Intro to Programming for Public Policy Week 7
The Web and Scraping

Eric Potash

May 7, 2018

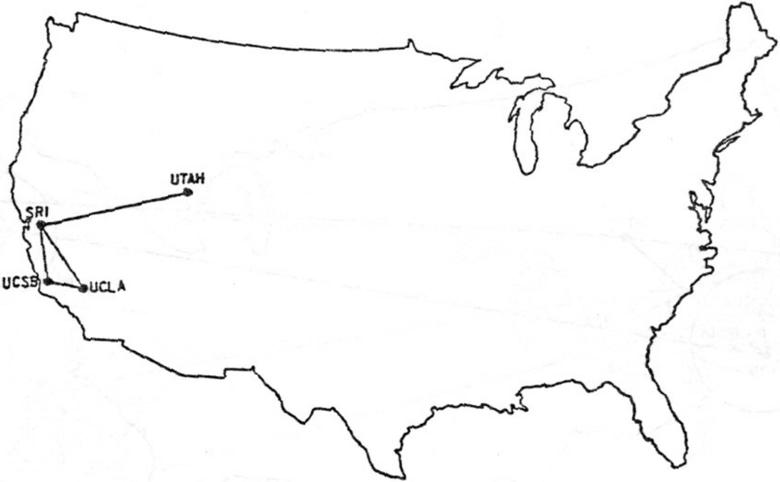
The Internet

The Internet and World Wide Web

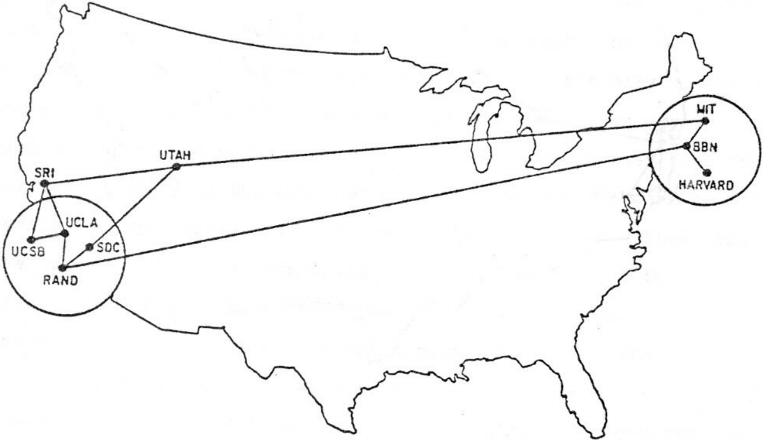
- The **Internet** is a *physical network* of cables and routers, and a set of protocols for moving information across that network.
- The **World Wide Web** (WWW) is an *information space* on the Internet. It combines several concepts:
  - Uniform Resource Locator (url)
  - Hypertext Transfer Protocol (http)
  - Hypertext Markup Language (html)

The Internet

- Started as a Defense Advanced Research Projects Agency (DARPA) ARPANET project to interconnect computers
- First transmission between nodes at UCLA in 1969; by 1970 reached across the US to Boston.
- Transmission Protocol/Internet Protocol (TCP/IP) developed during the 1970s
- ARPANET declared "operational" in 1975 and transferred to military
- Ethernet (transmission on wires) standard written 1981
- During 1980s, shift to National Science Foundation
  - NSFNET provided the backbone of Internet from 1985 to 1995
- Internet backbone privatized under Clinton in 1994



