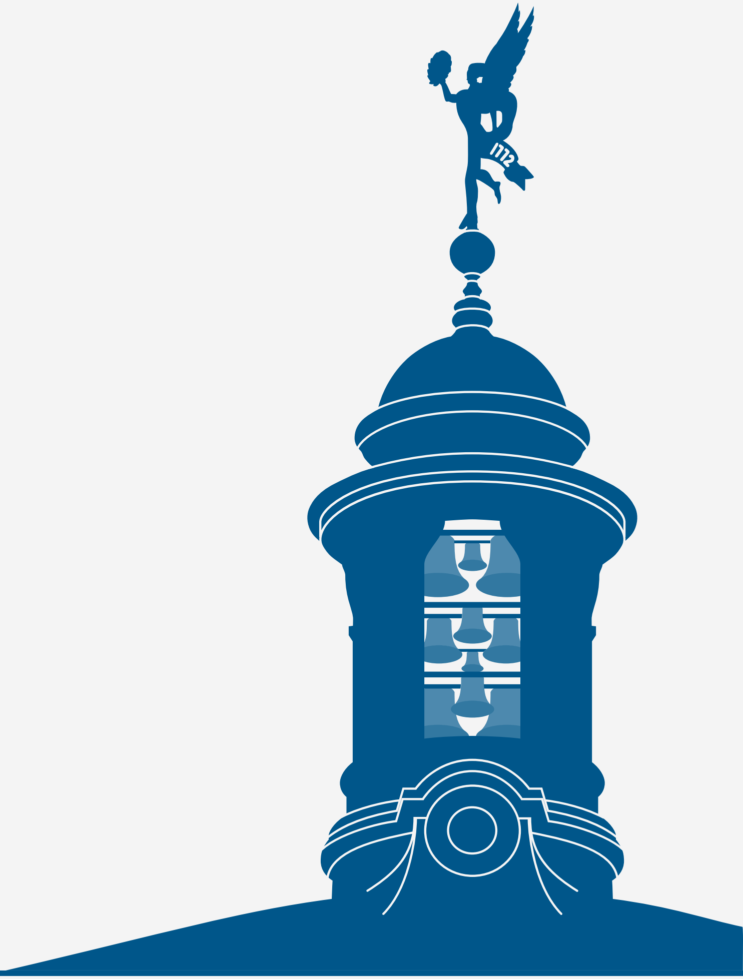
Automated Data Collection

Basics: Data access, data formats and data collection methods

Jakob Jünger, Chantal Gärtner



Goals

Practical skills

Basic Knowledge

Goals

Which Facebook post has the most comments?

How to collect posts and comments of Facebook pages?

Which country received the most news coverage last month?

How to collect online news?

