

Welcome

Please sign-up for an account at transportapi.com

Once logged in click dashboard->applications and
create a new application with an application id and api
key.

We will need these later.





Unlocking Data from the Web

<http://training.theodi.org/UnlockingData>

David Tarrant · @davetaz

Open Data Science

Day 1: Unlocking data from the web

Day 2: Data management and statistics

Day 3: Big data and data visualisation



Introductions

Your name

What is your favourite example/use of open data?

What do you want to do differently after the course?



Course aim

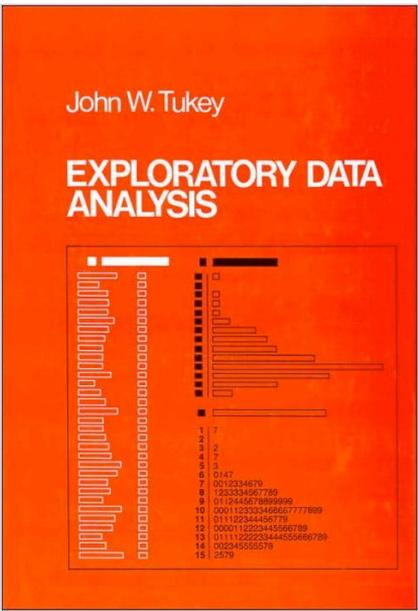
Equip you with the knowledge and tools
to help you upskill as modern data scientists.



Question

What is data science?





12:00 AM
January 1, 1962

John W. Tukey writes "The Future of Data Analysis"

I thought I was a **statistician**... but as I have watched mathematical statistics evolve, I have had cause to wonder and doubt... I have come to feel that my central interest is in **data analysis**... Data analysis, and the parts of statistics which adhere to it, must...**take on the characteristics of science rather than those of mathematics**... data analysis is intrinsically an empirical science... How vital and how important



12:00 AM
The International Association for Statistical Computing (IASC) is established.

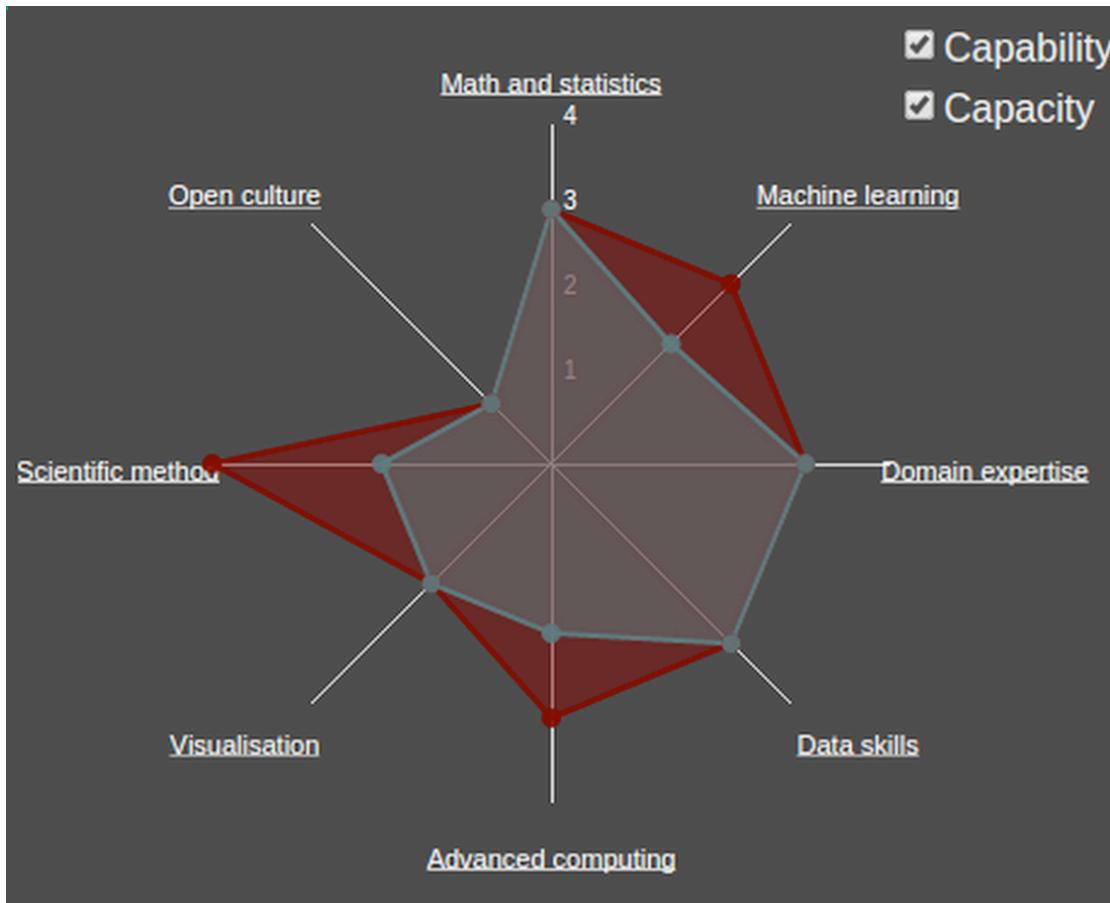
John W. Tukey writes
"The Future of Data
Analysis"

1962 1964 1966 1968 1972 1974 1976

History of data science



Data science

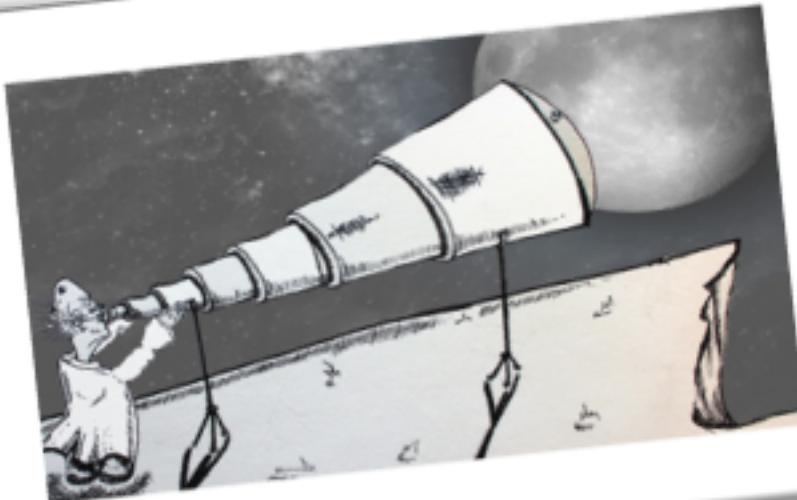




Computer Science, = Telescope science ?

"Computer science is no more about computers than astronomy is about telescopes"

-- *Edsger W. Dijkstra*



*"The computer is not our object of study,
It's our observational instrument"*

Thanks to Frank van Harmelen

Today

- Data skills
- Advanced computing
- Data visualisation



The “Data” bit of data science



Session 1

Unlocking data from the web

Session 2

Processing data

Session 3

Publishing insight

Session 1

Unlocking data from the web



Outcomes

List and identify the key structures and formats of data

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web



Outcomes

List and identify the key structures and formats of data

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web



Formats and Structures

What's the difference between a data structure and a data format?



Exercise

Exploring data formats



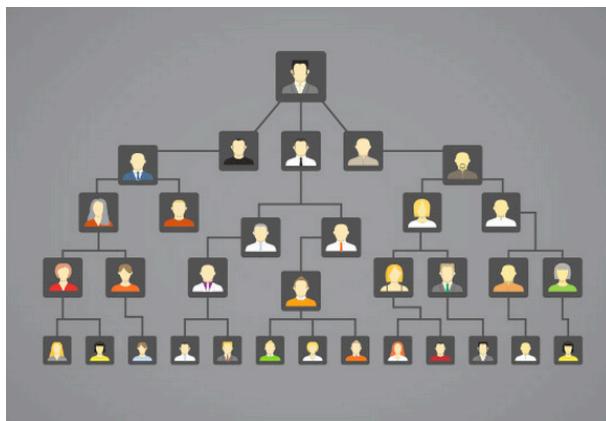
<http://training.theodi.org/UnlockingData/>



Structures



Tabular



Hierarchical



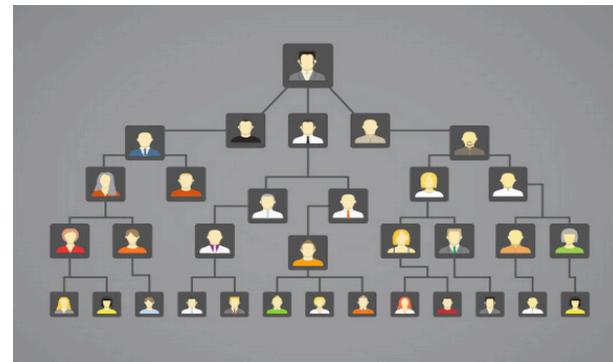
Network



Structures and formats



Tabular
csv, xls, json



Hierarchical
json, rdf*, xls**



Network
rdf*, json, html

- * rdf is not a format, it is a schema that can be serialised into many formats.
- ** xls can be abused to create hierarchical databases using multiple worksheets



Choose a structure and format:

