

Methods for unstructured data

Lecture 2: Web scraping

Helge Liebert

Roadmap and goals

- Use processing of web data to introduce various ideas along the way.
- Accessing data on the web: APIs.
- Gathering (semi-structured) web data and transforming it into structured data (*“web scraping”*).
- Give you a general understanding how web scrapers work.
- Foster an understanding of the development process.
- Point you towards the necessary tools so you can write your own.

1. Introduction
2. Static and dynamic websites
3. Application Programming Interfaces
4. The Document Object Model
5. Examples
6. Assignment


Guiding problem

- How to turn unstructured into structured data?
- Consider a situation where
 - You want to get data from the internet.
 - The data is in unstructured/semi-structured form.
 - Possibly embedded in a website.
 - You want to transform it into a differently structured format for further use.
 - You need to filter the available information.


[illegible]

...and this

NSCB - Good Governance
Our Campaigns - Candidates
www.nscb.gov.ph/ggi/details_prov.asp



REPUBLIC OF THE PHILIPPINES
PHILIPPINE STATISTICS AUTHORITY
SOLID • RESPONSIVE • WORLD-CLASS


Good Governance Index

GGI Indicators for the province of Misamis Oriental

Indicator	2005		2008	
	Value	Rank in	Value	Rank in
Economic Governance Index	107.08	37	86.12	71
Total Financial Resources Generated per Capita (in Millions)	171.05	45	233.43	50
Per Capita Tax and Non-Tax (in Millions)	71.66	1	73.87	7
Per Capita Total Deposits	122.61	4	N/A	N/A
Per Capita Expenditure on Social Services	212.80	22	90.22	67
Unemployment Rate	125.56	47	125.56	47
Underemployment Rate	55.31	53	55.31	53
Inflation Rate	66.37	14	37.46	37
Poverty Gap	87.56	34	72.64	42
Poverty Incidence	84.90	27	88.10	31
Political Governance Index	103.24	13	99.86	18
Crime Solution Efficiency Rate	103.91	8	96.76	23
Voter's Turnout Rate	102.56	31	102.56	31
Administrative Governance Index	128.00	23	154.38	13
A. Education Index	80.79	68	107.25	52
Elementary Teacher to Pupil Ratio	100.07	55	96.93	54
High School Teacher to Pupil Ratio	100.93	45	99.86	53
No. of Public Elem. Schools per 1000 School Age Population	56.02	85	96.64	45
No. of Public High Schools per 1000 School Age Population	51.16	70	118.54	60
Enrollment in Gov't. Elem. Sch. per 1000 School Age Population	57.08	75	95.78	9
Enrollment in Gov't. HS. Sch. per 1000 School Age Population	61.13	63	140.90	8
Cohort Survival Rate (Elementary)	93.57	46	83.99	49
Cohort Survival Rate (High School)	79.85	32	85.29	48
Elementary Pupil - Classroom Ratio	102.75	58	103.88	54
High School Pupil - Classroom Ratio	105.37	61	140.74	53
B. Health Index	229.66	19	212.88	21
Total Health Personnel per 1000 Population	193.11	19	170.55	23
% Birth less the 2500 grams	478.47	12	393.89	21
% of Household with access to safe water	109.09	48	131.13	1
% Barangay Health Station per 100,000 Population	137.99	30	151.93	21
C. Power and ICT Index	76.26	35	144.00	1
Power Index	124.83	1	124.83	1
ICT Index (Telephone Density per 1000 Population)	27.89	41	157.44	1
Good Governance Index	116.27	33	117.31	42

1 = highest/best, out of 79 provinces

Good Governance Index
[Main Page](#)
[Database](#)
[Technical Notes](#)
[Press Release](#)

...to this.