Schedule

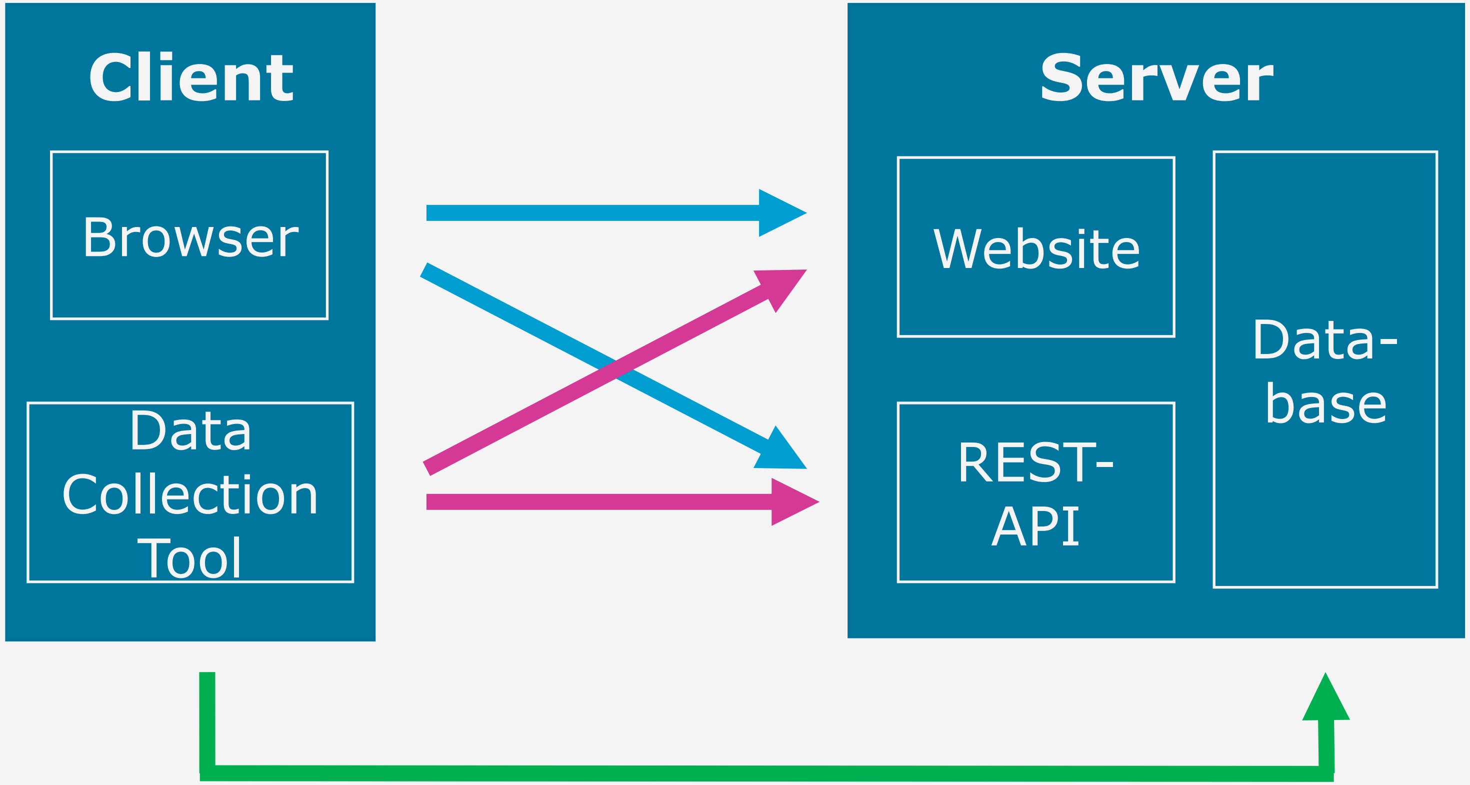
Day 1: Intro & APIs	
10:00 – 11:30	Introduction to automated data collection
11:30 – 11:45	*Break*
11:45 – 13:00	Practical Session: Using APIs with Facepager
13:00 – 14:00	*Lunch Break*
14:00 – 15:15	Practical Session: Data wrangling with python

Day 2: Webscraping	
10:00 – 11:30	Introduction to webscraping
11:30 – 11:45	*Break*
11:45 – 13:00	Practical Session: Webscraping with Python
13:00 – 14:00	*Lunch Break*
14:00 – 15:15	Practical Session: Webcrawling with Python

