scraper

How it works

Infrastructure

Scraper

scrape.bat

config.js

- Multiple Suites (one per site)
- URL's to scrape
- Scrapers
- All Javascript

Pjscrape.js

- Ouputs files as JSON
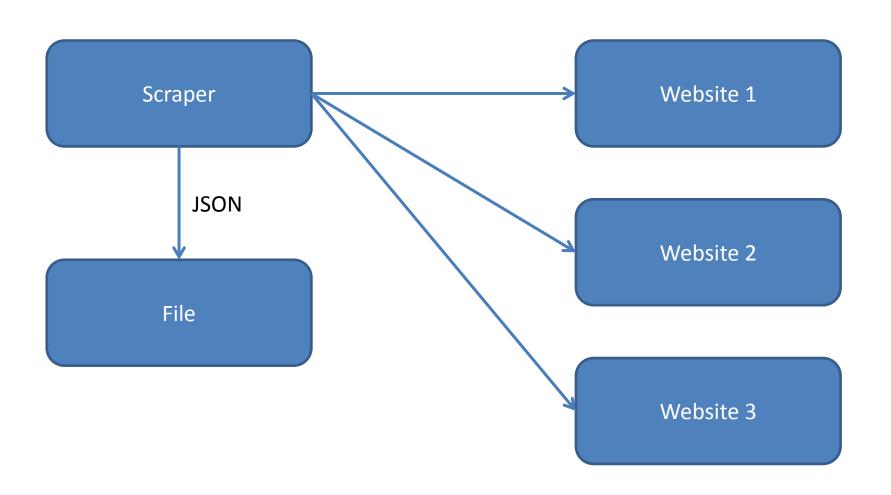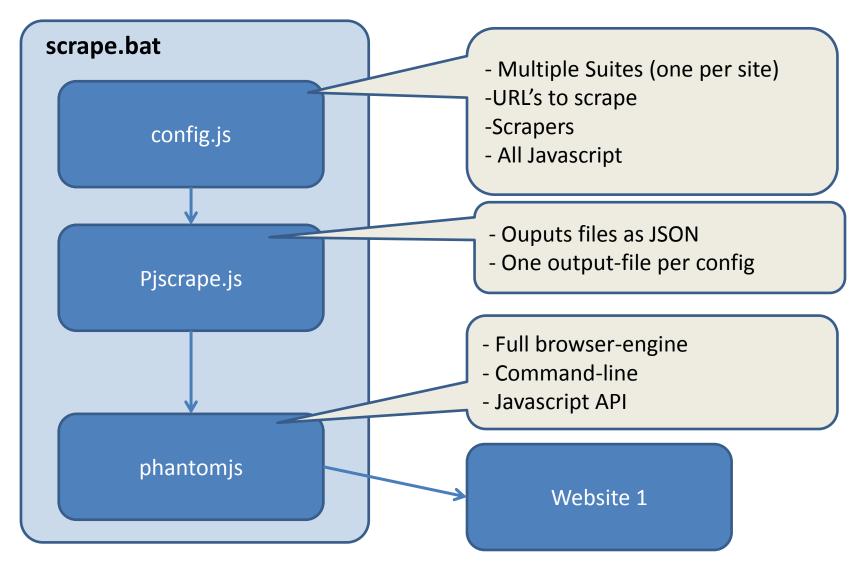- One output-file per config

phantomjs

- Full browser-engine
- Command-line
- Javascript API

Website 1

Config.js

- 120 lines of JS code
- Using jQuery
- Including logging, comments and 20 URL's
- Has to be changed when web-page layout changes
- Multiple config-files can be sent to pjscrape

URL's

- PhantomJS: http://phantomjs.org/
- Pjscrape.js: http://nrabinowitz.github.com/pjscrape/
- jQuery: http://api.jquery.com/