1. A list of events
2. A shopping catalog
3. Alternative names for companies



Outcomes

List and identify the key structures and formats of data

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web

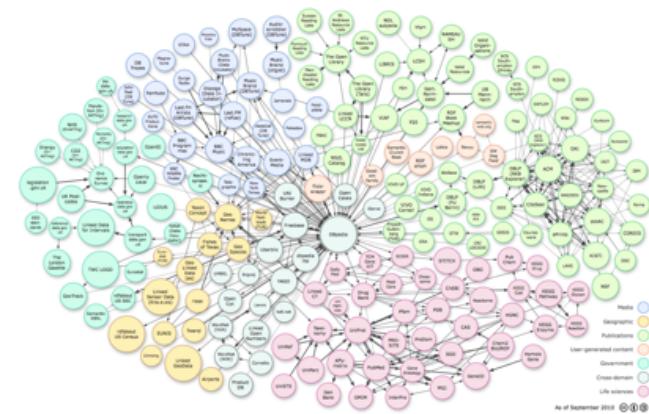


Approaches to publishing data

ON the web



IN the web

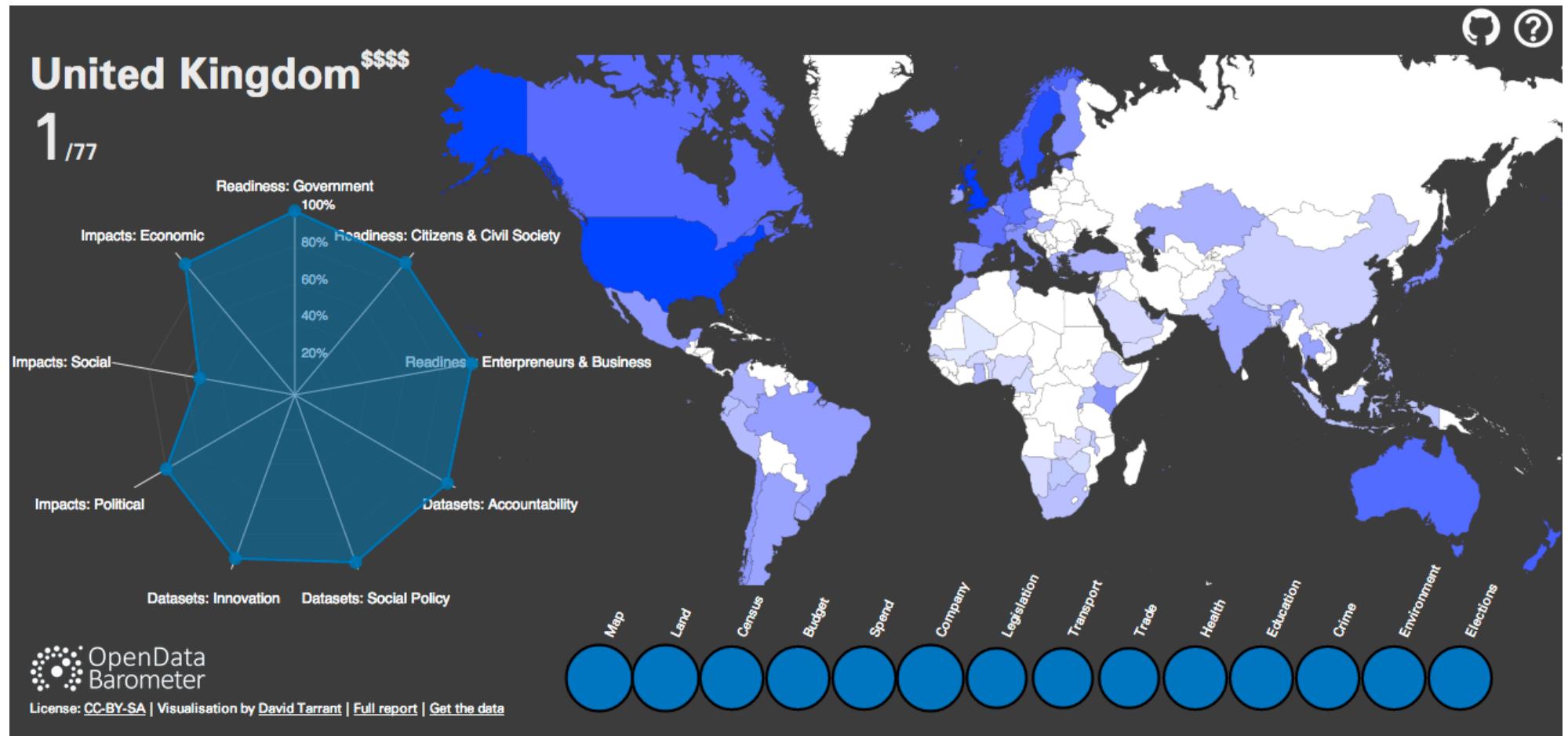


Finding data on the web (of documents)

- Government data
- Private sector data
- Google advanced
- Aggregators and portals
- Scraping



Government data



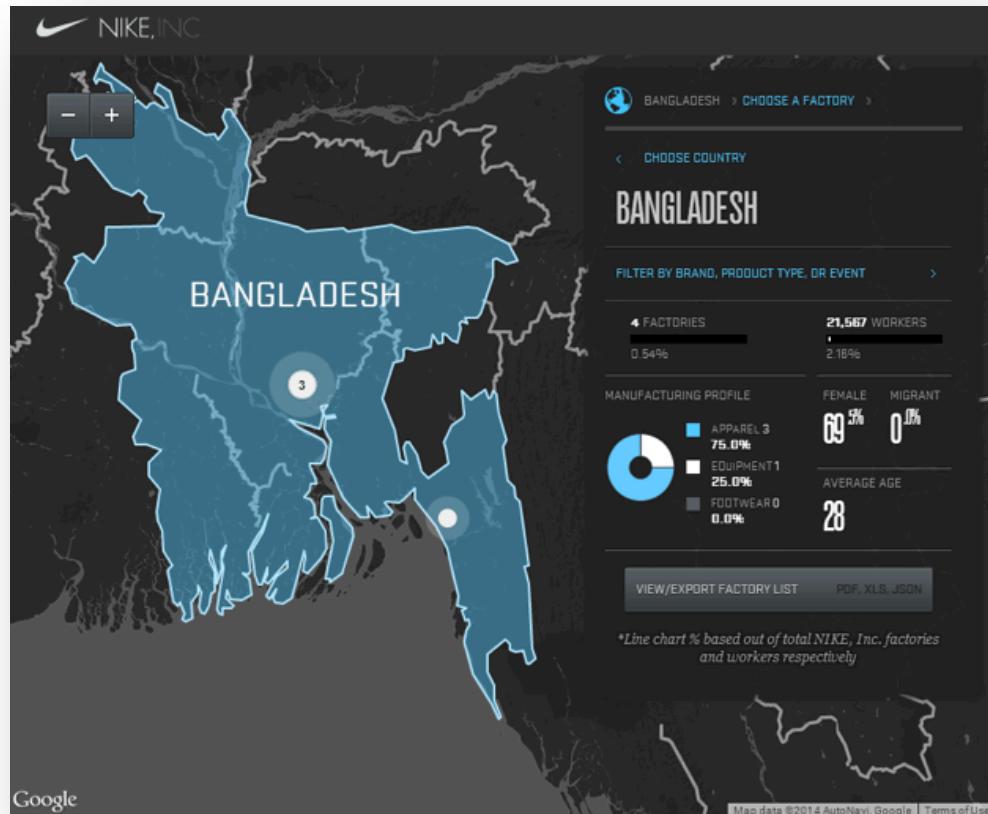
data.gov.XX

The screenshot shows the DATA.GOV.UK beta homepage with a search bar and navigation menu. A search result for 'Live traffic information from the Highways Agency' is displayed, showing a map of the United States and Mexico.

The screenshot shows the DATOS.GOB.MX BETA website. It features a map-based search interface for datasets, a search bar, and a section titled '127 datasets found'. Below this is a 'RATING SOCIAL PROGRAMS' section with a description of the database.

The screenshot shows the data.gov.my website, which is the official portal for Malaysia's open data. It includes sections for datasets, resources, mobile apps, and quick links.

Suppliers



<http://manufacturingmap.nikeinc.com/#>

You agree not to change or delete any ownership notices from materials downloaded or printed from the Platform. You agree not to modify, copy, translate, broadcast, perform, display, distribute, frame, reproduce, republish, download, display, post, transmit or sell any Intellectual Property or Content appearing on the Platform



Google advanced

Google site:gov filetype:xls

Web Images Maps Shopping More Search tools

About 4,150,000 results (0.22 seconds)

[\[xls\] Code List or Concept \(Acronym\)](#) ↗
www.acquisition.gov/short_codelistsTS.xls Share
File Format: Microsoft Excel - [View as HTML](#)
A, B, C, D, F, G, H, 1, Code List or Concept (Acronym), Definition, Authoritative or Maintenance Agency, Current Version, Update Frequency and/or ...

[\[xls\] Approps - Foreign Assistance.gov](#) ↗
www.foreignassistance.gov/Full_ForeignAssistanceData.xls
File Format: Microsoft Excel
A, B, C, D, E, F, G, H, 1, Planning Data, 2, 3, 4, Fiscal Year, Fiscal Year Type, Account Name, Agency Name, Operating Unit, Category, Sector, Amount ...

[\[xls\] TSB Monthly Cash Flow Projection](#) ↗
www.dia.mil/ia/awc/awcfo/cashflow.xls

site: Get results only from certain sites or domains

link: Find pages that link to a certain page

related: Find sites similar to one you already know

filetype: Find certain file types only

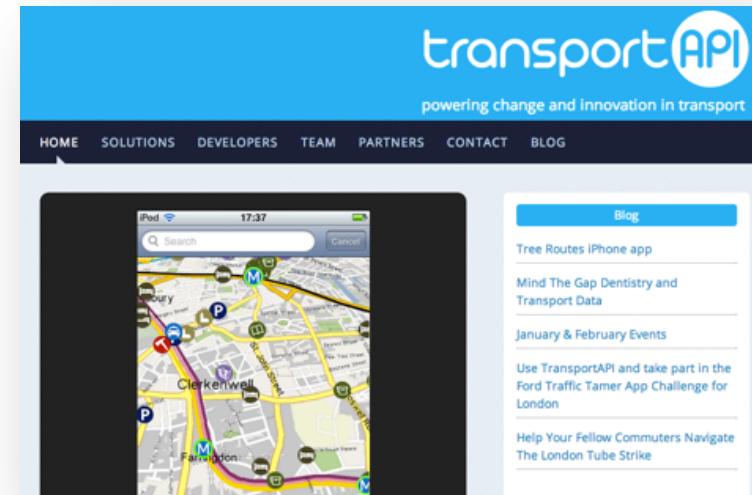


Aggregators and portals

Collect together data from across the web into one place.



enigma.io

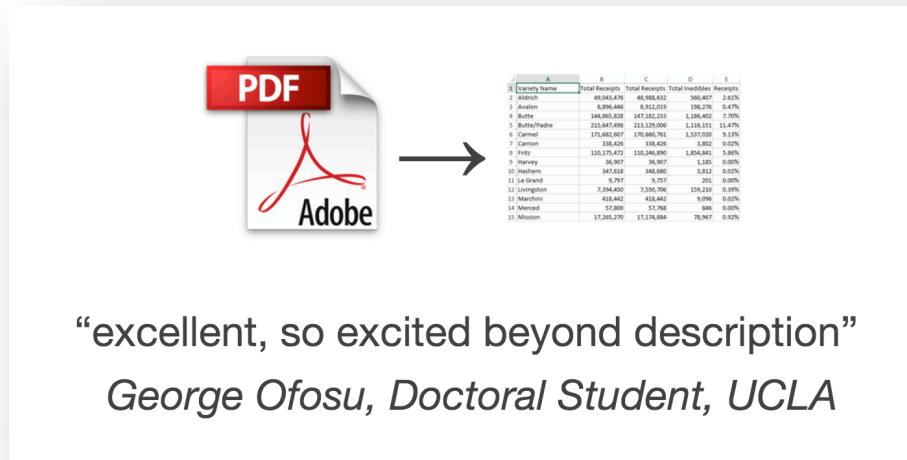


transportAPI



Scraping

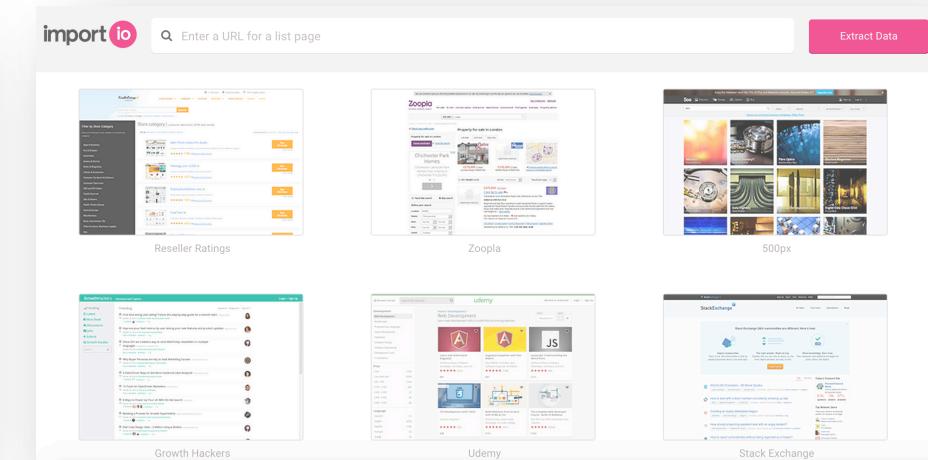
If you can't obtain usable data (csv, xls) then you may have to resort to scraping.



A diagram illustrating the process of data scraping. On the left, there is a white document icon with a red 'PDF' label and an Adobe logo. An arrow points from this icon to a screenshot of a Microsoft Excel spreadsheet. The spreadsheet contains a table with columns labeled 'Variety Name', 'Total Receipts', 'Total Receipts', 'Total Inedibles Receipts', and 'Receipts'. The data includes rows for various fruit varieties like Aldrich, Avocado, Butte, Butte/Padre, Cawein, Carrion, Fritz, Harkness, Le Grand, Liugong, Marchini, Merced, and Mission, along with their respective receipt counts and percentages.