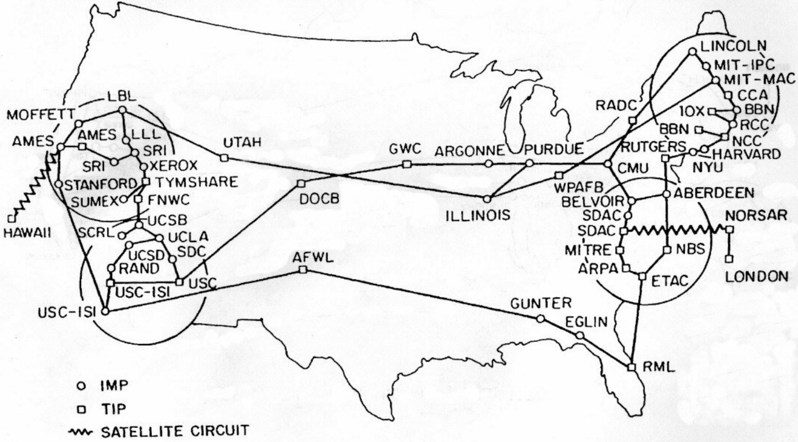
The ARPANET in December 1969



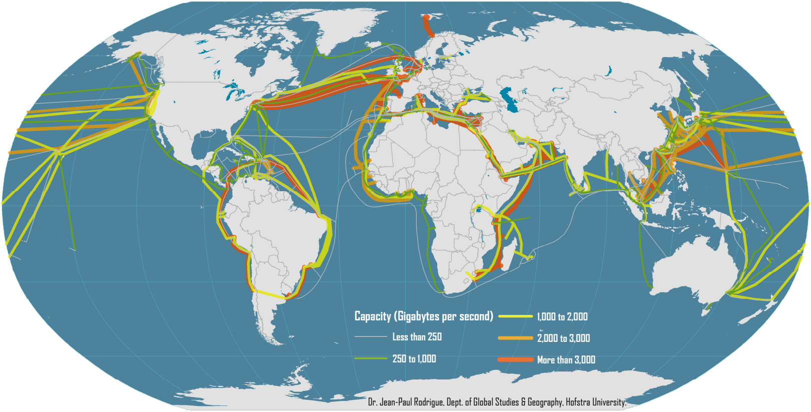
MIT
BBN
HARVARD
UTAH
SRI
UCLA
UCSB
SDC
RAND

ARPA NETWORK, GEOGRAPHIC MAP

JUNE 1975



Global Submarine Cable Network



Dr. Jean-Paul Rodrigue, Dept. of Global Studies & Geography, Hofstra University.

Source: Dataset encoded by Greg Mahlknecht, http://www.cablemap.info

Uniform Resource Locator

URLs are a system of globally unique identifiers for resources on the Web and elsewhere.

`scheme://host[:port]/path[?query][#fragment]`

- `scheme` is a protcol such as `http`, `https`, etc.
- `host` is something like `google.com` or `localhost`
- `:port` is optional, allows a single host to have separate websites
- `path` is the path to a particular resource like `index.html`
- More on queries later

Hypertext Transfer Protocol (HTTP)

HTTP Overview

- HTTP is the core communications protocol for retrieving data
- Consists of messages– requests and responses– sent between a client and a server

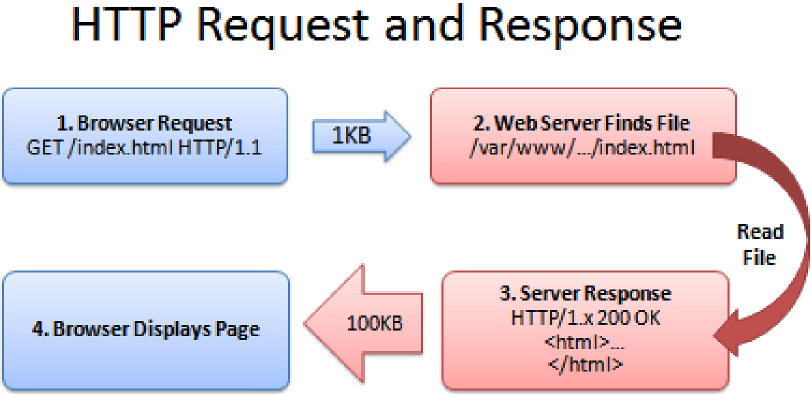


Figure 1: Source: opensourcevaristy.com

HTTP Request

```
GET /index.html HTTP/1.1
User-Agent: Mozilla/5.0
...
```

- First line contains:
  - HTTP method, here `GET`
  - Requested URL
  - HTTP version
- Rest of request may contain:
  - User-Agent: description of the client
    - Used e.g. to determine whether to serve mobile website version

HTTP Methods

- `GET`
  - Most common method, used to get data
- `POST`
  - Used to send data to server, e.g. form entries, search queries

HTTP Responses

```
HTTP/1.1 200 OK
...
```

- First line contains:
  - HTTP version
  - HTTP response code
- Rest of response contains:
  - Additional headers: `Server`, `Content-Type`, etc.
  - Requested Content

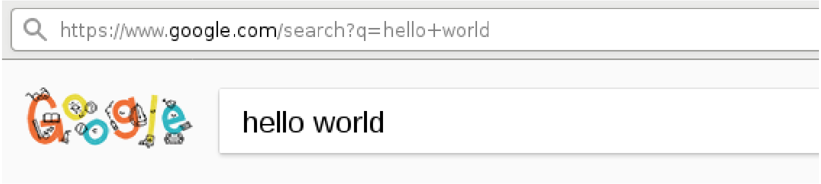
HTTP Response Codes

- 1xx: Informational
- 2xx: Success
  - 200: OK
- 3xx: Redirection
  - 301: Redirect
- 4xx: Errors
  - 404: File not found
  - 403: Forbidden