master.csv - LibreOffice Calc																							
File Edit View Insert Format Tools Data Window Help																							
Liberation Sans 10																							
A1																							
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R						
id	name	jahr	land	provinz	gruppe	popcat	pop	ehe	geb	lebgebo	lebgebo	todgebo	todgebo	geb	gebo	todgeb	todgebshare						
1	Aachen	1927	Preussen	Rheinprovinz	A	pop >= 100,000	156360	1423	2753	2366	387	80	7	2840	2446	87	3.0633902						
2	Aachen	1928	Preussen	Rheinprovinz	A	pop >= 100,000	150991	1446	2723	2260	463	66	17	2806	2326	83	2.9579473						
3	Aachen	1929	Preussen	Rheinprovinz	A	pop >= 100,000	155542	1558	2635	2192	443	49	23	2707	2241	72	2.659771						
4	Aachen	1930	Preussen	Rheinprovinz	A	pop >= 100,000	154634	1501	2635	2189	446	46	22	2703	2235	68	2.5157232						
5	Aachen	1931	Preussen	Rheinprovinz	A	pop >= 100,000	154400	1408	2429	2030	399	53	22	2504	2083	75	2.9952078						
6	Aachen	1932	Preussen	Rheinprovinz	A	pop >= 100,000	153834	1426	2305	1968	337	46	20	2371	2014	66	2.7836356						
7	Aachen	1933	Preussen	Rheinprovinz	A	pop >= 100,000	162990	1616	2371	2028	343	54	23	2448	2082	77	3.1454248						
8	Aachen	1934	Preussen	Rheinprovinz	A	pop >= 100,000	163839	1942	2841	2451	490	76	8	3025	2527	84	2.7768595						
9	Aachen	1935	Preussen	Rheinprovinz	A	pop >= 100,000	164180	1570	3048	2516	532	54	21	3123	2570	75	2.4015369						
10	Aachen	1936	Preussen	Rheinprovinz	A	pop >= 100,000	164105	1502	3012	2382	630	43	25	3080	2425	68	2.2077923						
12	Ahlen	1927	Preussen	Westfalen	D	15,000 <= pop < 30,000	23956	239	606	596	10	27	1	634	623	28	4.4164038						
13	Ahlen	1928	Preussen	Westfalen	D	15,000 <= pop < 30,000	24703	266	625	621	4	18	0	643	639	18	2.7993779						
14	Ahlen	1929	Preussen	Westfalen	D	15,000 <= pop < 30,000	25043	227	624	609	15	26	0	650	635	26	4						
15	Ahlen	1930	Preussen	Westfalen	D	15,000 <= pop < 30,000	25226	202	568	548	20	25	1	594	573	26	4.3771043						
16	Ahlen	1931	Preussen	Westfalen	D	15,000 <= pop < 30,000	25373	191	508	493	25	34	0	542	517	34	6.2730627						
17	Ahlen	1932	Preussen	Westfalen	D	15,000 <= pop < 30,000	25549	200	550	514	36	13	1	564	527	14	2.4822695						
18	Ahlen	1933	Preussen	Westfalen	D	15,000 <= pop < 30,000	25153	301	478	437	41	15	2	495	452	17	3.4343433						
19	Ahlen	1934	Preussen	Westfalen	D	15,000 <= pop < 30,000	25700	283	658	606	52	19	2	679	625	21	3.0927835						
20	Ahlen	1935	Preussen	Westfalen	D	15,000 <= pop < 30,000	25937	212	661	621	40	17	0	678	638	17	2.5073745						
21	Ahlen	1936	Preussen	Westfalen	D	15,000 <= pop < 30,000	26104	228	673	625	48	11	1	685	636	12	1.7518249						
22	Allenstein	1927	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	39315	241	892	859	33	14	3	909	873	17	1.870187						
23	Allenstein	1928	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	39323	293	943	898	45	24	2	969	922	26	2.6831784						
24	Allenstein	1929	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	39314	267	959	900	56	21	3	924	855	25	2.5406504						
25	Allenstein	1930	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	39345	283	862	810	52	14	5	881	824	19	2.1566403						
26	Allenstein	1931	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	39876	301	884	836	48	17	4	905	853	21	2.320442						
27	Allenstein	1932	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	40078	323	837	796	41	15	1	853	811	16	1.8757327						
28	Allenstein	1933	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	43079	340	890	854	36	14	4	908	868	18	1.9823788						
29	Allenstein	1934	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	43506	405	1062	1015	47	28	4	1094	1043	32	2.9250457						
30	Allenstein	1935	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	43881	363	1118	1051	67	21	4	1138	1067	20	1.7574693						
31	Allenstein	1936	Preussen	Ostpreussen	C	30,000 <= pop < 50,000	44566	364	1150	1037	82	16	5	1145	1058	26	2.2707424						
32	Altena	1927	Preussen	Westfalen	D	15,000 <= pop < 30,000	15931	131	240	226	14	5	0	245	231	5	2.0408163						
33	Altena	1928	Preussen	Westfalen	D	15,000 <= pop < 30,000	16333	163	244	229	15	11	2	257	240	13	5.0583658						
34	Altena	1929	Preussen	Westfalen	D	15,000 <= pop < 30,000	16464	139	257	248	9	6	1	264	254	7	2.6515152						
35	Altena	1930	Preussen	Westfalen	D	15,000 <= pop < 30,000	16498	139	242	220	22	16	1	259	236	17	5.6537064						
36	Altena	1931	Preussen	Westfalen	D	15,000 <= pop < 30,000	16407	138	196	186	10	5	2	203	191	7	3.4482758						
37	Altena	1932	Preussen	Westfalen	D	15,000 <= pop < 30,000	16115	122	192	176	15	9	0	201	185	9	4.477612						
38	Altena	1933	Preussen	Westfalen	D	15,000 <= pop < 30,000	16133	138	189	174	15	6	1	196	180	7	3.5712695						
39	Altena	1934	Preussen	Westfalen	D	15,000 <= pop < 30,000	16246	162	258	223	35	8	2	268	231	10	3.7313433						

