Web Data Collection with R

HTTP

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

Teaser

Teaser

URL example

https://en.wikipedia.org/wiki/Lion
http://r-datacollection.com
400 page not found 503 intenal server error 200 ok

things returned by httr functions

```
library(httr)

res <- GET("example.com")

names(res)
```

```
##  [1] "url"         "status_code" "headers"     "all_hea
##  [6] "content"     "date"        "times"       "request"
```

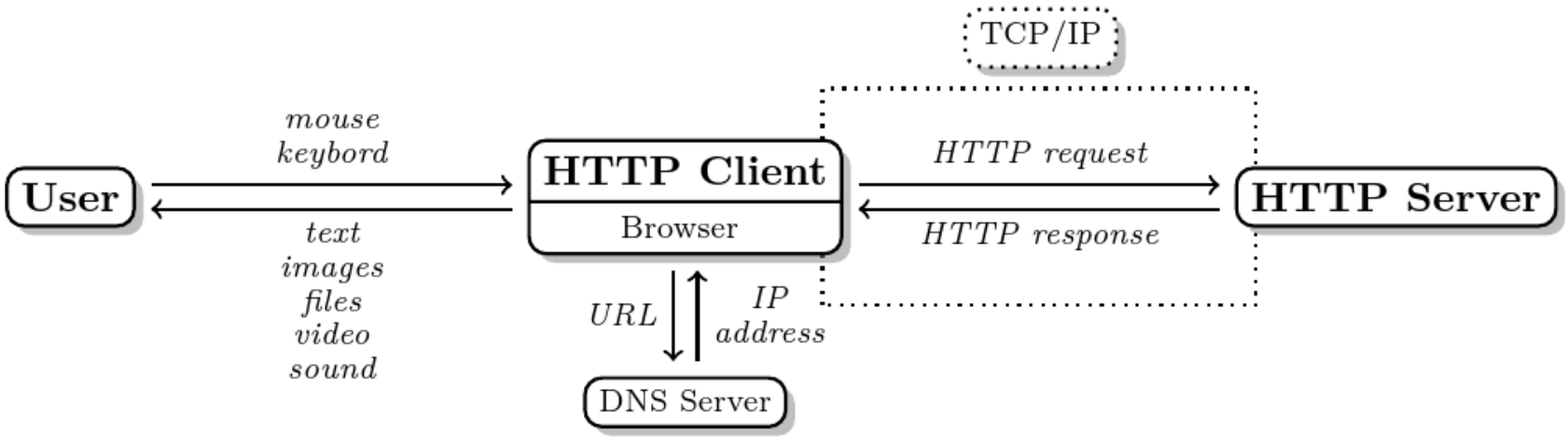
```
res$url
```

```
## [1] "HTTP://example.com/"
```

```
res$status_code
```

```
## [1] 200
```

client server communication



HTTP requests

1. Establishing connection

```
1  About to connect() to www.r-datacollection.com port 80 (#0)
2    Trying 173.236.186.125... connected
3  Connected to www.r-datacollection.com (173.236.186.125) port 80 (#0)
4  Connection #0 to host www.r-datacollection.com left intact
```