“excellent, so excited beyond description”
George Ofosu, Doctoral Student, UCLA

pdftables.com



A screenshot of the import.io web interface. At the top, there is a search bar with the placeholder "Enter a URL for a list page" and a pink "Extract Data" button. Below the search bar, there are six examples of scraped data from different websites:

- Reseller Ratings:** A screenshot of a website showing product reviews and ratings.
- Zoopla:** A screenshot of a real estate listing website.
- 500px:** A screenshot of a photo sharing website.
- Growth Hackers:** A screenshot of a blog or community website.
- Udemy:** A screenshot of an online learning platform.
- Stack Exchange:** A screenshot of a Q&A website.

magic.import.io



5-Stars



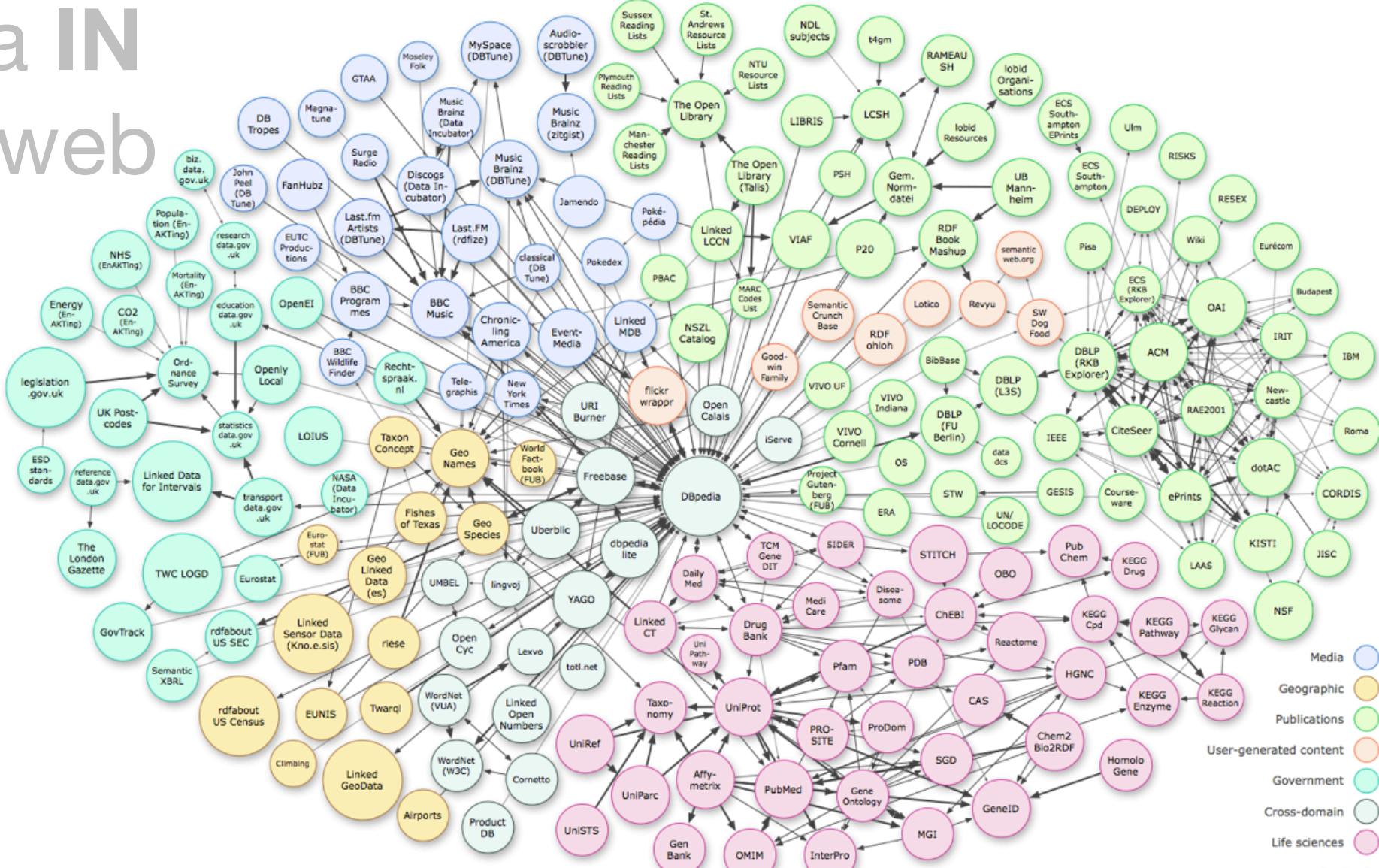
<http://5stardata.info/>

ON THE WEB



IN THE WEB

Data IN the web



As of September 2010

Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data **IN THE WEB**
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)



How the web should work,
but people forgot that Tim
put this in when he
invented it!



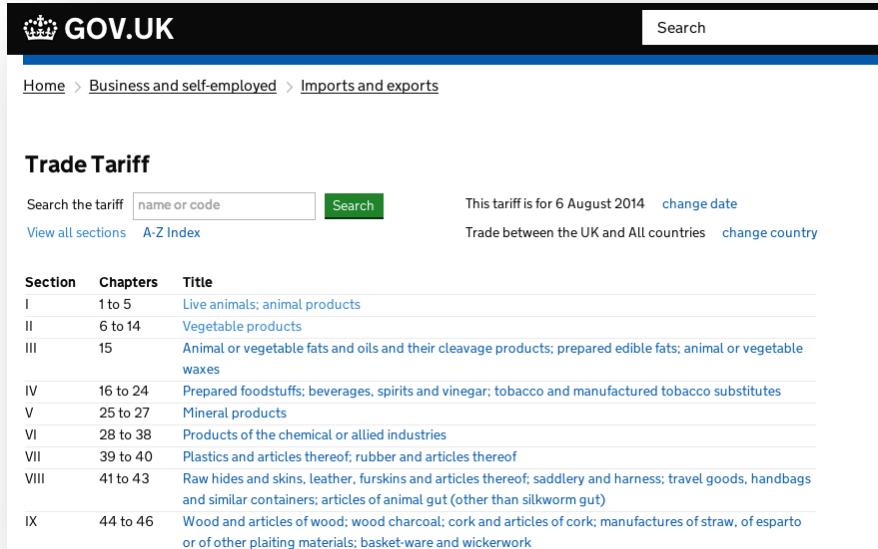
Duck typed data

If it looks like a duck
and quacks like a duck,
then it's probably a duck.

Basically, keep an eye out for tables,
lists and other stuff that looks like data.



1. Adding random extensions



The screenshot shows the 'Trade Tariff' section of the GOV.UK website. At the top, there's a search bar and a navigation menu with links to 'Home', 'Business and self-employed', and 'Imports and exports'. Below this, the title 'Trade Tariff' is displayed. A search bar allows users to 'Search by name or code' and a 'Search' button. To the right, a note says 'This tariff is for 6 August 2014' with a link to 'change date'. There are also links to 'View all sections' and 'A-Z Index'. The main content area is a table showing tariff sections and their titles:

Section	Chapters	Title
I	1 to 5	Live animals; animal products
II	6 to 14	Vegetable products
III	15	Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal or vegetable waxes
IV	16 to 24	Prepared foodstuffs; beverages, spirits and vinegar; tobacco and manufactured tobacco substitutes
V	25 to 27	Mineral products
VI	28 to 38	Products of the chemical or allied industries
VII	39 to 40	Plastics and articles thereof; rubber and articles thereof
VIII	41 to 43	Raw hides and skins, leather, furskins and articles thereof; saddlery and harness; travel goods, handbags and similar containers; articles of animal gut (other than silkworm gut)
IX	44 to 46	Wood and articles of wood; wood charcoal; cork and articles of cork; manufactures of straw, of esparto or of other plaiting materials; basket-ware and wickerwork

UK Trade Tariff



The screenshot shows the BBC Music and Programmes website for the TV show 'Doctor Who'. The header features the 'one' logo and the show's name. Below the header, a navigation bar includes links for 'Home', 'Episodes', 'Clips', 'Galleries', 'Latest News', 'Characters', 'Monsters', 'Fun and Games', and 'More'. A large image of the Doctor and a companion is prominently displayed. On the left, a box for 'On iPlayer' says 'This programme will be available shortly after broadcast'. On the right, a box for 'On TV' shows a preview of 'The Day of the Doctor'. The bottom of the page has a banner for the 'Day of the Doctor' event.

BBC Music and Programmes



Try using the following: .csv .json .xml .rss .rdf

2. Look for alternative links



Business Insight - NEWSASIA

NEWS TV WATCH LIVE

Wed, Aug 06 2014

ASIA PACIFIC SINGAPORE WORLD BUSINESS SPORT ENTERTAINMENT TECHNOLOGY HEALTH LIFESTYLE VIDEOS WEATHER MORE ▾

CHANGINGLIVES LUMINARY AWARDS START-UP

SCROLL DOWN!

SINGAPORE STORIES

Raise of up to 12% for Home Team officers, with sign-on bonuses of up to \$S30,000

SP (2) Ng (SP)

MEDIALCORP

Officers from the Home Team, both past and present were recognised at the Home Team National Day Observance Ceremony on Wednesday.

9 hours ago

Pay rise, special bonus for about 23,000 nurses

10 hours ago

50,000 openings on Jobs Bank for Singaporeans, PRs

1 hour ago

NUS University Town identified as a high-risk dengue cluster

10 hours ago

LIFESTYLE VIDEOS



2. Look for alternative links



 CHANNEL NEWSASIA MediaCorp News Group. © 2014 MediaCorp Pte Ltd. All Rights Reserved. Terms and Conditions Privacy Policy About MediaCorp Pte Ltd	NEWS Asia Pacific Singapore World Business Sport Entertainment Technology Health Lifestyle Videos Photos Special Reports Archives	TV Live TV TV Videos TV Schedule SERVICES Weather ADVERTISE WITH US Online Advertising Mobile Advertising TV Advertising Contact Sales	ABOUT US About Channel NewsAsia Our Logo Our Coverage Our Tagline Presenters and Correspondents Contact Us GET OUR NEWS       
---	---	---	---

RSS



3. Look for embedded data

ODI Experiment

Hidden data extractor

 open
data
institute

Hidden data extractor

Enter the URL of any webpage to see what JSON data is hidden within it.