Find

Find All Match Case

Sheet 1 of 1

Default

Sum=0

110%

Goal: Automation

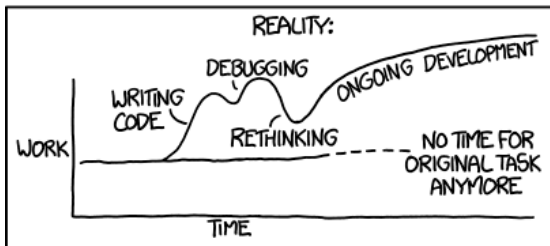
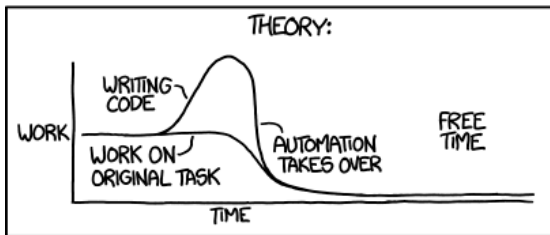
- Digitalization offers exciting data for research. But: *Data is messy*.
- Gathering or processing data often involves repetitive manual tasks.
- Disadvantages:
 - Manual tasks are often not well documented or reproducible ex post.
 - Manual work is frustrating and a huge time-sink.
 - Manual work may not be feasible with large data.

➡ Automation helps!

- Frees you to engage in other work.
- You learn new things.
- Should you encounter the same class of problem in the future, you already have a solution at hand.

Automation

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



Getting started—things to consider before you begin

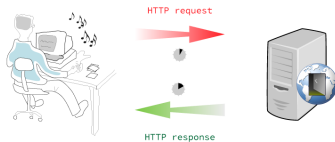
- Pick up the phone and try to get the data directly.
- Search if somebody has already faced the same or a similar problem.
- Does the site or service provide an API that you can access directly?
- Is there a wrapper for it?
- Is the website only online for a limited time? Do you want an original snapshot as a backup? Is it more convenient to filter your data offline?

Static and dynamic websites

Static vs. dynamic websites

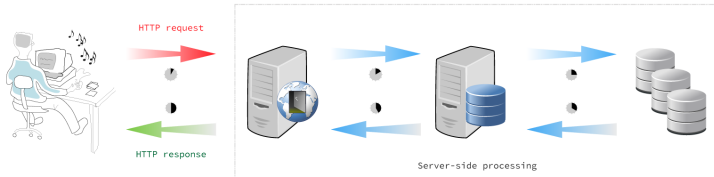
Scheme A

Static Website



Scheme B

Dynamic Website



Save an offline copy

- Use the shell utilities `wget` or `curl` to download the complete site.
- Also useful if you just want a set of files (e.g. pdf documents) from the same site directory.
- Convenient for static sites of limited size.
- Infeasible for large sites or sites that create content dynamically.

Examples

- Simple http GET request.
`wget http://www.google.com`
- Recursively download a website.
`wget -r http://www.some-site.com/some-subdir/`
- Download all pdfs from a site.
`wget -r -A.pdf http://url-to-webpage-with-pdfs/`
- Mirror a site offline and convert links for local browsing.
`wget --mirror -p --convert-links -p ./local-dir
http://target-website.com`

Application Programming Interfaces

- Data providers often offer Web APIs (*Application Programming Interface*) to access data.
- Allow programmable access to data via a defined set of HTTP messages. Similar to visiting a website: you specify a URL and information is sent to your machine.
- With a website, you receive code interpreted by your browser (HTML, CSS, JavaScript). With an API, you receive data.
- Usually in JSON (*JavaScript Object Notation*) or XML (*Extensible Markup Language*) format.