2. HTTP request

```
1  GET /index.html HTTP/1.1
2  Host: www.r-datacollection.com
3  Accept: */*
```

HTTP responses

3. HTTP response

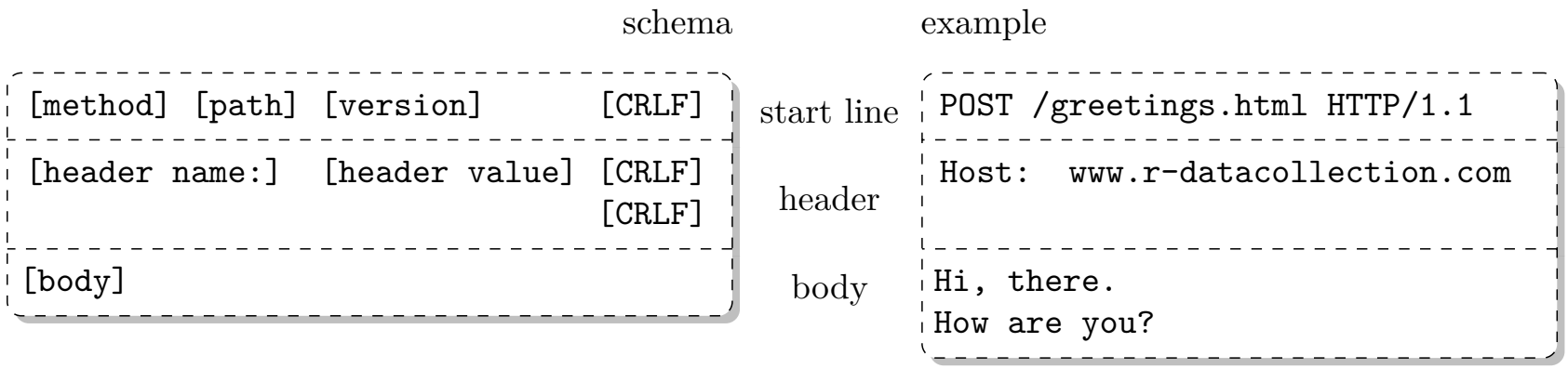
```
1  HTTP/1.1 200 OK
2  Date: Thu, 27 Feb 2014 09:40:35 GMT
3  Server: Apache
4  Vary: Accept-Encoding
5  Content-Length: 131
6  ...

8  <!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML//EN">
9  <html> <head>
10 <title></title>
11 </head>
12 ...
```

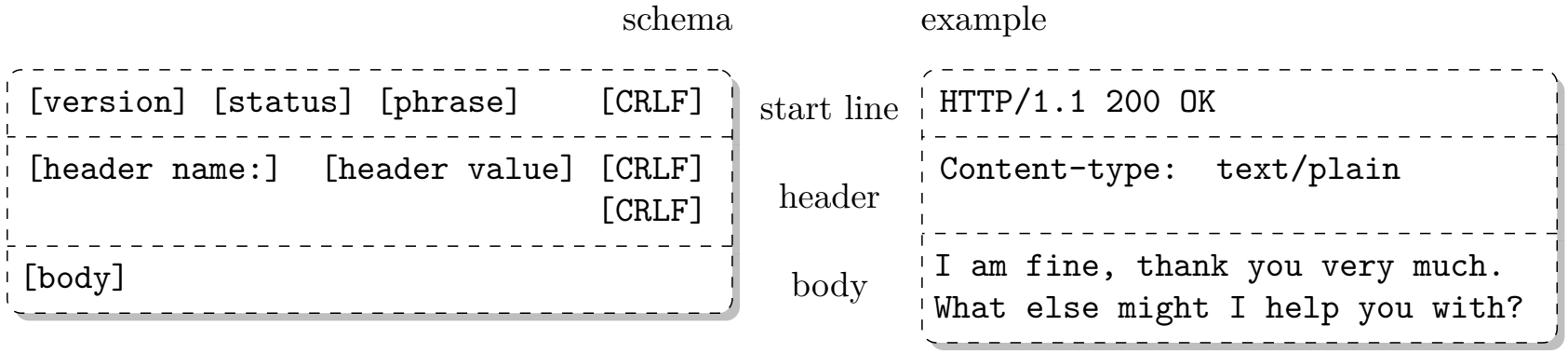
4. Closing connection

```
1  Closing connection #0
```

HTTP requests



HTTP responses



HTTP methods

Method	Description
GET	Retrieves resource from server
POST	Retrieves resource from server using the message body to send data or files to the server
HEAD	Works like *GET*, but server responds only with start line and header, no body
PUT	Stores the body of the request message on the server
DELETE	Deletes a resource from the server
TRACE	Traces the route of the message along its way to the server
OPTIONS	Returns list of supported HTTP methods
CONNECT	Establishes a network connection

GET and POST

- the same but GET puts all the information in the query string while POST puts the information in the body of the request

Error codes

- 1xx : information
- 2xx : ok
- 3xx : redirect
- 4xx : client error
- 5xx : server error