HTTP GET Request Parameters

```
/index.php?name1=value1&name2=value2
```

- Query string with parameters sent in the URL of a GET request
- Parameter names and values are like a python dictionary
- Shouldn't use with sensitive data

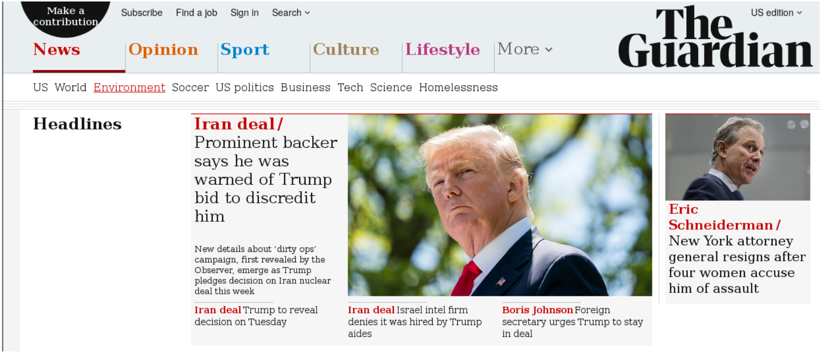


https://www.google.com/search?q=hello+world

hello world

Hypertext Markup Language (HTML)

Web browser



Make a contribution

Subscribe Find a job Sign in Search

US edition

The Guardian

News Opinion Sport Culture Lifestyle More

US World Environment Soccer US politics Business Tech Science Homelessness

Headlines

Iran deal / Prominent backer says he was warned of Trump bid to discredit him

New details about 'dirty ops' campaign, first revealed by the Observer, emerge as Trump pledges decision on Iran nuclear deal this week

Iran deal Trump to reveal decision on Tuesday

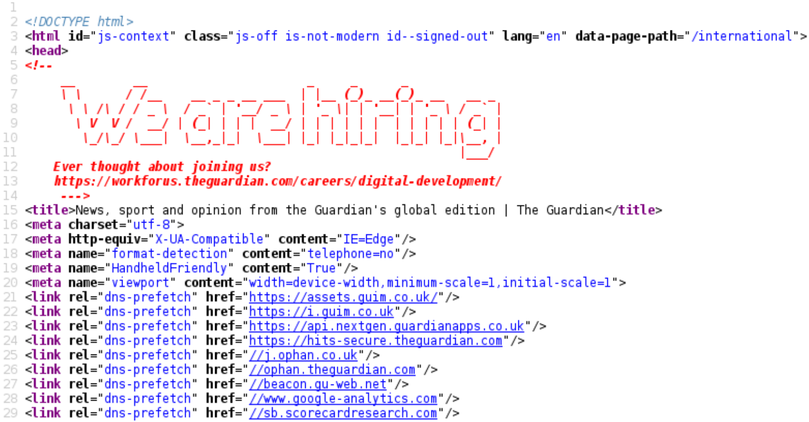
Iran deal Israel intel firm denies it was hired by Trump aides

Boris Johnson Foreign secretary urges Trump to stay in deal

Eric Schneiderman / New York attorney general resigns after four women accuse him of assault

Figure 2: guardian.com rendered in a web browser

HTML sourcecode



```html

<!DOCTYPE html>
<html id="js-context" class="js-off is-not-modern id--signed-out" lang="en" data-page-path="/international">
<head>
<!--
     __        __                     _     _      _             
     \ \      / /__    __ _ _ __ ___  | |__ (_)_ __(_)_ __   __ _ 
      \ \ /\ / / _ \  / _` | '__/ _ \ | '_ \| | '__| | '_ \ / _` |
       \ V  V /  __/ | (_| | | |  __/ | | | | | |  | | | | | (_| |
        \_/\_/ \___|  \__,_|_|  \___| |_| |_|_|_|  |_|_| |_|\__, |
                                                            |___/ 
    Ever thought about joining us?
    https://workforus.theguardian.com/careers/digital-development/
     --->
<title>News, sport and opinion from the Guardian's global edition | The Guardian</title>
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=Edge"/>
<meta name="format-detection" content="telephone=no"/>
<meta name="HandheldFriendly" content="True"/>
<meta name="viewport" content="width=device-width,minimum-scale=1,initial-scale=1">
<link rel="dns-prefetch" href="https://assets.guim.co.uk/"/>
<link rel="dns-prefetch" href="https://i.guim.co.uk"/>
<link rel="dns-prefetch" href="https://api.nextgen.guardianapps.co.uk"/>
<link rel="dns-prefetch" href="https://hits-secure.theguardian.com"/>
<link rel="dns-prefetch" href="//j.ophan.co.uk"/>
<link rel="dns-prefetch" href="//ophan.theguardian.com"/>
<link rel="dns-prefetch" href="//beacon.gu-web.net"/>
<link rel="dns-prefetch" href="//www.google-analytics.com"/>
<link rel="dns-prefetch" href="//sb.scorecardresearch.com"/>
```

