

A Fast-Track-Overview on Web Scraping with R

ECPR WSMT Bamberg

Peter Meißner

<https://github.com/petermeissner>
<http://pmeissner.com>

presented: 2016-02-29 / last update: 2016-02-19

Introduction

THE WEB

- ▶ web pages (e.g. `http://example.com`, `http://ecpr.eu/`)
- ▶ web formats (XML, HTML, JSON, ...)
- ▶ web frameworks (HTTP, URL, APIs, ...)
- ▶ social media (Twitter, ...)
- ▶ web data (page views, ip-addresses, ...)

Introduction

phase	problems	examples
download	protocols procedures	HTTP, HTTPS, POST, GET, ... cookies, authentication, forms, ...
extraction	parsing extraction cleansing	translating HTML (XML, JSON, ...) into R getting the relevant parts cleaning up, restructure, combine