Web APIs

- Often just two steps:
 1. Construct the URL query that serves as the API request.
 2. Process the response message the API sends back.
- Examples:
 - <https://api.kivaws.org/v1/loans/newest.html>
 - <https://api.kivaws.org/v1/loans/newest.json>
 - <https://api.kivaws.org/v1/loans/search.json?sector=Agriculture&country=VN>
 - <https://www.theyworkforyou.com/api/getMPs?&key=someapikeyhere&output=js>
- Libraries may offer wrappers for APIs: **WDI**, **wbstats**, **twfy**, **pvsR**, Google Maps, OpenStreetMap/OSRM, ...
- Sometimes it is possible to reverse engineer a site's internal API rather than scraping the HTML.

The Document Object Model

HTML and the Document Object Model

- Extracting information from the web requires a basic understanding of HTML and the associated Document Object Model (DOM).
- HTML elements provide the structure and content of web pages.
- Consist of **<start>** and **</end>** tags, with content in between.
`<tagname>Content here</tagname>`
- A page consists of nested elements.
- The **html** element is the outer-most element, nesting the **head** and **body** elements, which in turn have nested elements.
- Nesting structure of elements can be represented by a tree (DOM).

Document Object Model

- The DOM is a programming interface for HTML and XML documents.
- Provides a structured representation of the document.
- A document as a group of nodes, each node representing a part of the document.
- Allows programmatic access to the tree to change the structure, style and content of the document.
- Connects web pages to scripts or programming languages.

A simple HTML page

A simple HTML page:

```
<html>
  <head>
    <title>My Web Page</title>
  </head>
  <body>
    <h1>Welcome To My Web Page</h1>
  </body>
</html>
```

How a browser renders this page:

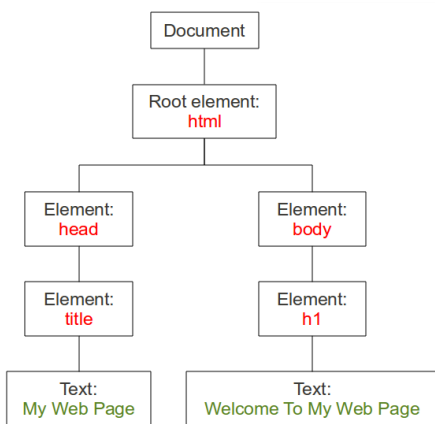


HTML and the DOM

A simple HTML page:

```
<html>
  <head>
    <title>My Web Page
  </title>
</head>
<body>
  <h1>Welcome To My Web Page
</h1>
</body>
</html>
```

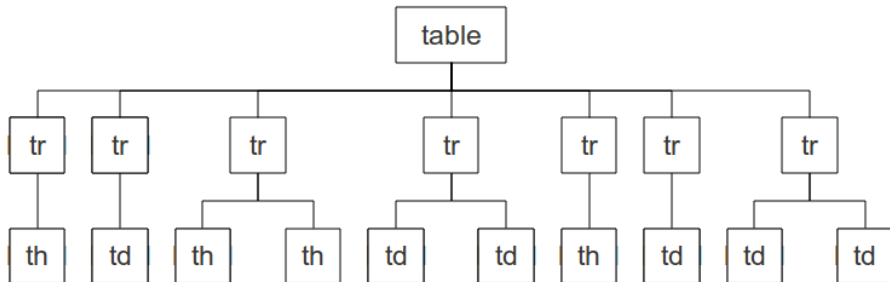
Corresponding node tree:



DOM node trees

- HTML DOM views a document as a tree structure called node tree.
- Everything in an HTML document is a node.
 - The entire document is a document node
 - Every HTML element is an element node
 - Every HTML attribute is an attribute node
 - Text content in the HTML elements is a text node
- Nodes can be accessed through the tree.
- Nodes may be assigned unique id attributes.

Example: An HTML table element



- Tables are represented by a top-level **table** element.
- The **table** element nests **tr** (*table row*) elements.
- These nest **th** (*table header*) and **td** (*table data*) element cells.

HTML and the DOM

- HTML tags can have attributes and text content.

```
<tag attribute="value" attribute2="value">Text content.</tag>
```

- Example page:

```
<html>
  <head>
    <title>My Web Page</title>
  </head>
  <body>
    <h1>Welcome To My Web Page</h1>
    
    <a href="pagelink.html" id="pageid">Check this other page.</a>
  </body>
</html>
```