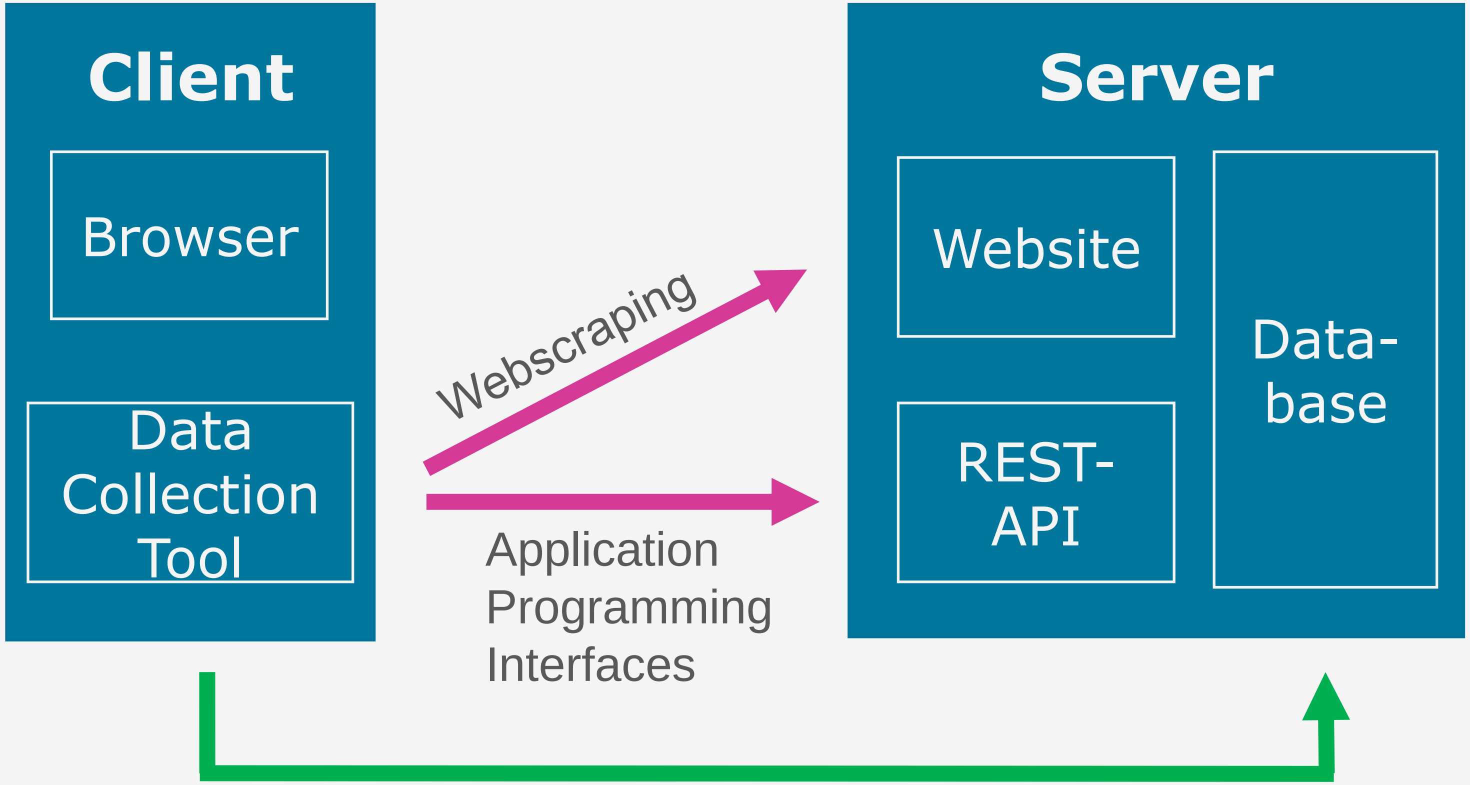
Schedule

Day 3: Webcrawling and advanced techniques	
9:00 – 10:30	Challenges of automatic data collection
10:30 – 11:00	*Break*
11:00 – 12:30	Browser automating using Selenium
12:30 – 13:30	*Lunch Break*
13:30 – 14:30	Recap and open questions

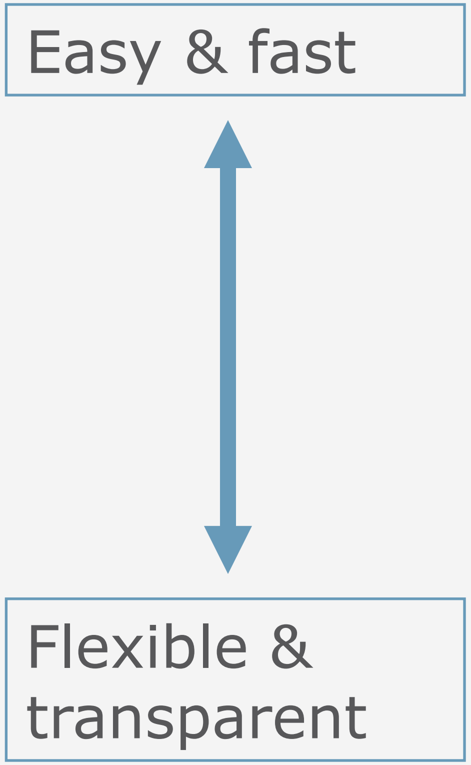
Automated Data Collection on the Web



Automated Data Collection on the Web



Tools & Co



Easy & fast

Flexible & transparent

- Option 1: Commercial **Services**
- Option 2: **Tools** with user interfaces
  - Local: Facepager, Rapidminer
  - Server: DMI Tools
- Option 3: **Packages** for R or Python
  - Example: twitterR
- Option 4: **Frameworks**
  - Example: Scrapy
- Option 5: Develop **scripts**
  - Python: requests, beautifulsoup, selenium
  - R: Rvest, Rselenium

Important Data Formats

XML/HTML

```
<p class="important">Content of a
paragraph</p>
```

- Markup language
- Text structured by elements ("tags")
- Hierarchical structure