Submit

Try these

[Products from Marks and Spencer UK](#)

[Products from ASOS](#)



Finding data on the web (of data)

1. Add random extensions (.xml, .json, .csv etc)
2. Look for alternative links (rss feeds etc)
3. Look for embedded data **IN THE WEB**
4. Do some content negotiation
5. Spot the API
6. Scrape (or search google again)



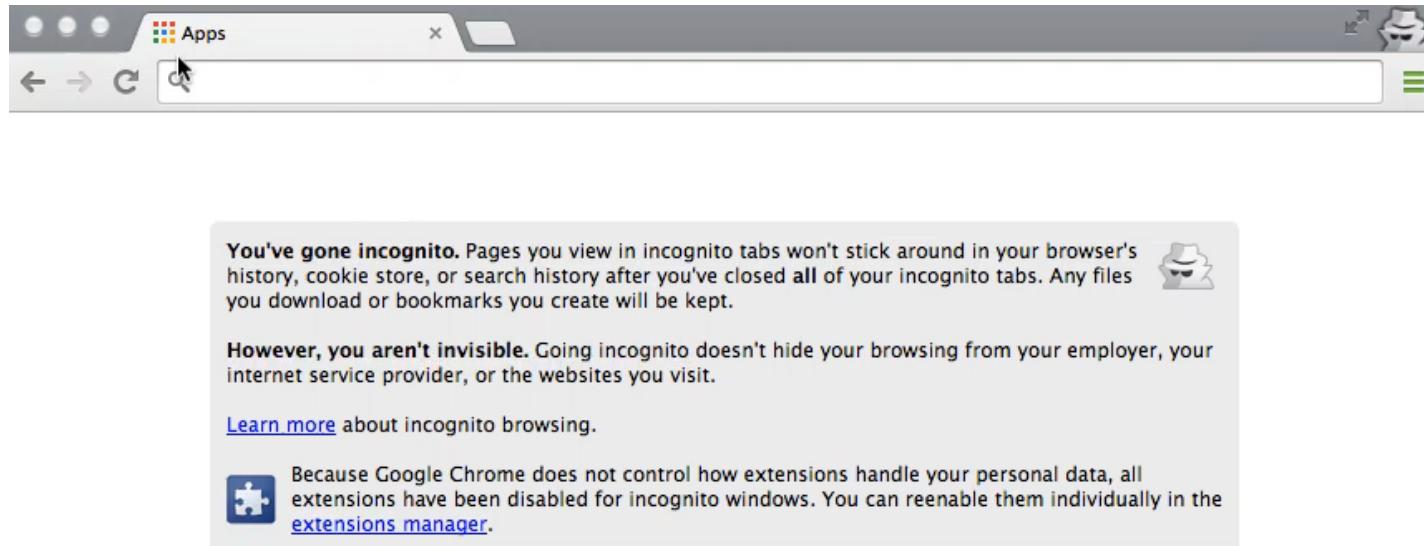
How the web should work,
but people forgot that Tim
put this in when he
invented it!



Content Negotiation



How we use the web



What decisions were taken for you?

- Google.com redirected to google.co.uk
- Searching google.co.uk assumed you wanted english
- Clicking the link assumed you wanted the pretty html page representing Albert Einstein
- So all of the application layer is hidden



In Spanish?

Albert Einstein - Wikipedia

en.wikipedia.org/wiki/Albert_Einstein

Create account Log in

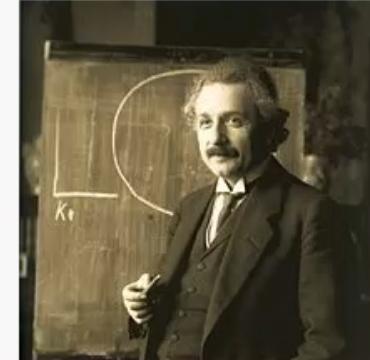
Article Talk Read View source View history Search

Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For other uses, see [Albert Einstein \(disambiguation\)](#) and [Einstein \(disambiguation\)](#).

Albert Einstein (/ælˈbɛrt ˈaɪnʃtɪn/; German: [albɛrt ˈaɪnʃtɛn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist. He developed the general theory of relativity, one of the two pillars of modern physics (alongside quantum mechanics).^{[2][3]} He is best known for his mass–energy equivalence formula $E = mc^2$ (which has been dubbed "the world's most famous equation").^[4] He received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect".^[5] The latter was pivotal in establishing quantum theory.



Albert Einstein

CC BY SA

- So what did we need to do?
- Search on the page and hope!



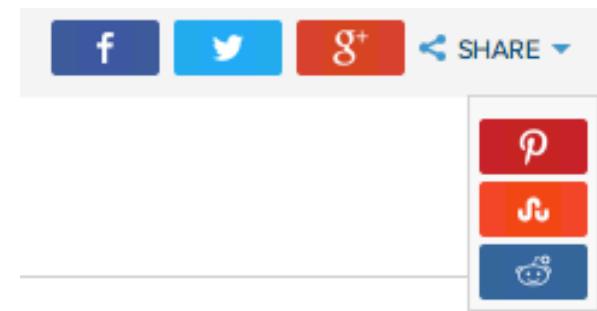
Common links

RSS Feed



Languages

Social channels



Printable version



Print version

Albert Einstein - Wikipedia

es.wikipedia.org/wiki/Albert_Einstein

Crear una cuenta Iniciar sesión

Artículo Discusión Leer Editar Ver historial Buscar

WIKIPEDIA La enciclopedia libre

Portada Portal de la comunidad Actualidad Cambios recientes Páginas nuevas Página aleatoria Ayuda Donaciones Notificar un error

▼ Imprimir/exportar Crear un libro Descargar como PDF Versión para imprimir

► Herramientas

▼ Otros proyectos

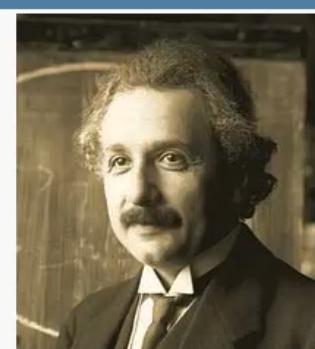
Albert Einstein

Para otros usos de este término, véase [Einstein \(desambiguación\)](#).

Albert Einstein (en alemán ['albet 'aɪnʃtaɪn]; Ulm, Imperio alemán, 14 de marzo de 1879 - Princeton, Estados Unidos, 18 de abril de 1955) fue un físico alemán de origen judío, nacionalizado después suizo y estadounidense. Es considerado como el científico más importante del siglo XX. Manuel Alfonseca cuantifica la importancia de 1000 científicos de todos los tiempos y, en una escala de 1 a 8, Einstein y Freud son los únicos del siglo XX en alcanzar la máxima puntuación;¹ asimismo califica a Einstein como «el científico más popular y conocido del siglo XX».²

En 1905, cuando era un joven físico desconocido, empleado en la Oficina de Patentes de Berna, publicó su

Albert Einstein 😊



CC BY SA

Common links

RSS Feed



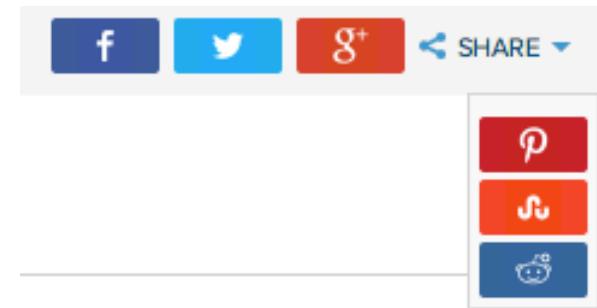
Languages



Printable version



Social channels



How a machine does it

Postman W Albert Einstein - Wikipedia en.wikipedia.org/wiki/Albert_Einstein NEW ABP

Create account Log in

WIKIPEDIA The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Wikimedia Shop

Interaction Help About Wikipedia Community portal Recent changes Contact page Tools

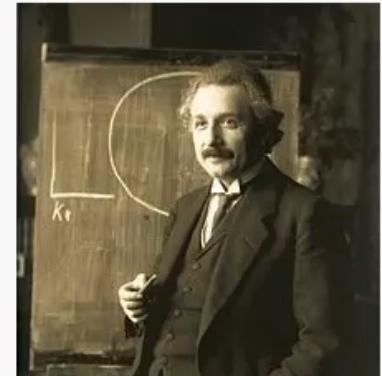
Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For other uses, see [Albert Einstein \(disambiguation\)](#) and [Einstein \(disambiguation\)](#).

Albert Einstein (/ælbert ˈaɪnstaɪn/; German: [albert ˈaɪnʃtaɪn] ([listen](#)); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist. He developed the [general theory of relativity](#), one of the two pillars of [modern physics](#) (alongside [quantum mechanics](#)).^{[2][3]} He is best known for his [mass–energy equivalence formula](#) $E = mc^2$ (which has been dubbed "the world's most famous equation").^[4] He received the [1921 Nobel Prize in Physics](#) "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect".^[5] The latter was pivotal in establishing [quantum theory](#).

Near the beginning of his career, Einstein thought that



Albert Einstein



A Web Request



http://en.wikipedia.org/wiki/Albert_Einstein

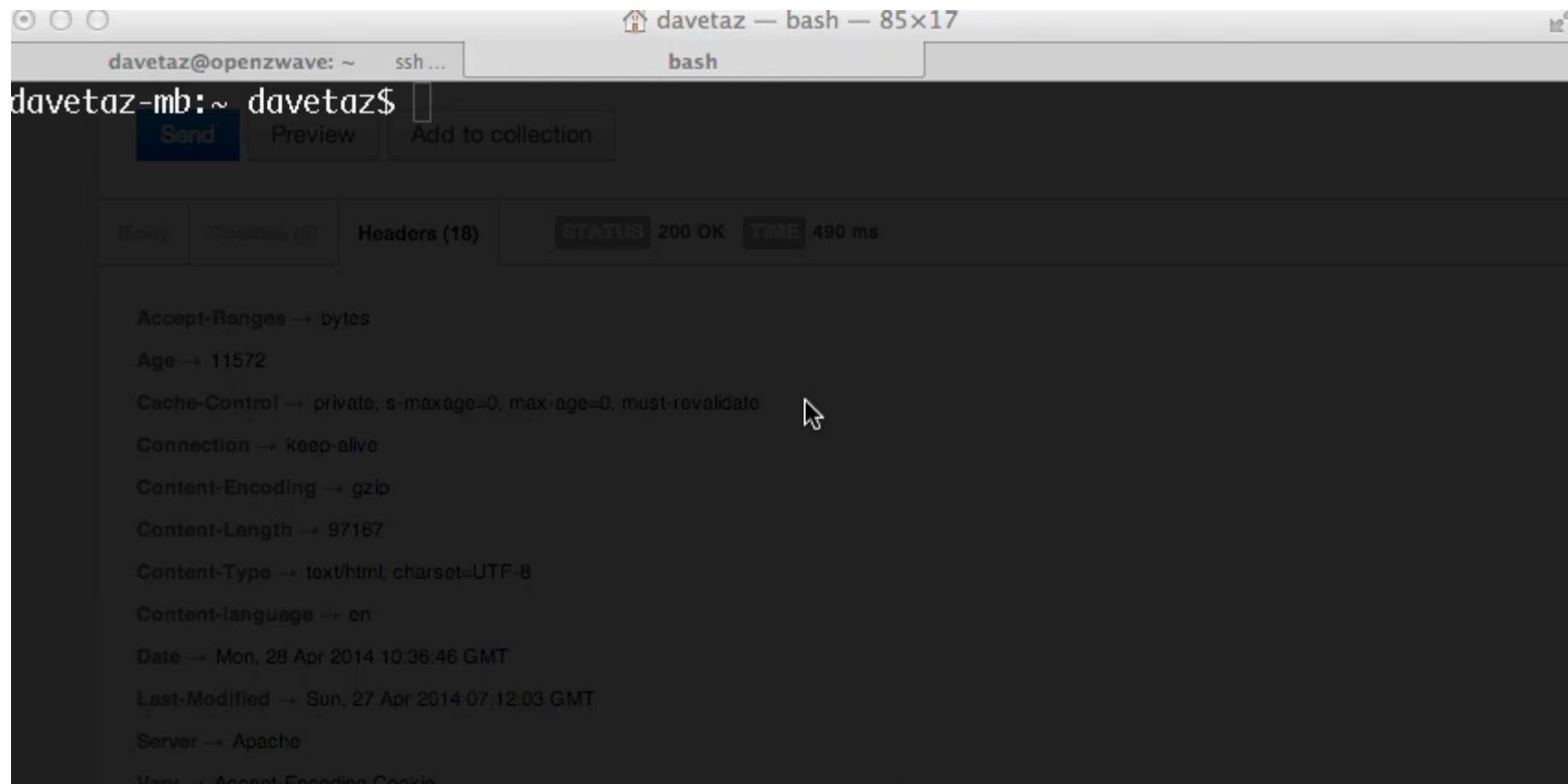
- **GET / HTTP/1.1**

```
telnet en.wikipedia.org 80
```

```
> GET /wiki/Albert_Einstein HTTP/1.1
> Host: en.wikipedia.org
```



A Web Request



A screenshot of a terminal window titled "davetaz — bash — 85x17". The window shows a command-line interface with the prompt "davetaz@openzwave: ~". Below the prompt, there are three buttons: "Send", "Preview", and "Add to collection". The main area displays a list of HTTP headers from a response. The headers listed are:

- Accept-Ranges → bytes
- Age → 11572
- Cache-Control → private, s-maxage=0, max-age=0, must-revalidate
- Connection → keep-alive
- Content-Encoding → gzip
- Content-Length → 97167
- Content-Type → text/html; charset=UTF-8
- Content-language → en
- Date → Mon, 28 Apr 2014 10:36:46 GMT
- Last-Modified → Sun, 27 Apr 2014 07:12:03 GMT
- Server → Apache
- Vary → Accept-Encoding, Cookie

The "STATUS" field shows "200 OK" and the "TIME" field shows "490 ms".