Government and bureaucrats tend to show an insensitivity to their parents and their recent suffering from a lost child, and produce broad disclaimers in the IMR reports that the information has not been properly reported, resulting in these discrepancies. Little has been done to address the underlying structural problems of the vital registry systems in respect to the lack of reporting from parents in rural areas who have created a gap between the official and popular meanings of child death.^[116] It is also argued that the bureaucratic separation of vital deaths resulting from cultural death rituals is to blame for the inaccuracy of the infant mortality rate (IMR). Vital death registers often fail to recognize the cultural implications and importance of infant deaths. It is not to be said that vital registry systems are not an accurate representation of a region's socio-economic situation, but it is only the case if those few statistics are valid, which is unfortunately not always the circumstance. "Popular death reporters" is an alternative method for collecting and processing statistics on infant and child mortality may benefit from "popular death reporters" who are culturally linked to infants may be able to provide more accurate observations on the incidence of infant mortality.^[117] According to ethnographic data, "popular death reporters" refers to people who hold more knowledge of childbirth, including the grower-gilder, gablekeeper, midwife, popular healers etc. — all key participants in mortuary rituals.^[118] By conducting the mortuary death rituals, registers, and asking "popular death reporters" can increase the validity of child mortality rates, but there are many barriers that can reflect the validity of our statistics of infant mortality. One of three barriers are political decisions. Numbers are exaggerated when international funds are being doled out and underrepresented during relief.^[119]

The humanitarian separation of vital death reporting and cultural death rituals stems in part due to structural violence^[120] Individuals living in rural areas of Brazil need to travel large capital villages and travel in order to report infant birth to a Brazilian Association League office. The negative financial aspects deter registration, so often individuals are of lower income and cannot afford such expenses.^[121] Similar to the lack of both reporting, families in rural Brazil face difficult choices about installing visiting structural arrangements when choosing to report infant mortality. Financial constraints such as reliance on local implementations may also lead to skewed infant mortality data.^[122]

In developing countries such as Brazil the deaths of impoverished infants are regularly unrecorded into the countries vital registry system. This causes a false statistically. Culturally violent and contextual solutions can be used to ground the meaning of mortality from a statistical standpoint. In northeastern Brazil they have accomplished this standpoint while conducting an ethnographic study combined with an alternative method by surveying infant deaths.^[123] These types of techniques can develop quality ethnographic data that will ultimately help the amount of the magnitude of infant mortality in the region. Political economic reasons have been seen to skew the infant mortality data in the past when government Canada declared his presidency campaign on reducing the infant mortality rate during his term in office. By using this new way of surveying, these instances can be minimized and removed, overall creating accurate and sound data.^[124]

Epidemiology

See also: List of countries by infant mortality rate

For the world, and for both less developed countries (LDC) and more developed countries (MDC), IMR declined significantly between 1960 and 2001. According to the *State of the World's Mothers* report by Save the Children, the world IMR declined from 126 in 1960 to 57 in 2001.^[125] However, IMR was, and remains, higher in LDCs. In 2001, the IMR for LDCs (91) was about 10 times as large as it was for MDCs (8). On average, for LDCs, the IMR is 17 times as higher than that of MDCs. Also, while both LDCs and MDCs made significant reductions in infant mortality rates, reductions among less developed countries are, on average, much less than those among the more developed countries. The top and bottom five countries by this measure (taken from *The World Factbook* 2012 estimated)^[126] are shown below.

Rank	Country	Infant mortality rate (deaths/1,000 live births)
1	Alghanistan	121.83
2	Niger	109.98
3	Mali	109.98
4	Somalia	102.72
5	Central African Republic	87.17
216	Sweden	2.74
217	Singapore	2.65
220	Bermuda	2.47
221	Japan	2.21
222	Morocco	1.80

According to O'Neil, Gertzel, Puffer and Savelbergh's Barth's histories, however, age subject to a number of errors, including omission of deaths and age misreporting errors.^[127]

United States

The infant mortality rate in the US decreased by 2.3% to a historic low of 582 infant per one 100,000 live births in 2014.^[128]

Of the 27 most developed countries, the U.S. has the highest Infant Mortality Rate, despite spending much more on health care per capita^{[129][130]}. Significant racial and socio-economic differences in the United States affect the IMR. In contrast with other developed countries, which have more homogeneous populations, in particular, IMR varies greatly by race in the U.S. The average IMR for the white population is therefore not a fair representation of the value variations that exist between segments of the population. Many theories have been explained as to why these racial differences exist with socio economic factors usually coming out as a reasonable explanation. However, more studies have concluded against this notion, and the largest advancement is around the idea of stress and how it affects pregnancy.^[131]

In the 1950s, the infant mortality rate in the United States was estimated at 29.8 per 1,000 babies born to whites and 34.0 per 1,000 African Americans, but rates have significantly declined in the West in modern times. This declining rate has been mainly due to modern improvements in basic health care, technology, and medicine.^[132] In the last century, the infant mortality rate has decreased by 93%.^[133] Overall, the rates have decreased drastically from 20 deaths in 1970 to six deaths in 2003 per every 1,000 live births. In 2003, the leading causes of infant mortality in the United States were congenital anomalies, disorders related to immaturity, SIDS, and maternal complications. Babies born with low birth weight increased proportionally to 19.2%, this reflected the amount of low birth weights concluding that 12.4% of births from mothers were low birth weight compared with 7.7% of such births from non-mothers.^[134]