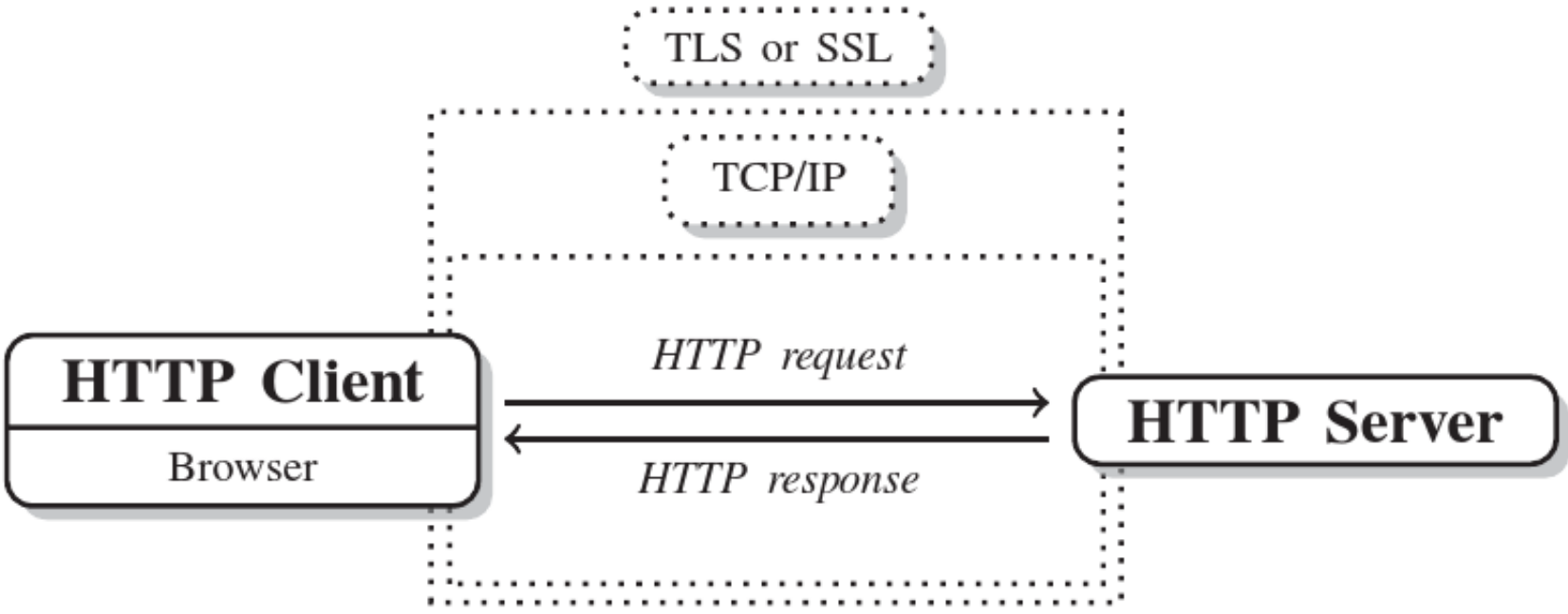
Cookies

1. client request to server
2. server response with header field "Set-Cookie: sessionid=1234; path=/; domain=r-datacollection.com; expires=Mon, 31-Dec-2035 23:00:01 GMT"
3. client request repeating the cookie values: `Cookie: sessionid=1234`

Identification

- useragent (e.g. R 3.2.3 / httr 1.0.0)
- referer (Last_Page_I_Visited.html)
- from (e.g. bot@botnet.com)
- cookie (user=0112343asas)

HTTPs


TLS or SSL
TCP/IP
HTTP Client
Browser
HTTP request
HTTP response
HTTP Server

HTTP and httr

- httr is a wrapper package to the curl package
- httr functions mirror HTTP methods (GET, POST, PUT, DELETE, ...)
- httr tries to have very reasonable defaults (e.g. handles cookies by default, follows redirections, ...)
- httr functions return the whole communication (request, response)

HTTP and httr - queries

```
library(httr)
library(rvest)
```

```
## Loading required package: xml2
```