The same

Postman W Albert Einstein - Wikipedia en.wikipedia.org/wiki/Albert_Einstein Create account Log in

Article Talk Read View source View history Search

WIKIPEDIA The Free Encyclopedia

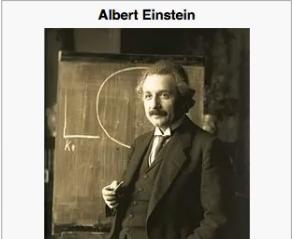
Albert Einstein

From Wikipedia, the free encyclopedia

"Einstein" redirects here. For other uses, see [Albert Einstein \(disambiguation\)](#) and [Einstein \(disambiguation\)](#).

Albert Einstein (/ælərt ˈaɪnʃtɪn /; German: [albert ˈaɪnʃtaɪn] (listen); 14 March 1879 – 18 April 1955) was a German-born theoretical physicist. He developed the general theory of relativity, one of the two pillars of modern physics (alongside quantum mechanics).^{[2][3]} He is best known for his mass–energy equivalence formula $E = mc^2$ (which has been dubbed "the world's most famous equation").^[4] He received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect".^[5] The latter was pivotal in establishing quantum theory.

Near the beginning of his career, Einstein thought that



davetaz@openwave: ~ ssh ... bash

davetaz-mb: ~ davetaz\$

Send Preview Add to collection

Status Headers (18) Status 200 OK Time 490 ms

Accept-Ranges → bytes
Age → 11572
Cache-Control → private, s-maxage=0, max-age=0, must-revalidate
Connection → keep-alive
Content-Encoding → gzip
Content-Length → 97167
Content-Type → text/html; charset=UTF-8
Content-Language → en
Date → Mon, 28 Apr 2014 10:36:46 GMT
Last-Modified → Sun, 27 Apr 2014 07:12:03 GMT
Server → Apache
More → About Encoding Cache



Headers

The headers define the properties of an HTTP transaction.

The screenshot shows the Postman application interface. At the top, there's a header bar with tabs for 'Normal', 'Basic Auth', 'Digest Auth', 'OAuth 1.0', and 'No environment'. Below the header, the URL is set to 'http://en.wikipedia.org/wiki/Albert_Einstein' with a 'GET' method selected. There are buttons for 'URL params' and 'Headers (0)'. The main body of the interface shows a table with 'Header' and 'Value' columns, and a 'Manage presets' button. Below this is a row with 'Send' (highlighted in blue), 'Preview', 'Add to collection', and 'Reset' buttons. At the bottom, there are tabs for 'Body', 'Cookies (6)', 'Headers (18)', 'STATUS 200 OK', and 'TIME 428 ms'. Under the 'Body' tab, there are buttons for 'Pretty', 'Raw', 'Preview', and 'JSON/XML'. The 'Pretty' tab is selected, displaying the HTML source code of the Wikipedia page for Albert Einstein. The code includes meta tags for charset, title, and generator, along with links for alternate, apple-touch-icon, shortcut icon, search, and EditURI.

```
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
  <head>
    <meta charset="UTF-8" />
    <title>Albert Einstein - Wikipedia, the free encyclopedia</title>
    <meta http-equiv="X-UA-Compatible" content="IE=EDGE" />
    <meta name="generator" content="MediaWiki 1.24wmf1" />
    <link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Albert_Einstein" />
    <link rel="apple-touch-icon" href="//bits.wikimedia.org/apple-touch/wikipedia.png" />
    <link rel="shortcut icon" href="//bits.wikimedia.org/favicon/wikipedia.ico" />
    <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)" />
    <link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
    <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
```



How a machine does it

I have an address

I would like a different version:

Accept:



application/rss+xml

Accept-Language:



es-es



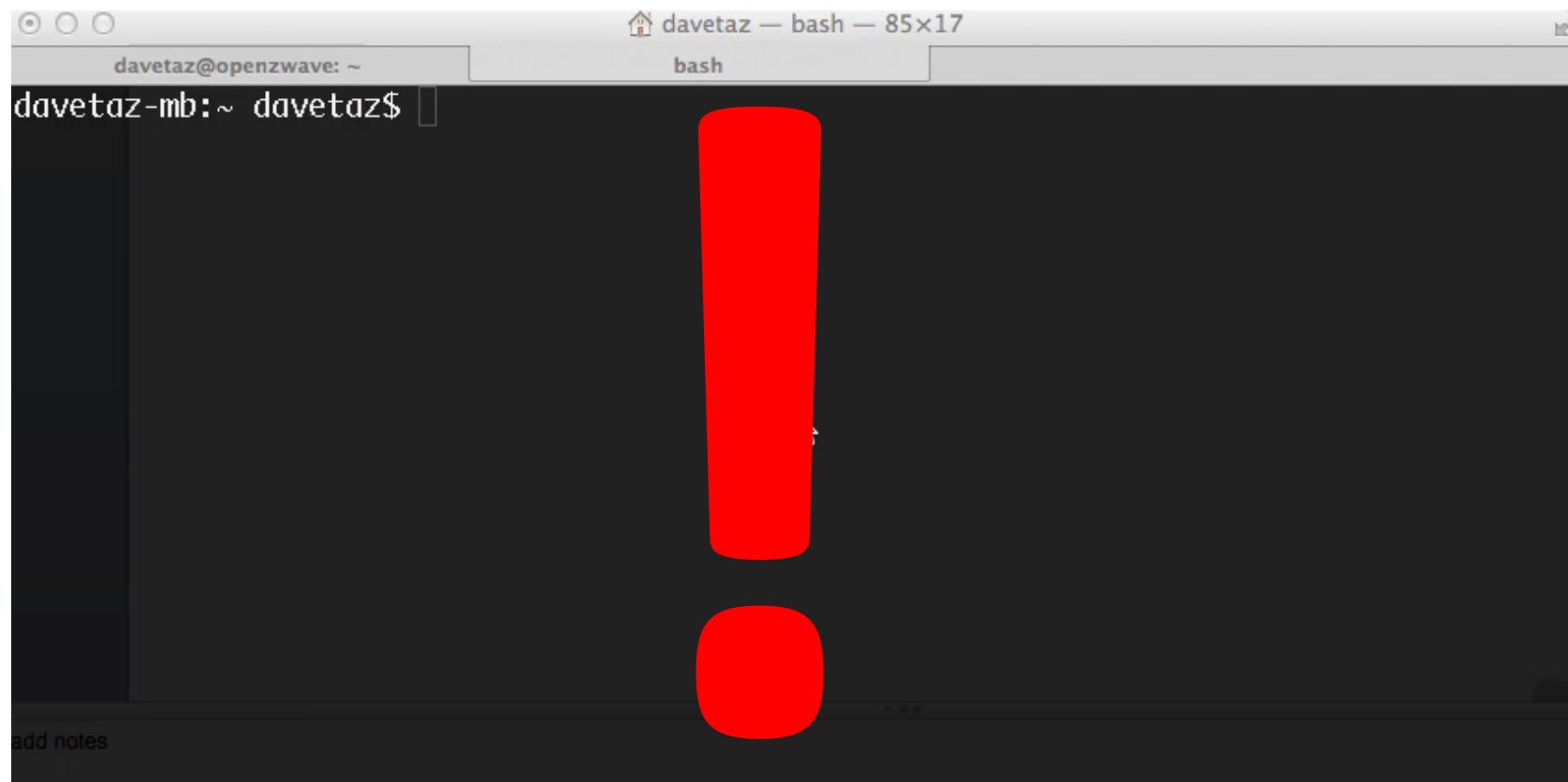
Spanish please

The screenshot shows the Postman application interface. The header bar includes tabs for 'Postman' and 'Albert Einstein - Wikipedia'. The URL field contains 'chrome-extension://fdmmpjlgnpjigdojojpjoooidkmcomcm/index.html'. Below the URL, there are tabs for 'Normal', 'Basic Auth', 'Digest Auth', 'OAuth 1.0', and 'No environment'. The main request area shows a GET method being used to access 'http://en.wikipedia.org/wiki/Albert_Einstein'. There are buttons for 'Send', 'Preview', 'Add to collection', and 'Reset'. The response status is '200 OK' with a time of '474 ms'. The 'Body' tab is selected, displaying the HTML code of the Wikipedia page for Albert Einstein. The code is as follows:

```
1 <!DOCTYPE html>
2 <html lang="en" dir="ltr" class="client-nojs">
3   <head>
4     <meta charset="UTF-8" />
5     <title>Albert Einstein - Wikipedia, the free encyclopedia</title>
6     <meta http-equiv="X-UA-Compatible" content="IE=EDGE" />
7     <meta name="generator" content="MediaWiki 1.24wmf1" />
8     <link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Albert_Einstein" />
9     <link rel="apple-touch-icon" href="//bits.wikimedia.org/apple-touch/wikipedia.png" />
10    <link rel="shortcut icon" href="//bits.wikimedia.org/favicon/wikipedia.ico" />
11    <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia
12 (en)" />
13    <link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
14    <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
```

At the bottom left, there is a Creative Commons BY-SA license logo.

What happened?