JSON `{"Name": „Al Jazeera"}`

- Key value pairs and lists
- Hierarchical structure

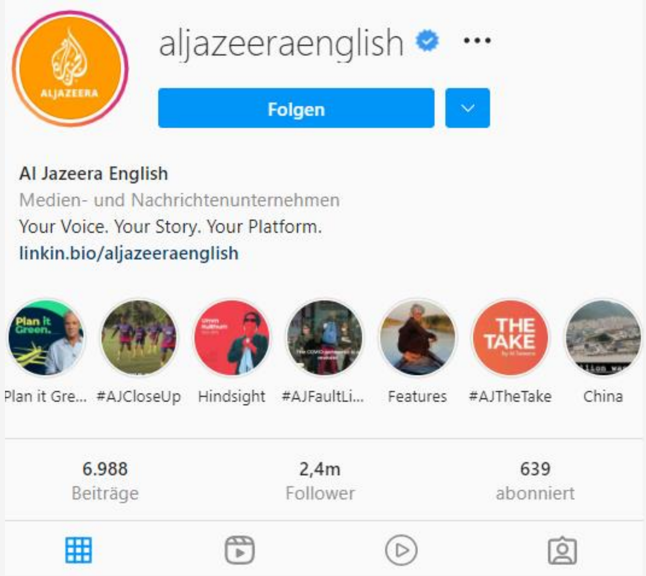
CSV

```
Post;Author;Text\n
```

- Tabular format, each row is a record, first row is header
- Fields separated by comma or semicolon

Example: https://www.instagram.com/aljazeeraenglish/?__a=1

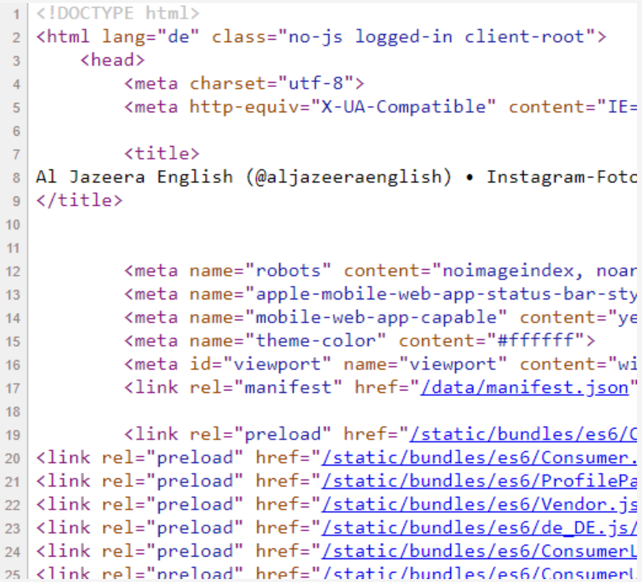
User Interface
(Browser)



Al Jazeera English
Medien- und Nachrichtenunternehmen
Your Voice. Your Story. Your Platform.
linkin.bio/aljazeeraenglish

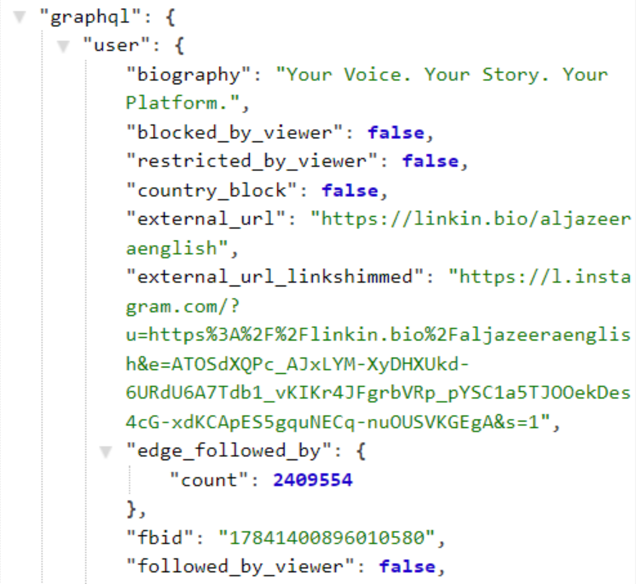
6.988 Beiträge
2,4m Follower
639 abonniert

HTML
(Webscraping)