According to the New York Times, "the main reason for this high rate is preterm delivery, and there was a 15% increase in such births from 2000 to 2006." Between 2007 and 2011, however, the preterm birth rate has decreased every year. In 2011 there was an 11.7% rate of babies born before the 37th week of gestation, down from a high of 12.29% in 2006.^[135]

Economic disparities on labor and delivery and neonatal care are relatively high in the United States. A conventional birth averages \$899,775 with a C-section costing \$11,554.^[136] Preterm births in the US have been estimated to cost \$51,700 per child, with a total yearly cost of \$26.2 billion^[137] Despite this spending, several reports state that infant mortality rate in the United States is significantly higher than in other developed nations.^{[138][139]} Estimates vary; the *CIA's World Factbook* ranks the US 58th internationally in 2014, with a rate of 6.17, while the UN figures from 2009-2011 put the US 34th.

African-American differences in measurement could play a substantial role in the disparity between the US and other nations. A non-white live birth in the US could be regarded as a sibling in similarly developed nations like Japan, Sweden, Norway, India, the Netherlands, and France – thereby reducing the infant death count.^[140]

Neonatal intensive care is also likely to be applied in the US to marginally viable infants, although such interventions have been found to increase both costs and disability. A first-year investigation of the *Born Alive Infant Protection Act* of 2002 based on universal resuscitation of infants born between 20–23 weeks increased the neonatal spending burden by \$153.1 million while simultaneously decreasing early-life expenditures by \$29.3 million.^[141]

The vast majority of research conducted in the late twentieth and early twenty-first century indicates that African-Americans are more than twice as likely to die in a first year of life than white infants. Although following a decline from 13.63 to 11.46 deaths per 1000 live children from 2005 to 2010, non-Hispanic black children continued to report a rate 2.2 times as high as that for non-Hispanic white children.^[142]

Contemporary research findings have demonstrated that nationwide racial disparities in infant mortality are linked to the experiential state of the mother and that these disparities cannot be fully accounted for by socio-economic, behavioral or genetic factors.^[143] The Hispanic paradox, an effect observed in other health indicators, appears in the infant mortality rate as well. Hispanic mothers have an IMR comparable to non-Hispanic white mothers, despite lower educational attainment and economic status. A study in North Carolina, for example, concluded that "white women who did not complete high school have a lower infant mortality rate than black college graduates."^[144] According to Maternal & Child Welfare (A Coronary Artery Risk Development in Young Adults) study, "self-reported experience of racial discrimination were associated with pre-term and low-birthweight deliveries, and such experiences may contribute to black-white disparities in prenatal outcomes."^[145] Likewise, dozens of population-based studies indicate that "The subjective, or perceived experience of racial discrimination is strongly associated with an increased risk of infant death and with poor health behaviors for future generations of African Americans."^[146]

Wikipedia on infant mortality

W Infant mortality - Wikipedia × +

← → ↻ 🔒 https://en.wikipedia.org/wiki/Infant_mortality

Epidemiology [edit]

See also: *List of countries by infant mortality rate*

For the world, and for both less developed countries (LDCs) and more developed countries (MDCs), IMR declined. However, IMR was, and remains, higher in LDCs. In 2001, the IMR for LDCs (91) was about 10 times as large as that for MDCs. However, countries are, on average, much less than those among the more developed countries. ^[*clarification needed*]

A factor of about 67 separates the countries with the highest and lowest reported infant mortality rates. The top and

Rank	Country	Infant mortality rate (deaths/1,000 live births)
1	Afghanistan	121.63
2	Niger	109.98
3	Mali	109.08
4	Somalia	103.72
5	Central African Republic	97.17
218	Sweden	2.74
219	Singapore	2.65
220	Bermuda	2.47
221	Japan	2.21
222	Monaco	1.80

According to Guillot, Gerland, Pelletier and Saabneh "birth histories, however, are subject to a number of error

Fetching a table from Wikipedia

```
library("rvest")

# 1) fetch and parse the website
page <- read_html("https://en.wikipedia.org/wiki/Infant_mortality")
# 2) extract the html node containing the table
table <- html_node(page,
                    xpath = "//*[@id='mw-content-text']/div/table[2]")
# 3) extract the table as a data frame
mrates <- html_table(table)
```

Inspecting the HTML source

- Convenient with modern browsers: Use the developer tools (right-click *Inspect*).
- Look at the HTML source to grasp the structure.
- Find out how to navigate the site.
- Find the element(s) you want to extract.
- Get the Xpath expression or CSS selector to extract elements.