Common links

RSS Feed

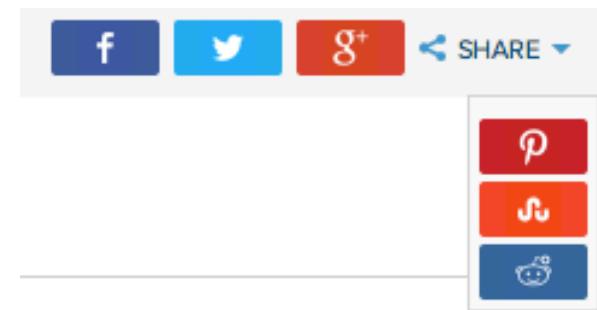


Languages



Printable version

Social channels



Error Codes

- 100: Go
- 200: Got it
- 300: Not me
- 400: Your problem
- 500: Maybe it's my problem



Try again



application/atom+xml

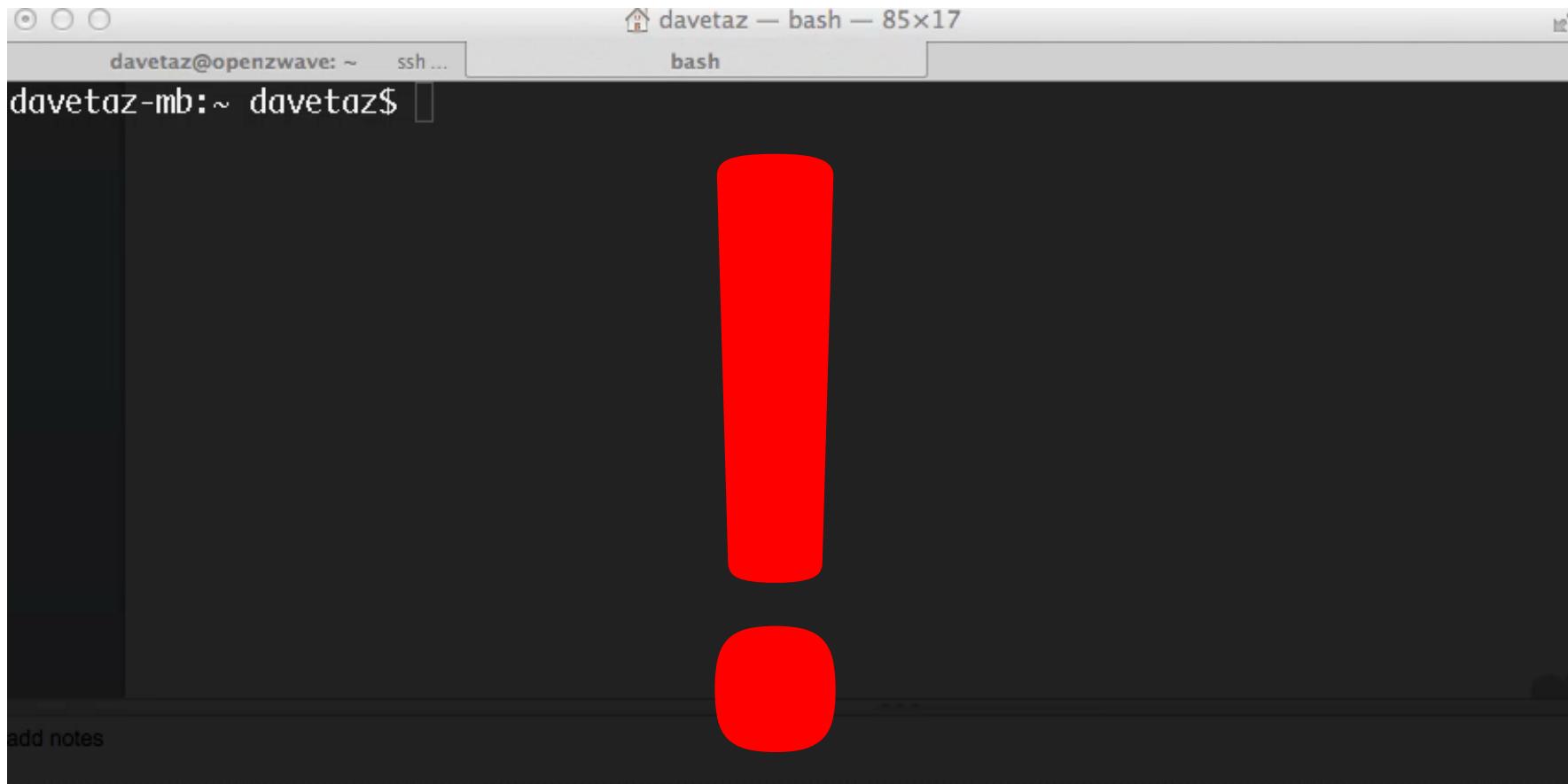
The screenshot shows the Postman application interface. At the top, there's a header bar with tabs for 'Normal', 'Basic Auth', 'Digest Auth', 'OAuth 1.0', and 'No environment'. Below the header, the URL 'http://www.wikipedia.org/wiki/Albert_Einstein' is entered, and the method is set to 'GET'. There are buttons for 'URL params' and 'Headers (0)'. Under the 'Header' section, there's a single entry: 'Content-Type' with the value 'application/atom+xml'. Below the header section are buttons for 'Send', 'Preview', 'Add to collection', and 'Reset'. The status bar at the bottom shows 'STATUS 200 OK' and 'TIME 517 ms'. The main body area displays the raw HTML code of the Wikipedia page for Albert Einstein. On the far left, there are icons for Creative Commons Attribution-ShareAlike (CC BY SA) and a right-pointing arrow.

```
1 <!DOCTYPE html>
2 <html lang="en" dir="ltr" class="client-nojs">
3   <head>
4     <meta charset="UTF-8" />
5     <title>Albert Einstein - Wikipedia, the free encyclopedia</title>
6     <meta http-equiv="X-UA-Compatible" content="IE=EDGE" />
7     <meta name="generator" content="MediaWiki 1.24wmf1" />
8     <link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Albert_Einstein" />
9     <link rel="apple-touch-icon" href="//bits.wikimedia.org/apple-touch/wikipedia.png" />
10    <link rel="shortcut icon" href="//bits.wikimedia.org/favicon/wikipedia.ico" />
11    <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia
(en)" />
12    <link rel="EditURI" type="application/rsd+xml" href="https://en.wikipedia.org/w/api.php?action=rsd" />
13    <link rel="copyright" href="https://creativecommons.org/licenses/by-sa/3.0/" />
```

and again



application/atom+xml



A screenshot of a terminal window titled "davetaz — bash — 85x17". The window shows a command-line interface with the prompt "davetaz-mb:~ davetaz\$". A large red exclamation mark (!) is displayed prominently in the center of the terminal window. At the bottom left of the terminal window, there is a small text "add notes".



Common links



RSS Feed



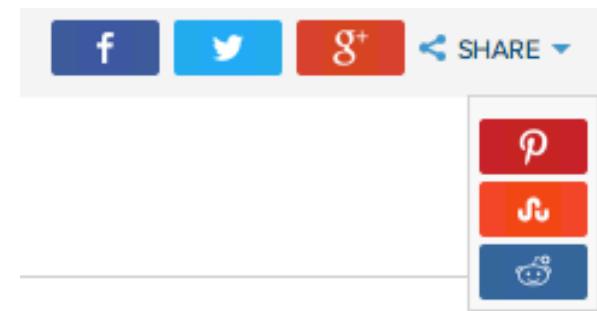
Languages



Printable version



Social channels



So what do you have?

The screenshot shows the Postman application interface. The header bar includes the Postman logo, a tab for 'Albert Einstein - Wikipedia', and various extension icons. The main interface shows a GET request to http://www.wikipedia.org/wiki/Albert_Einstein. The 'Body' tab is selected, displaying the HTML source code of the Wikipedia page for Albert Einstein. The code includes standard HTML tags like <html>, <head>, and <body>, along with meta tags for charset, title, and generator, as well as links for alternate, apple-touch-icon, shortcut icon, and search.

```
1 <!DOCTYPE html>
2 <html lang="en" dir="ltr" class="client-nojs">
3   <head>
4     <meta charset="UTF-8" />
5     <title>Albert Einstein - Wikipedia, the free encyclopedia</title>
6     <meta http-equiv="X-UA-Compatible" content="IE=EDGE" />
7     <meta name="generator" content="MediaWiki 1.24wmf1" />
8     <link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/Albert_Einstein" />
9     <link rel="apple-touch-icon" href="//bits.wikimedia.org/apple-touch/wikipedia.png" />
10    <link rel="shortcut icon" href="//bits.wikimedia.org/favicon/wikipedia.ico" />
11    <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia
(en)" />
12    <link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
13    <link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
```



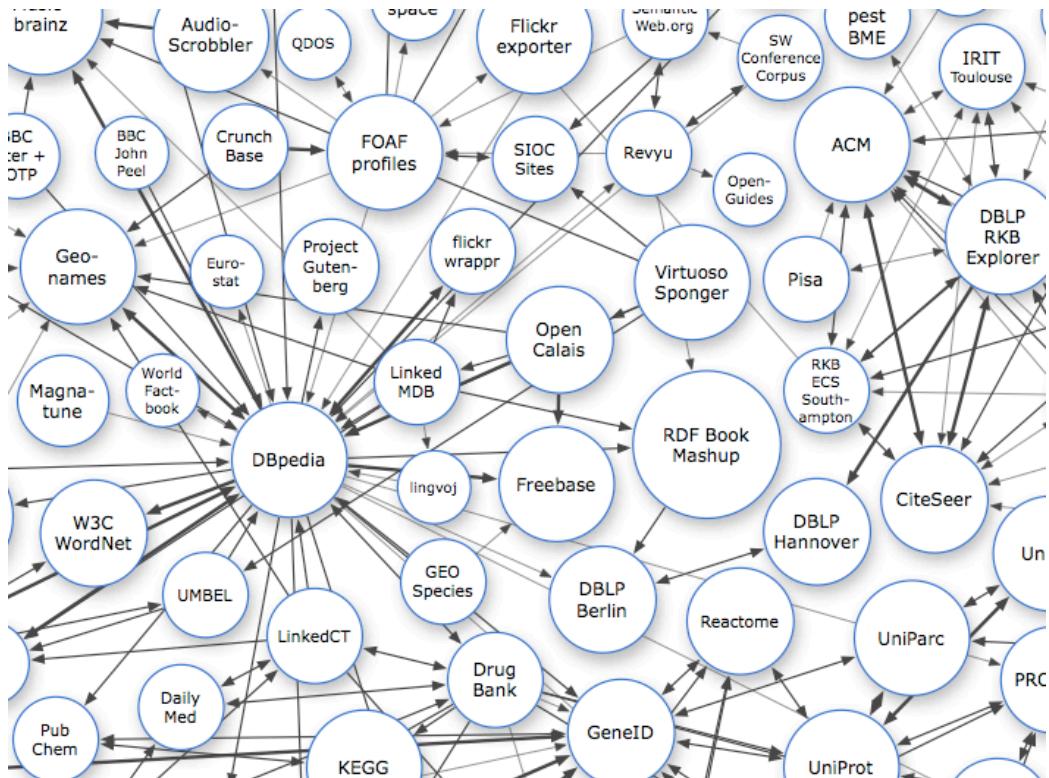
The <link rel=...> tag

So who and what is the <link rel=...> tag for:

1. To stop the number of HTTP headers forever expanding?
2. To add custom headers? (Even though that is what X-* headers are)
3. Should machines have to search all the possible link-rel headers related to a resource and other links. (recurse, recurse, recurse)

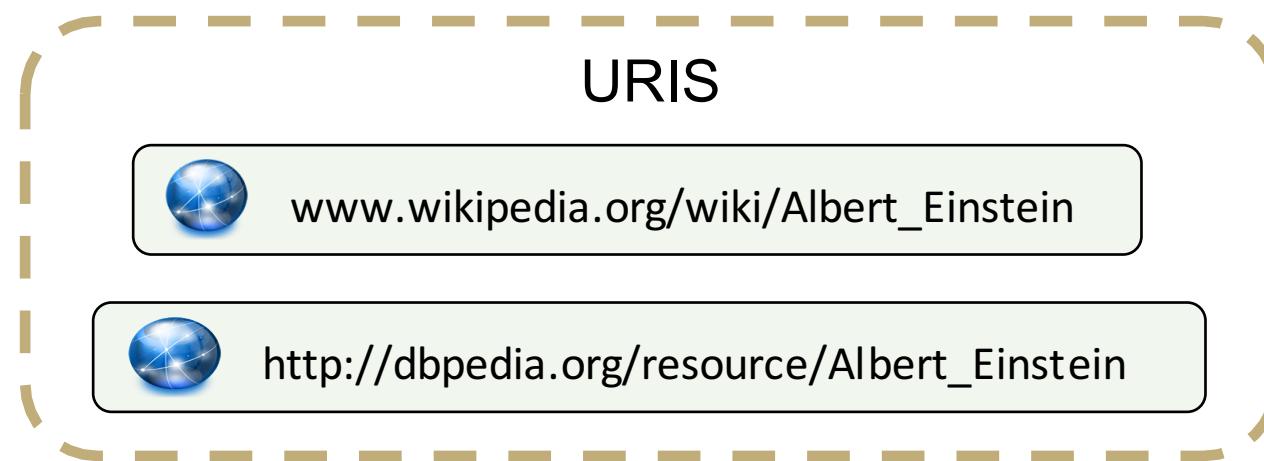


It's not all doom and gloom



The data version of wikipedia,
available in a machine friendly way.