Figure 3: guardian.com HTML sourcecode

HTML overview

- Language that webpages are written in
- Consists of *tags*
- Most tags come in pairs (opening and closing):
  - `<html></html>`
  - `<head></head>`
  - `<body></body>`
  - `<a></a>`
- Some do not:
  - `<img>`
  - `<br>`
- Whitespace doesn't matter (unlike Python)

Basic HTML webpage

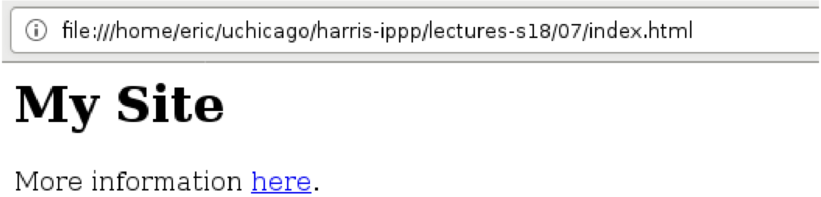
```html
<html>

<head>
    <title>My Title</title>
</head>

<body>
<h1>My Site</h1>
<p>
More information <a href="main.html">here</a>.
</p>

</body>
</html>
```

Basic webpage



file:///home/eric/uchicago/harris-ippp/lectures-s18/07/index.html

My Site

More information here.

Figure 4: Webpage rendered in browser

HTML links

```
<a href="main.html">here</a>
```

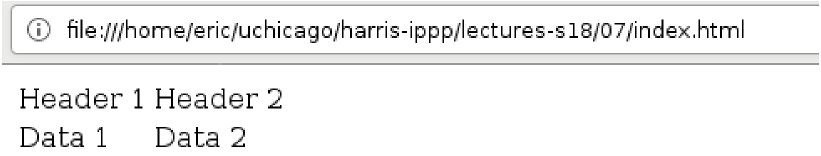
- `a` is the tag
- `here` is the content
- `href` is an attribute whose value is `main.html`
    - attributes and their values form something like a dictionary

More HTML tags

- `<title>`: page title
- `<body>`: page body
- `<h1>`: largest header
- `<h6>`: smallest header
- `<p>`, `<br>`: paragraph, break
- `<em>`, `<b>`: emphasis (italics), bold
- `<ol>`, `<ul>`, `<li>`: (Un-)ordered list, list element
- `<a href="">`: Hyperlink
- `<table>`, `<tr>`, `<th>`, `<td>`: table row, header, and cell
- `<img src="">`: image
- `<div>`: block

HTML table

```
<table>
  <tr>
    <td>Header 1</td>
    <td>Header 2</td>
  </tr>
  <tr>
    <td>Data 1</td>
    <td>Data 2</td>
  </tr>
</table>
```



file:///home/eric/uchicago/harris-ippp/lectures-s18/07/index.html

Header 1	Header 2
Data 1	Data 2

Other web technologies

- Javascript
  - `.js` files sourced or code directly embedded in `<script>...</script>` tags
  - Used to make content dynamic
- CSS
  - `.css` files or code embedded in `<style></style>` tags
- Server-side languages
  - Ruby, PHP, Python (Django), Java, etc.

Web Scraping

Web scraping overview

- Scraping is the process of programmatically extracting information from websites
- Anything that you can view in a web browser can *potentially* be scraped

Why scrape?

Some websites offer services (APIs) that allow you to get data directly. So why scrape?

- Not all websites provide an API
- Not all of a website's content is available through its API
- APIs often use tokens to limit the amount of data that can be requested
    - With scraping there is, in principle, no limit

When to scrape?