HTML elements visualized

w Infant mortality - Wikipedia x

← → ↻ 🔒 https://en.wikipedia.org/wiki/Infant_mortality#Epidemiology

Epidemiology [edit]

See also: *List of countries by infant mortality rate*

For the world, and for both less developed countries (LDCs) and more developed countries (MDCs), IMR declined significantly between 1960 and 2001. According to the [State of the World's Mothers report](#) by [Save the Children](#), the world IMR declined from 126 in 1960 to 57 in 2001.^[10]

However, IMR was, and remains, higher in LDCs. In 2001, the IMR for LDCs (91) was about 10 times as large as it was for MDCs (8). On average, for LDCs, the IMR is 17 times as higher than that of MDCs. Also, while both LDCs and MDCs made significant reductions in infant mortality rates, reductions among less developed countries are, on average, much less than those among the more developed countries. [clarification needed]

A factor of about 67 separate countries with the highest and lowest reported infant mortality rates.

The top and bottom five countries by this measure (taken from [The World Factbook's 2012](#) `table.wikitable` | 300x320
by `josiah` and shown below.

Rank	Country	Infant mortality rate (deaths/1,000 live births)
1	Afghanistan	121.63
2	Niger	109.98
3	Mali	109.08
4	Somalia	103.72
5	Central African Republic	97.17
216	Sweden	2.74
219	Singapore	2.65
220	Bermuda	2.47
221	Japan	2.21
222	Monaco	1.80

According to Guillot, Gerland, Pelletier and Saabneh "birth histories, however, are subject to a number of errors, including omission of deaths and age misreporting errors."¹⁰³

United States [\[edit \]](#)

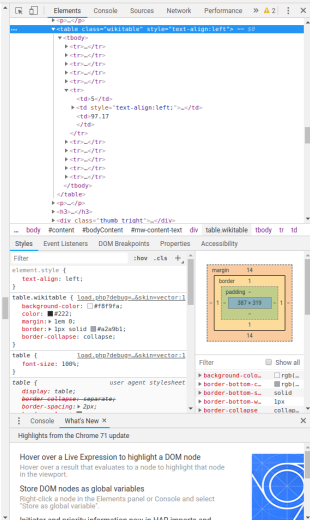
The infant mortality rate in the US decreased by 2.3% to a historic low of 582 infant deaths per 100,000 live births in 2014.^[104]

Of the 27 most developed countries, the US has the highest Infant Mortality Rate, despite spending much more on health care per capita^[data not needed]. Significant racial and socio-economic differences in the United States affect the IMR, in contrast with other developed countries, which have more homogeneous populations. In particular, IMR varies greatly by race in the US. The average IMR for the whole country is therefore not a fair representation of the wide variations that exist between segments of the population. Many theories have been proposed as to why these racial differences exist, with socio-economic factors usually coming out as the most plausible explanation. However, more studies

World historical and predicted infant mortality rates per 1,000 births (1950-2050)

UN, medium variant, 2008 rev. ⁽¹⁰⁰⁾

Years	Rate	Years	Rate
1950–1955	152	2000–2005	52
1955–1960	136	2005–2010	47
1960–1965	116	2010–2015	43
1965–1970	100	2015–2020	40
1970–1975	91	2020–2025	37
1975–1980	83	2025–2030	34
1980–1985	74	2030–2035	31
1985–1990	65	2035–2040	28
1990–1995	61	2040–2045	25
1995–2000	57	2045–2050	23



Infant mortality rates from Wikipedia

```
<table class="wikitable" style="text-align:left">
  <tbody>
    <tr>
      <th>Rank</th>
      <th>Country</th>
      <th>Infant mortality rate <br> (deaths/1,000 live births)</th>
    </tr>
    <tr>
      <td>1</td>
      <td style="text-align:left;"><a href="/wiki/Afghanistan" title="Afghanistan">Afghanistan</a></td>
      <td>121.63</td>
    </tr>
    <tr>
      <td>2</td>
      <td style="text-align:left;"><a href="/wiki/Niger" title="Niger">Niger</a></td>
      <td>109.98</td>
    </tr>
    <tr>
      <td>3</td>
      <td style="text-align:left;"><a href="/wiki/Mali" title="Mali">Mali</a></td>
      <td>109.08</td>
    </tr>
    <tr>
      <td>4</td>
      <td style="text-align:left;"><a href="/wiki/Somalia" title="Somalia">Somalia</a></td>
      <td>103.72</td>
    </tr>
    ...
  </tbody>
</table>
```