Wikipedia & dbpedia



Postman Google

https://www.google.co.uk/?gws_rd=cr&ei=_kdeU8q3D47-PMaKgagP

+David Mail Images

Share

1



Google Search I'm Feeling Lucky

Updated Privacy & Terms Settings Use Google

Advertising Business About



davetaz — bash — 85x17

davetaz@openzwave: ~ ssh ...

bash

bash

davetaz-mb:~ davetaz\$ wikipedia.org/wiki/Albert_Einstein

es.wikipedia.org/wiki/Albert_Einstein

URIS



www.wikipedia.org/wiki/Albert_Einstein



http://dbpedia.org/resource/Albert_Einstein



add notes



Common links



RSS Feed



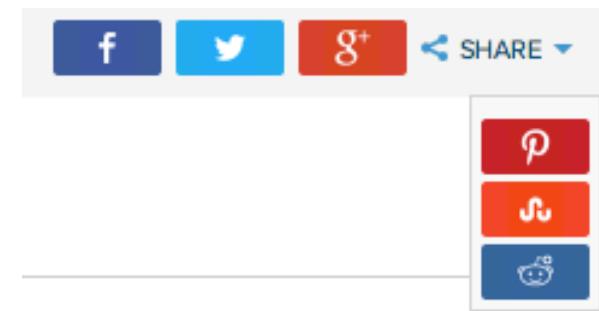
Languages



Printable version



Social channels



Linked data

Amazing but hard to publishing and use.

EEE Building

<http://id.southampton.ac.uk/building/32> ← This is the URI

Detail	Facilities	Services	Energy
--------	------------	----------	--------

Site: Highfield Campus
Construction: 2006
Architect: John McAslan & Partners
Features: Building 32 is non-residential

[View Disability Report for this Building](#)

Occupants

Electronics & Computer Science
Southampton Education School
Agents, Interactions & Complexity
Web & Internet Science
Leadership School Improve & Effectiveness
Lifelong & Work-Related Learning
Mathematics & Science Education
Social Justice & Inclusive Education
Teaching Only Staff
Deanery



©2010 Francois-Xavier Beckers (CC-BY)



ton.ac.uk/building/32
→ rooms:Building, <http://id.southampton.ac.uk/ns/UoSBuilding>
→ "EEE Building"
occupant → Electronics & Computer Science, Southampton Education School, Agents, Interactions & Complexity, Web & Internet Science, ...show 8 more...
tion → "32"^^<http://id.southampton.ac.uk/ns/building-code-scheme>
ations:within → Highfield Campus
→ <http://www.soton.ac.uk/estates/ourestate/buildings/highfield/32.html>
→ "50.9364157"^^xsd:float
→ "-1.395905"^^xsd:float
southampton.ac.uk/ns/disabledGoPage → <http://www.disabledgo.com/en/access-guide/building-32>
ations:easting → "442544"^^xsd:integer
ations:northing → "115392"^^xsd:integer
Organization → University of Southampton
southampton.ac.uk/ns/ombielName → "Bldg 32 (EEE)"
eature → Building 32 is non-residential
southampton.ac.uk/ns/buildingDate → "2006"
southampton.ac.uk/ns/buildingArchitect → John McAslan & Partners
spatial → "POLYGON((-1.3961073411331264 50.93683868764933,-1.3958347895092957 50.9368567227702,-1.3956958407975968 50.936065737417,-1.3959558923017397 50.93603859197583,-1.3961073411331264 50.93683868764933))"
southampton.ac.uk/ns/electricityTimeSeries → "elec/b32/ekw"
← is spatialrelations:within of ← 32 / 3077, 32 / 1015, Physical and Applied Science Faculty Deanery, Social and Human Sciences Faculty Deanery, ...show 54 more...
← is foaf:depicts of ← <http://data.southampton.ac.uk/image-archive/buildings/raw/32.jpg>, <http://data.southampton.ac.uk/image-archive/buildings/1000/32.jpg>, <http://data.southampton.ac.uk/image-archive/buildings/800/32.jpg>, <http://data.southampton.ac.uk/image-archive/buildings/600/32.jpg>, ...show 5 more...
← is event:place of ← RAE Solent Branch Christmas Special Lecture - The Red Arrows



Browsing the web of data

[RDF Browser](#) | [SPARQL Browser](#)

Quick and Dirty RDF browser

URI:

[Browse RDF](#)

Here are some suggestions:

- <http://education.data.gov.uk/id/school/118217>
- <http://eprints.ecs.soton.ac.uk/id/eprint/10053>
- <http://id.southampton.ac.uk/building/59>
- <http://data.totl.net/playingcards/>
- <http://dbpedia.org/resource/Southampton>



Application Programming Interfaces (APIs)



Exercise

What is an API?
Any examples?



What is an API

Defines how one application
can **consistently** interact with
another.



Teaching nerds passion

```
function makeOut(passionLevel, partsOfBody) {  
    for (each partOfBody in partsOfBody) {  
        partOfBody.kiss(passionLevel);  
        lookIntoEyes();  
        sighDeeply();  
    }  
    moanDaintily();  
    sleep();  
}
```



Thanks to Pamela Fox (blog.pamelafox.org)

Making your call

```
function makeOut(passionLevel, partsOfBody) {  
  for (each partOfBody in partsOfBody) {  
    partOfBody.kiss(passionLevel);  
    lookIntoEyes();  
    s    makeOut(10, ["neck", "ear", "mouth"]);  
  }  
  moanDaintily();  
  sleep();  
}
```



Thanks to Pamela Fox (blog.pamelafox.org)

The Webs API

```
class Photos {  
    function search(api_key, tags) {  
        if (!validKey(api_key)) {return(401);}  
        ...  
        ...  
        return results;  
    }  
}
```



Using the Webs API

http://api.flickr.com/services/rest/?method=flickr.photos.search&api_key=a8c42ef2ac8d88aed7e351f95e93160f&tags=fox



Evolution of Web APIs

- Long URLs are not easy to understand
- Flickr API using query parameters to define everything!

?method=...&auth=...&query=...

- There are better ways



Web APIs (before)

http://api.flickr.com/services/rest/?method=flickr.photos.comments.getList&api_key=a8c42ef2ac8d88aed7e351f95e93160f&photo_id=13992401185



A better solution?

`https://www.flickr.com/photos/{user}/{photo}/comments
/tags
/licence`

`https://www.flickr.com/search?tags=...`



API Recap

A **promise** by one application to service another.

The underlying code can change separately to the API...

So far we have looked at READ only



Discussion

What's the biggest
problem with APIs like
flickr?





Aggregator/Enricher

- Take data from hundreds of transport companies
- Bring it together
- Align it under one easy to use API



Exercise

Using REST APIs to GET data



<http://training.theodi.org/UnlockingData/>



Session 2

Processing data



Outcomes

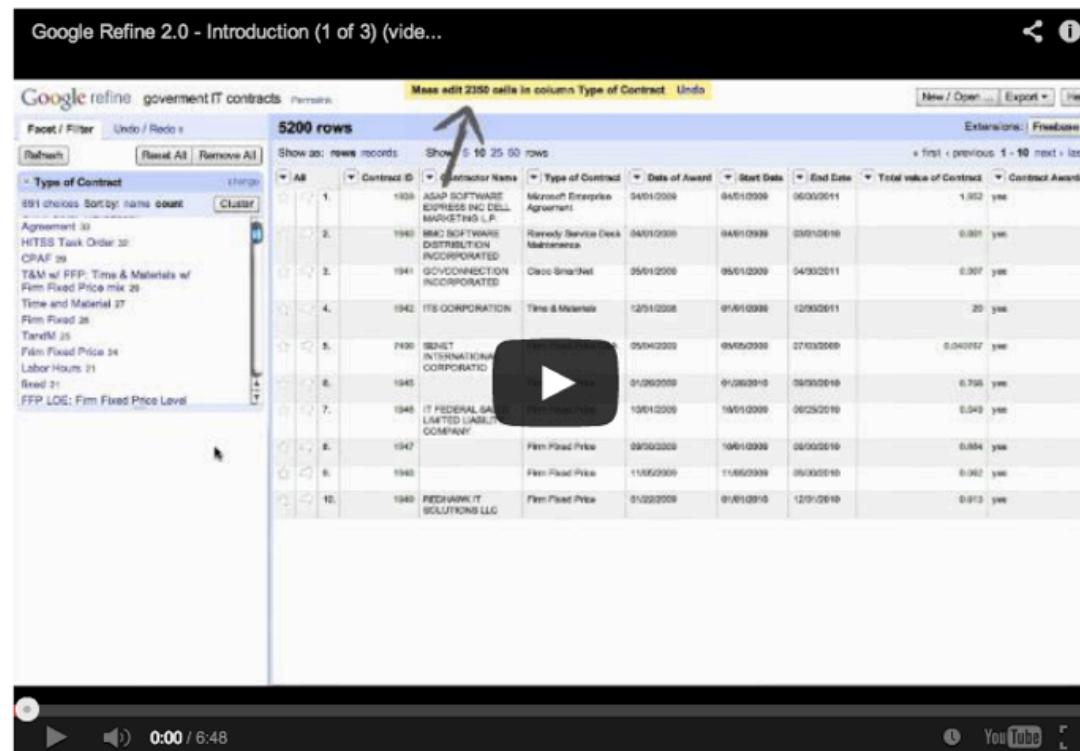
Clean and enrich data

Create a number of data processing pipelines

Translate data between a number of formats



Introducing Open Refine



The screenshot shows the Google Refine 2.0 interface with the title "Google Refine 2.0 - Introduction (1 of 3) (vide...)" at the top. A yellow bar at the top indicates "Mass edit 2150 cells in column Type of Contract: Undo". The main area displays a table with 5200 rows of contract data. The columns include Contract ID, Contractor Name, Type of Contract, Date of Award, Start Date, End Date, Total value of Contract, and Contract Awarded. A facet sidebar on the left lists various types of contracts such as "Agreement", "ITBS Task Order", "CPAF", "T&M w/ FFP: Time & Materials w/ Firm Fixed Price mix", "Time and Material", "Firm Fixed", "TandM", "Firm Fixed Price", "Labor Hours", "Breed", and "FFP LOE: Firm Fixed Price Level". A large black play button is overlaid on the table. At the bottom, there is a video player with controls, showing "0:00 / 6:48" and the YouTube logo.

Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	End Date	Total value of Contract	Contract Awarded
1. 1938	ASAP SOFTWARE EXPRESS INC CELL MARKETING L.P.	Microsoft Enterprise Agreement	04/01/2008	04/01/2008	05/01/2011	1,952	yes
2. 1940	RMC SOFTWARE DISTRIBUTION INCORPORATED	Randy Service Desk Maintenance	04/01/2008	04/01/2008	03/01/2010	0.001	yes
3. 1941	GOVCONNECTION INCORPORATED	Class Standard	05/01/2008	05/01/2008	04/30/2011	0.007	yes
4. 1942	ITS CORPORATION	Time & Materials	12/01/2008	01/01/2009	02/01/2011	20	yes
5. 1943	ISINET INTERNATIONAL CORPORATION	Software License	05/04/2008	05/05/2008	07/03/2009	0.04072	yes
6. 1945		Time & Materials	01/05/2009	01/05/2009	05/05/2010	0.705	yes
7. 1946	IT FEDERAL SA LIMITED LIABILITY COMPANY	Software License	10/01/2008	10/01/2008	08/25/2010	0.049	yes
8. 1947		Firm Fixed Price	09/09/2008	10/01/2008	01/01/2010	0.004	yes
9. 1948		Firm Fixed Price	11/05/2008	11/05/2008	06/03/2010	0.002	yes
10. 1949	RECHAWK IT SOLUTIONS LLC	Firm Fixed Price	01/03/2009	01/01/2010	12/01/2010	0.012	yes

<http://openrefine.org>



Data processing pipelines

ODI Experiment

Refine AutoBot

Refine AutoBot

Enter the URL of the CSV file for cleaning.