Scraping is good for:

- Downloading all `.mp3` files linked from a site
- Constructing a pandas DataFrame from a table on Wikipedia
- Parsing articles from a small news website

Scraping is not ideal for:

- Extracting information from your emails (python has POP/IMAP libraries)
- Analyzing tweets (Twitter provides an API)

Ethical considerations

- Credit all sources
  - Publishing scraped content can be a copyright violation
- Don’t overload websites
  - Most sites will block you before you can do this
- Obey `robots.txt`
  - Most sites have a file describing which areas of the site are prohibited from being scraped
- You are not anonymous on the web
  - Unless you take explicit steps (VPN, Tor, etc.) to do so

Scraping from the command line

- We already saw one basic way to scrape using the command line program `curl`:

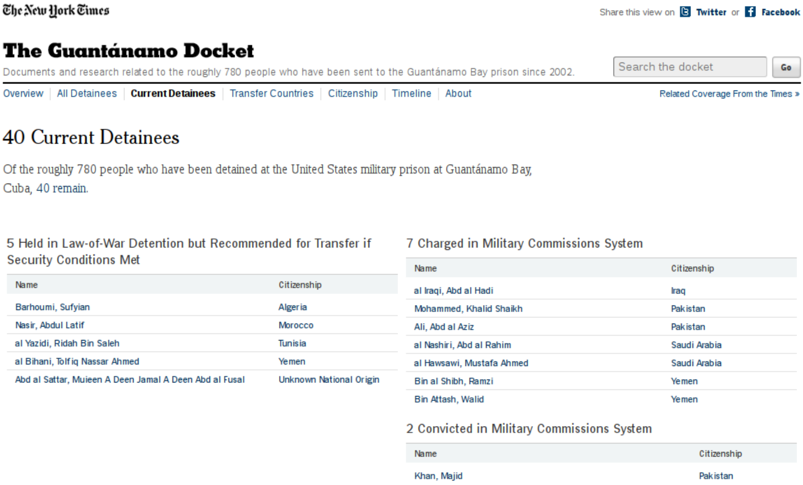
```
$ curl https://data.cityofchicago.org/api/views/xzkq-xp
  -o salaries.csv
```

- There's a similar program called `wget`:

```
$ wget https://data.cityofchicago.org/api/views/xzkq-xp
```

- The advantage of `wget` is that it can follow links and download them, too

Example site to scrape



The New York Times
Share this view on Twitter or Facebook

The Guantánamo Docket

Documents and research related to the roughly 780 people who have been sent to the Guantánamo Bay prison since 2002.

Search the docket
Go
Overview | All Detainees | Current Detainees | Transfer Countries | Citizenship | Timeline | About
Related Coverage From the Times »

40 Current Detainees

Of the roughly 780 people who have been detained at the United States military prison at Guantánamo Bay, Cuba, 40 remain.

5 Held in Law-of-War Detention but Recommended for Transfer if Security Conditions Met

Name	Citizenship
Barhoumi, Sufyian	Algeria
Nasir, Abdul Latif	Morocco
al Yazidi, Ridah Bin Saleh	Tunisia
al Bihani, Tolfiq Nassar Ahmed	Yemen
Abd al Sattar, Muieen A Deen Jamal A Deen Abd al Fusal	Unknown National Origin

7 Charged in Military Commissions System

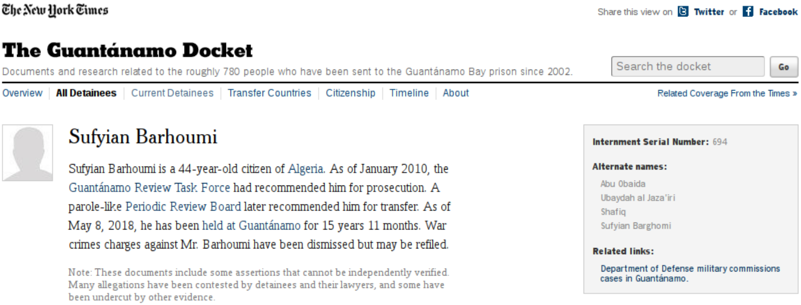
Name	Citizenship
al Iraqi, Abd al Hadi	Iraq
Mohammed, Khalid Shaikh	Pakistan
Ali, Abd al Aziz	Pakistan
al Nashiri, Abd al Rahim	Saudi Arabia
al Hawsawi, Mustafa Ahmed	Saudi Arabia
Bin al Shibh, Ramzi	Yemen
Bin Attash, Walid	Yemen

2 Convicted in Military Commissions System

Name	Citizenship
Khan, Majid	Pakistan

https://www.nytimes.com/interactive/projects/guantanamo/detainees/current

Subpage



The Guantánamo Docket

Documents and research related to the roughly 780 people who have been sent to the Guantánamo Bay prison since 2002.