CSS selectors and XPath expressions

```
# fetch and parse the website
page <- read_html("https://en.wikipedia.org/wiki/Infant_mortality")
# list the table nodes
html_nodes(page, "table")
# using xpath expressions or css selectors is equivalent
table <- html_node(page,
                    xpath = "//*[@id='mw-content-text']/div/table[2]")
table <- html_node(page,
                    css = "#mw-content-text > div > table:nth-child(121)")
```

Examples

The general structure

- There is no universal recipe. But most programs follow a certain structure.
 1. Open a website mimicking a browser and navigate it (optional).
 2. Get the page source HTML and feed it to a parser.
 3. Extract the elements you need.
 4. Filter and arrange them as needed and save them.
 5. Repeat 1.-4. as needed.

Navigating to another page

```
# Open infant mortality page
session <- session("https://en.wikipedia.org/wiki/Infant_mortality")
# Goto page on Somalia
session <- session_follow_link(session, "Somalia")
# Read the source
page <- read_html(session)
# Extract html
table <- html_node(page,
  xpath = "//*[@id='mw-content-text']/div/table[4]")
regions <- html_table(table)
```

Filtering links

```
# read wiki page
page <- read_html("https://en.wikipedia.org/wiki/Infant_mortality")
# get the links
wikilinks <- html_attr(html_nodes(page, "a"), "href")
# use regex to filter internal links:
#   select only articles, no files or category pages,
#   matching with mortality or somalia
links <- grep("^(?!.*:)(/wiki/. *Mortality)|(/wiki/. *Somalia)", wikilinks,
              ignore.case = TRUE, value = TRUE, perl = TRUE)
links <- unique(links)
# go to first selected article page and process it
session <- jump_to(session, links[1])
page <- read_html(session)
html_nodes(page, "title")
```

~~A more elaborate example~~ An example of impermanence

- Phillippine Statistics Authority *Good Governance Index*.
- Available at ~~<http://nap.psa.gov.ph/ggi/default.asp>~~.
- Available at <https://web.archive.org/web/20190915135458/http://nap.psa.gov.ph/ggi/default.asp>.
- Get all GGI data tables for all municipalities.
- Save them in a local data file for further analysis.
- How would you go about this?

Another example

- WHO venomous snakes distribution and species risk categories
- Database form link:
<https://apps.who.int/bloodproducts/snakeantivenoms/database/SearchFrm.aspx>
- Collect snake data for all countries.
- Getting dropdown options and initial form submission straightforward.
- So is table extraction.
- Navigating further links is tricky.

If simple navigation fails

- Some web pages cannot be navigated easily with simpler requests.
- Often due to hidden Javascript or other server-side processing.
- In this case, resort to Selenium (`library(rselenium)` in R).
- Under the hood, Selenium relies on a complete browser running in a container.
- Slower and comes with substantial overhead costs.
- Only use when necessary.

- Start simple and expand your program incrementally.
- Keep it simple. Do not overengineer the problem. Do not repeat yourself.
- Limit the number of iterations for test runs. Use print statements to inspect objects.
- Write tests to verify things work as intended.
- If your program requires complex monitoring/validation procedures or threading for performance, use Python.

Final remarks

- Sometimes small programs can go a long way.
- Do not lose sight of your ultimate goal. Time is valuable.
- Do not engage in perfectionism, focus on GTD.
- Identify everyday tasks that you can optimize.

Assignment

Assignment, part I

Choose **one** of the following.

Open Data Tanzania

<https://tanzania.opendataforafrica.org/>

- ➡ For each region in Tanzania, get the basic facts (top of the page: capital, languages, area, ...), population and mortality statistics (tables under links 'Population projections, total' and 'Deaths').

Books to sell

<https://books.toscrape.com/index.html>

- ➡ For every category, collect all books with title, price, availability, stock and product description.

Swiss Tax Calculator

<https://swisstaxcalculator.estv.admin.ch/#/taxburden/income-wealth-tax>

- ➡ Collect municipal income tax burden statistics for the following parameters: All payers, atheist, geographical comparison (all municipalities), all years, income 50'000 CHF.

Assignment, part I

- Submission deadline is November 7, 2021.
- Submit code only, no data.
- Comment your code or submit a short description alongside.
- A proof-of-concept restricted to the first couple of regions/... is fine.
- Accounts for 20% of the final grade.

Next lecture: Text as data.