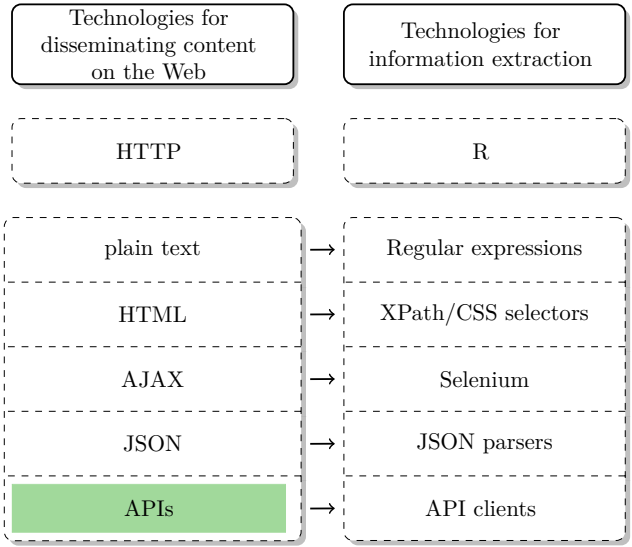
Introduction to Web Scraping with R

APIs

Simon Munzert | IPSDS

Technologies of the World Wide Web



What are APIs?

What are APIs?

Definition

- **A**pplication **P**rogramming **I**nterface
- "data search engine": you pose a request, the API answers with a bulk of data
- let you/your program query a provider for specifc data
- common data formats: XML, JSON
- many popular web services provide APIs (Twitter, Google, Facebook, Wikipedia, ...)

What are APIs?

Definition

- **A**pplication **P**rogramming **I**nterface
- "data search engine": you pose a request, the API answers with a bulk of data
- let you/your program query a provider for specifc data
- common data formats: XML, JSON
- many popular web services provide APIs (Twitter, Google, Facebook, Wikipedia, . . .)

Why we should care about APIs

- provide instant access to clean data
- free us from building manual scrapers
- API usage implies mutual agreement about data collection process

Example

Example

Google Maps API

- Google provides access to powerful location services
- free service (at least when used modestly)
- input/output: places, names, coordinates, maps, ...
- see also: https://developers.google.com/maps/documentation/

Example

Google Maps API

- Google provides access to powerful location services
- free service (at least when used modestly)
- input/output: places, names, coordinates, maps, ...
- see also: https://developers.google.com/maps/documentation/

Potential use cases

- geocode observations based on address, city, post code, ...
- calculate distances between observations
- map observations

Example

Access the API with R

- the `ggmap` package provides high-level functions to access the API
- in this case, all we have to do is to figure out how the R function works – the communication with the API is processed completely in the background

Example

Access the API with R

- the ggmap package provides high-level functions to access the API
- in this case, all we have to do is to figure out how the R function works – the communication with the API is processed completely in the background

R code

```
3  library(ggmap)
   Google Maps API Terms of Service: http://developers.google.com/maps/terms.
   Please cite ggmap if you use it: see citation('ggmap') for details.
```

end

Example

Access the API with R

- the ggmap package provides high-level functions to access the API
- in this case, all we have to do is to figure out how the R function works – the communication with the API is processed completely in the background

R code

```
5  library(ggmap)
   Google Maps API Terms of Service: http://developers.google.com/maps/terms.
   Please cite ggmap if you use it: see citation('ggmap') for details.
```

end

R code

```
6  geocode("Berlin, Germany")
   Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Berlin,%20
   Germany&sensor=false
         lon      lat
   1 13.40495 52.52001
```

end

Example

Excerpt from raw JSON behind the call

```
{
   "results" : [
      {
         "address_components" : [
            {
               "long_name" : "Berlin",
               "short_name" : "Berlin",
               "types" : [ "locality", "political" ]
            },
         ],
         "formatted_address" : "Berlin, Germany",
            "location" : {
               "lat" : 52.52000659999999,
               "lng" : 13.404954
            }
         },
         "place_id" : "ChIJAVkDPzdOqEcRcDteW0YgIQQ",
         "types" : [ "locality", "political" ]
      }
   ]
}
```

Example

Map the location

Example

Map the location

R code

```
8  get_googlemap("Berlin, Germany", zoom =
   12, maptype = "hybrid") %>% ggmap()
```

end

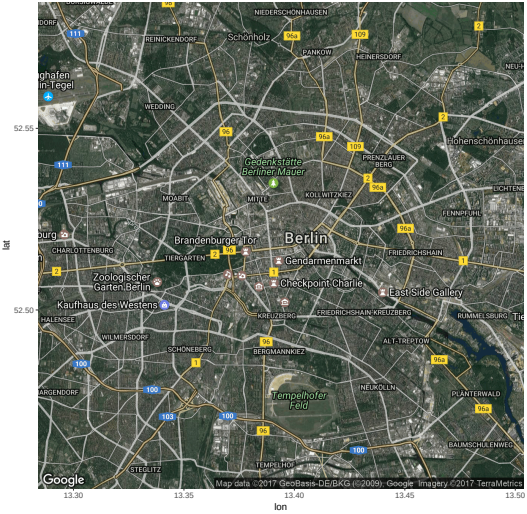
Example

Map the location

R code

```
9  get_googlemap("Berlin, Germany", zoom =
   12, maptype = "hybrid") %>% ggmap()
```

end



Summary

Summary

Advantages of API use

- collecting data from the web using APIs provided by the data owner represents **the gold standard of web data retrieval**
- pure data collection without 'layout waste'
- standardized data access
- de facto automatic agreement of data owner
- robustness of calls

http://maxpixel.freegreatpicture.com/photo-1461569

Summary

Advantages of API use

- collecting data from the web using APIs provided by the data owner represents **the gold standard of web data retrieval**
- pure data collection without 'layout waste'
- standardized data access
- de facto automatic agreement of data owner
- robustness of calls

http://maxpixel.freegreatpicture.com/photo-1461569

Disadvantages of API use

- (sometimes) requires knowledge of API architecture
- dependent upon API suppliers
- use not always free