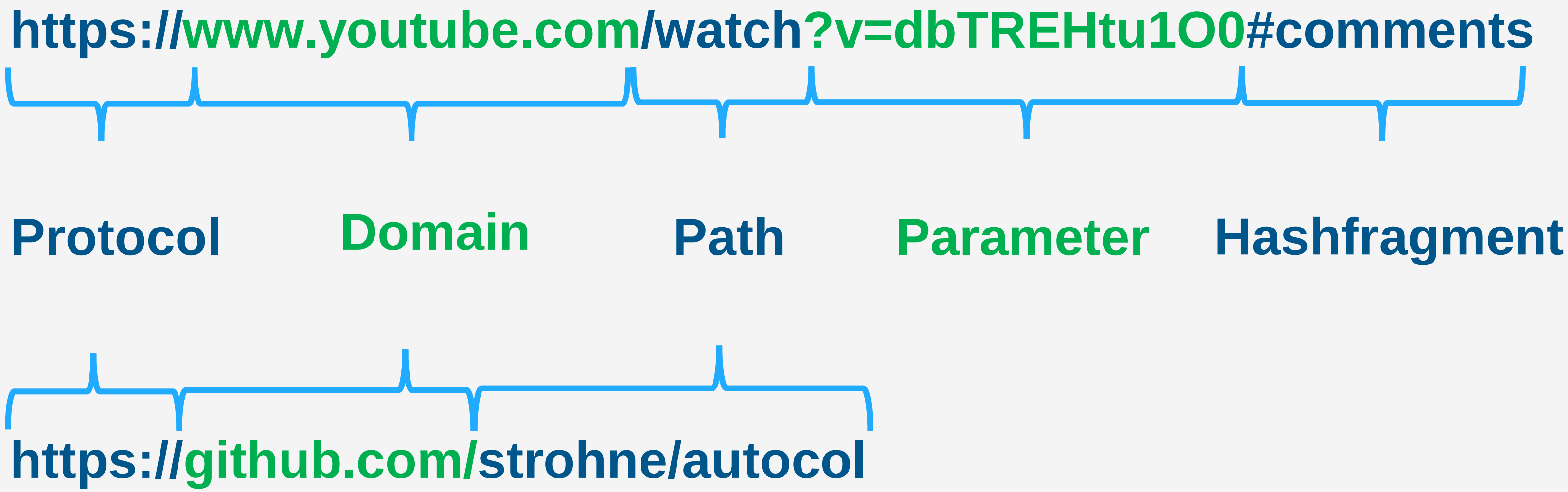
```html
<!DOCTYPE html>
<html lang="de" class="no-js logged-in client-root">
    <head>
        <meta charset="utf-8">
        <meta http-equiv="X-UA-Compatible" content="IE=

        <title>
Al Jazeera English (@aljazeeraenglish) • Instagram-Fotc
</title>


        <meta name="robots" content="noimageindex, noar
        <meta name="apple-mobile-web-app-status-bar-sty
        <meta name="mobile-web-app-capable" content="ye
        <meta name="theme-color" content="#ffffff">
        <meta id="viewport" name="viewport" content="wi
        <link rel="manifest" href="/data/manifest.json"

        <link rel="preload" href="/static/bundles/es6/C
<link rel="preload" href="/static/bundles/es6/Consumer.
<link rel="preload" href="/static/bundles/es6/ProfilePa
<link rel="preload" href="/static/bundles/es6/Vendor.js
<link rel="preload" href="/static/bundles/es6/de_DE.js/
<link rel="preload" href="/static/bundles/es6/ConsumerL
<link rel="preload" href="/static/bundles/es6/ConsumerL
```

JSON
(API)



```
"graphql": {
    "user": {
        "biography": "Your Voice. Your Story. Your
        Platform.",
        "blocked_by_viewer": false,
        "restricted_by_viewer": false,
        "country_block": false,
        "external_url": "https://linkin.bio/aljazeer
        aenglish",
        "external_url_linkshimmed": "https://l.insta
        gram.com/?
        u=https%3A%2F%2Flinkin.bio%2Faljazeeraenglis
        h&e=ATOSdXQPc_AJxLYM-XyDHXUkd-
        6URdU6A7Tdb1_vKIKr4JFgrbVRp_pYSC1a5TJOOekDes
        4cG-xdKCApES5gquNECq-nuOUSVKGEgA&s=1",
        "edge_followed_by": {
            "count": 2409554
        },
        "fbid": "17841400896010580",
        "followed_by_viewer": false,
```

URLs



https://www.youtube.com/watch?v=dbTREHtu1O0#comments

Protocol

Domain

Path

Parameter

Hashfragment

https://github.com/strohne/autocol