HTTP and httr - queries

```
url <- "http://www.r-datacollection.com/materials/http/GETe

GET(url) %>% content(as="text") %>% cat()
```

```
## No encoding supplied: defaulting to UTF-8.

## Please specify your name!
## Please specify your age!
```

```
GET(url, query = list(name="Joy", age="22"))  %>%
  content(as="text") %>% cat()
```

```
## No encoding supplied: defaulting to UTF-8.

## Hello Joy!
## You are 22 years old.
```

HTTP and httr - useragent

```
url <- "http://www.r-datacollection.com/materials/http/retu

GET(url) %>% content(as="text") %>% cat()

## No encoding supplied: defaulting to UTF-8.

## GET /materials/http/return.php HTTP/1.1
## Connection: close
## Accept: application/json, text/xml, application/xml, */*
## Accept-Encoding: gzip, deflate
## Host: www.r-datacollection.com
## User-Agent: libcurl/7.35.0 r-curl/0.9.6 httr/1.1.0
## Authorization:
##
```

HTTP and httr - useragent

```
GET(url, user_agent("httr")) %>% content(as="text") %>% cat

## No encoding supplied: defaulting to UTF-8.

## GET /materials/http/return.php HTTP/1.1
## Connection: close
## Accept: application/json, text/xml, application/xml, */*
## Accept-Encoding: gzip, deflate
## Host: www.r-datacollection.com
## User-Agent: httr
## Authorization:
##
```

HTTP and httr - auto parsing

```
url <- "http://www.r-datacollection.com/materials/html/OurF

GET(url)  %>% content() %>%
  html_nodes("title") %>%
  html_text()
```

```
## No encoding supplied: defaulting to UTF-8.

## [1] "First HTML"
```

```
GET(url)  %>% content(as="text")
```

```
## No encoding supplied: defaulting to UTF-8.

## [1] "<!DOCTYPE html>\n <html>\n   <head>\n     <title>Fi
```

HTTP and httr - follow location

```
url <- "http://www.r-datacollection.com/materials/http/redi
try(readLines(url))

GET(url)  %>% content(as="text")
```

```
## No encoding supplied: defaulting to UTF-8.

## [1] "<html>\r\n <head>\r\n  <title>redirected</title>\r\
```

HTTP and httr - (auto) cookies

```
url <- "http://www.r-datacollection.com/materials/http/Cook

content(GET(url))
```

```
## No encoding supplied: defaulting to UTF-8.

## {xml_document}
## <html>
## [1] <body>\n  <p>Hallo, who are you?</p>\n</body>
```

```
content(GET(url))
```

```
## No encoding supplied: defaulting to UTF-8.

## {xml_document}
## <html>
## [1] <body>\n  <p>Ah, nice to meet you again. </p>\n</bod
```

HTTP and httr - methods

```
url <- "http://www.r-datacollection.com/materials/http/retu

GET(url, query=list(a=1,b=3)) %>% content(as="text") %>% ca

## No encoding supplied: defaulting to UTF-8.

## <pre>
## parameters submitted via GET: (name) : (value)
## a  : 1
## b  : 3
##
##
## parameters submitted via POST: (name) : (value)
##
## </pre>
```

HTTP and httr - methods

```
url <- "http://www.r-datacollection.com/materials/http/retu

POST(url, body = list(a=1,b=3)) %>% content(as="text") %>%

## No encoding supplied: defaulting to UTF-8.

## <pre>
## parameters submitted via GET: (name) : (value)
##
##
## parameters submitted via POST: (name) : (value)
## a   : 1
## b   : 3
##
## </pre>
```

HTTP and httr - headers

```
url <- "http://www.r-datacollection.com/materials/http/retu

GET(url, add_headers(from="bot@botnet.de", referrer="botnet
  content(as="text") %>% cat()
```

```
## No encoding supplied: defaulting to UTF-8.
## GET /materials/http/return.php HTTP/1.1
## Connection: close
## Referrer: botnet.de
## From: bot@botnet.de
## Accept: application/json, text/xml, application/xml, */*
## Cookie: id=d10b3aaa226397acafc2275ba02a2586
## Accept-Encoding: gzip, deflate
## Host: www.r-datacollection.com
## User-Agent: libcurl/7.35.0 r-curl/0.9.6 httr/1.1.0
## Authorization:
##
```