Sufyian Barhoumi

Sufyian Barhoumi is a 44-year-old citizen of Algeria. As of January 2010, the Guantánamo Review Task Force had recommended him for prosecution. A parole-like Periodic Review Board later recommended him for transfer. As of May 8, 2018, he has been held at Guantánamo for 15 years 11 months. War crimes charges against Mr. Barhoumi have been dismissed but may be refiled.

Note: These documents include some assertions that cannot be independently verified. Many allegations have been contested by detainees and their lawyers, and some have been undercut by other evidence.

Internment Serial Number: 694

Alternate names:

Abu Obaida
Ubaydah al Jaza'iri
Shafiq
Sufyian Barghomi

Related links:

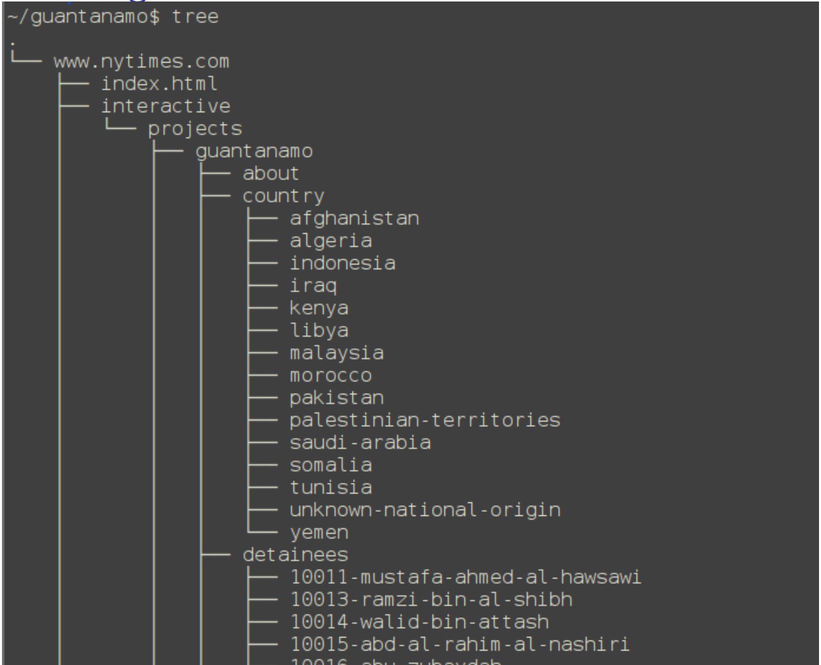
Department of Defense military commissions cases in Guantánamo.

Recursive wget

```
$ wget -r -l 1 https://www.nytimes.com/interactive/projects
...
2018-05-07 23:18:40 (8.70 MB/s) - ‘www.nytimes.com/interact

FINISHED --2018-05-07 23:18:40--
Total wall clock time: 1.0s
Downloaded: 66 files, 1.3M in 0.09s (14.3 MB/s)
```

Recursive wget results



```
~/guantanamo$ tree
.
└── www.nytimes.com
    ├── index.html
    ├── interactive
    │   └── projects
    │       ├── guantanamo
    │       │   ├── about
    │       │   ├── country
    │       │   │   ├── afghanistan
    │       │   │   ├── algeria
    │       │   │   ├── indonesia
    │       │   │   ├── iraq
    │       │   │   ├── kenya
    │       │   │   ├── libya
    │       │   │   ├── malaysia
    │       │   │   ├── morocco
    │       │   │   ├── pakistan
    │       │   │   ├── palestinian-territories
    │       │   │   ├── saudi-arabia
    │       │   │   ├── somalia
    │       │   │   ├── tunisia
    │       │   │   ├── unknown-national-origin
    │       │   │   └── yemen
    │       │   ├── detainees
    │       │   │   ├── 10011-mustafa-ahmed-al-hawsawi
    │       │   │   ├── 10013-ramzi-bin-al-shibh
    │       │   │   ├── 10014-walid-bin-attash
    │       │   │   ├── 10015-abd-al-rahim-al-nashiri
    │       │   │   │   10016-abu-zubaydah
```