Enter the refine operation history.

Submit

The screenshot shows a web-based application titled "Refine AutoBot". At the top left is the title "Refine AutoBot". To its right is a small orange box containing the text "ODI Experiment". On the far right is the logo for the "open data institute" with the letters "ODI" in white inside a dark blue square. The main body of the application has a light gray background. It contains two input fields. The first field is labeled "Enter the URL of the CSV file for cleaning." and has a placeholder URL "http://theodi.github.io/refine-autobot/". The second field is labeled "Enter the refine operation history." and is currently empty. At the bottom center is a large blue button with the word "Submit" in white. The entire application is set against a dark gray background.

<http://theodi.github.io/refine-autobot/>



The Open Database Of The Corporate World

We have information on
70,597,888 companies

Aggregator/Enabler

search companies search officers

SEARCH

Filter by jurisdiction

1,298 Abu Dhabi (UAE)

144,755 Alaska (US)

40,157 Albania

899,455 Arizona (US)

46,537 Aruba

165,582 Bahamas

99,185 Bahrain

88,563 Bangladesh

... [View all](#)

Just released:
OpenCorporates API v0.3

Corporate network data,
financial accounts, complex
filters, and more. [Read more](#)

Get data access to over
60 million companies

- | | | |
|--|---|--|
| Open data | Quality data | Unique data |
| <ul style="list-style-type: none">• All the data on the world's largest
corporate network• Available as either raw data
or through open data API commercially• All data includes source, allowing
cross-checking of data | <ul style="list-style-type: none">• Data is from primary public sources• Fully processed, providing
consistent and reliable data• Many users on OpenCorporates
with quality controls for free• Measured with | <ul style="list-style-type: none">• Open, transparent and highly
granular• Many unique datasets and sources• Extend your existing data or built
new services |

Announcing Open LEIs

Today, OpenCorporates
announces a new sister website,
[Open LEIs](#), a user-friendly
interface on the emerging Global
Legal Entity Identifier System.

[Read more](#)

OPENLEIs

A BETA VIEW ON THE LEI SYSTEM

New! Just added: Open
corporate network data

[Read more](#) about this important
new feature



Exercise

Enriching a dataset containing
company names (e.g. transactions)
with company data from
OpenCorporates



Session 3

Publishing insight



Outcome

Build simple web pages that bring together a number of datasets to reveal new insight



Human Readable Web



World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) .

What's out there?

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

Help

on the browser you are using

Software Products

A list of W3 project components and their current state. (e.g. [Line Mode](#) ,[X11 Viola](#) , [NeXTStep](#) , [Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

Technical

Details of protocols, formats, program internals etc

Bibliography

Paper documentation on W3 and references.

People

A list of some people involved in the project.

History

A summary of the history of the project.

How can I help ?

If you would like to support the web..

Getting code

Getting the code by [anonymous FTP](#) , etc.



Question

What is significant about the first web page?

What actually did Tim invent?



Markup Links & Tags



World Wide Web

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) .

[What's out there?](#)

Pointers to the world's online information, [subjects](#) , [W3 servers](#), etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) , [X11 Viola](#) , [NeXTStep](#) ,
[Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

If you would like to support the web..

[Getting code](#)

Getting the code by [anonymous FTP](#) , etc.



Markup

Links & Tags

<h>World Wide Web </h>

<p>The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents. </p>

<p>Everything there is online about W3 is linked directly or indirectly to this document, including an [executive summary](#) of the project, [Mailing lists](#) , [Policy](#) , November's [W3 news](#) , [Frequently Asked Questions](#) . </p>

<list>[What's out there?](#)

Pointers to the world's online information, [subjects](#) , [W3 servers](#) , etc.

[Help](#)

on the browser you are using

[Software Products](#)

A list of W3 project components and their current state. (e.g. [Line Mode](#) , [X11 Viola](#) , [NeXTStep](#) , [Servers](#) , [Tools](#) , [Mail robot](#) , [Library](#))

[Technical](#)

Details of protocols, formats, program internals etc

[Bibliography](#)

Paper documentation on W3 and references.

[People](#)

A list of some people involved in the project.

[History](#)

A summary of the history of the project.

[How can I help ?](#)

If you would like to support the web..

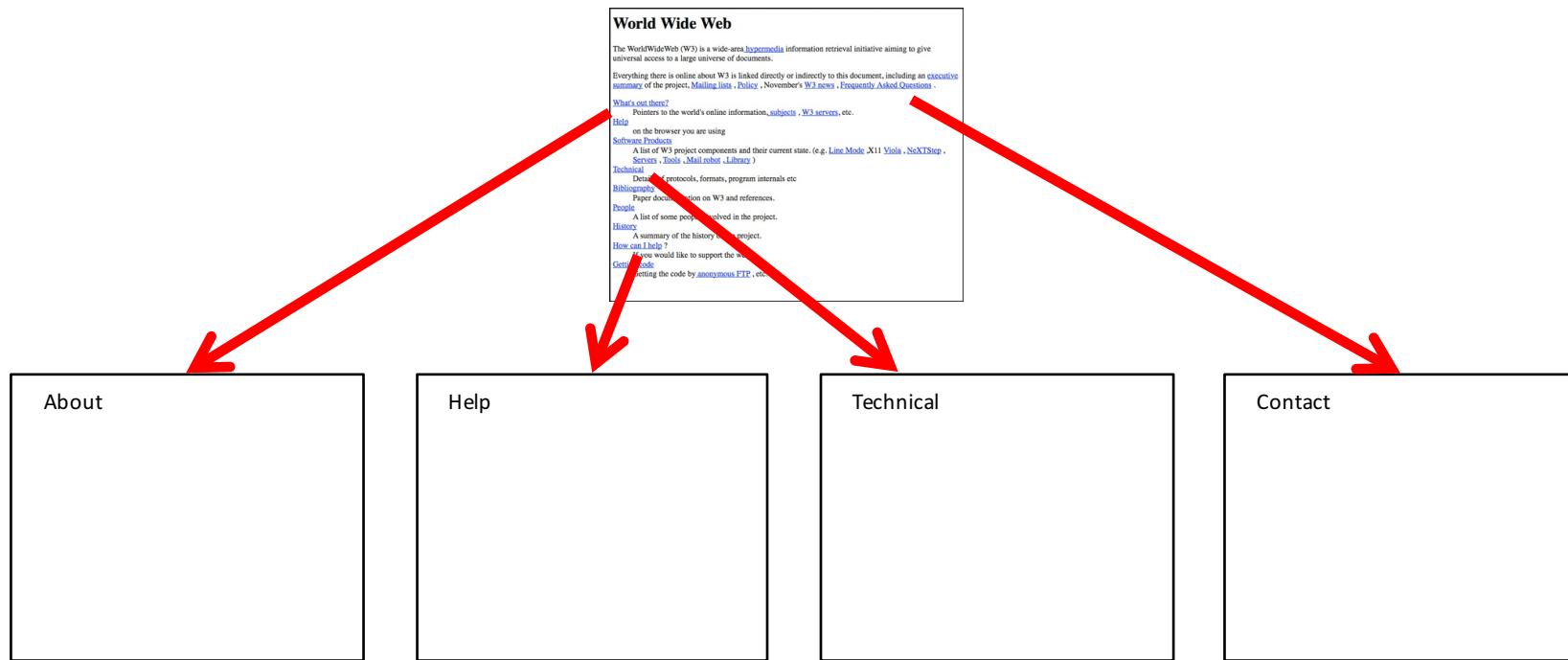
[Getting code](#)

Getting the code by [anonymous FTP](#) , etc.

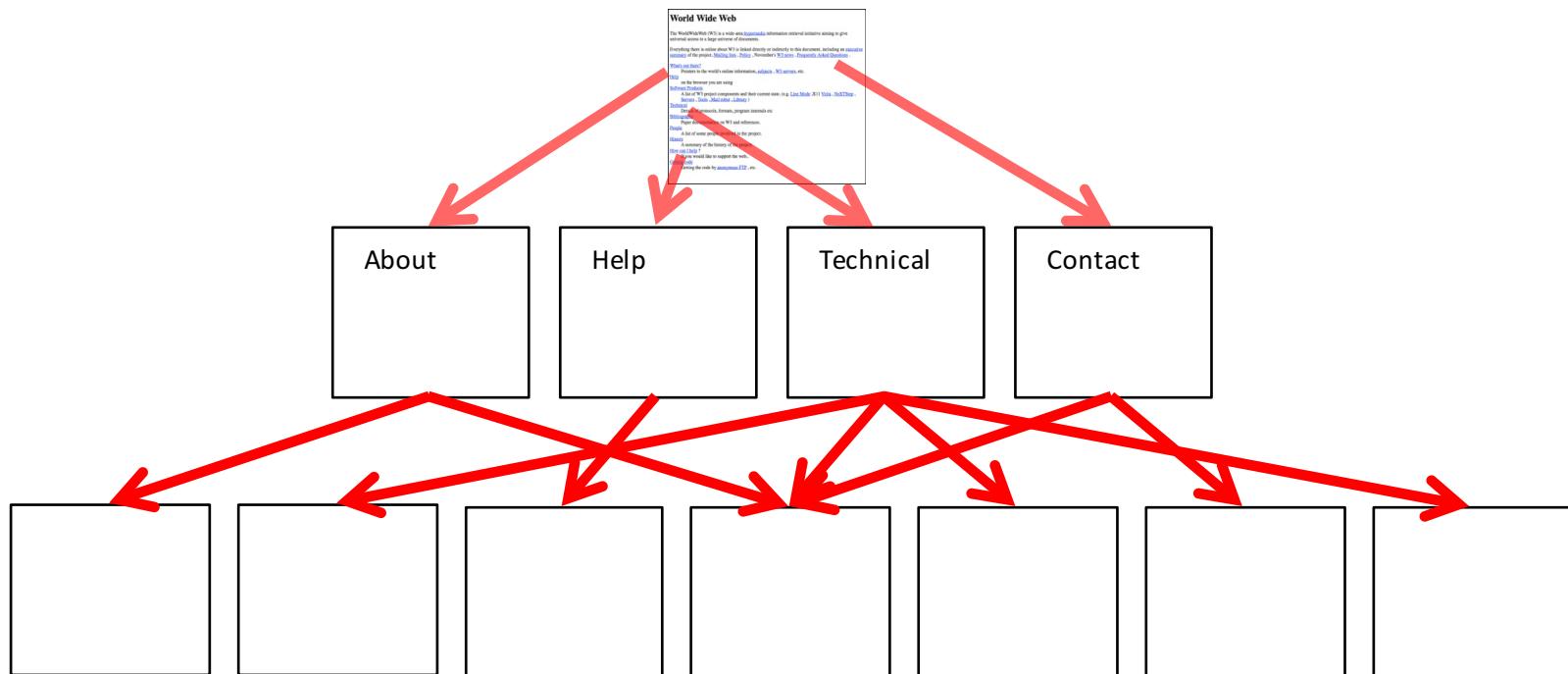
</list>



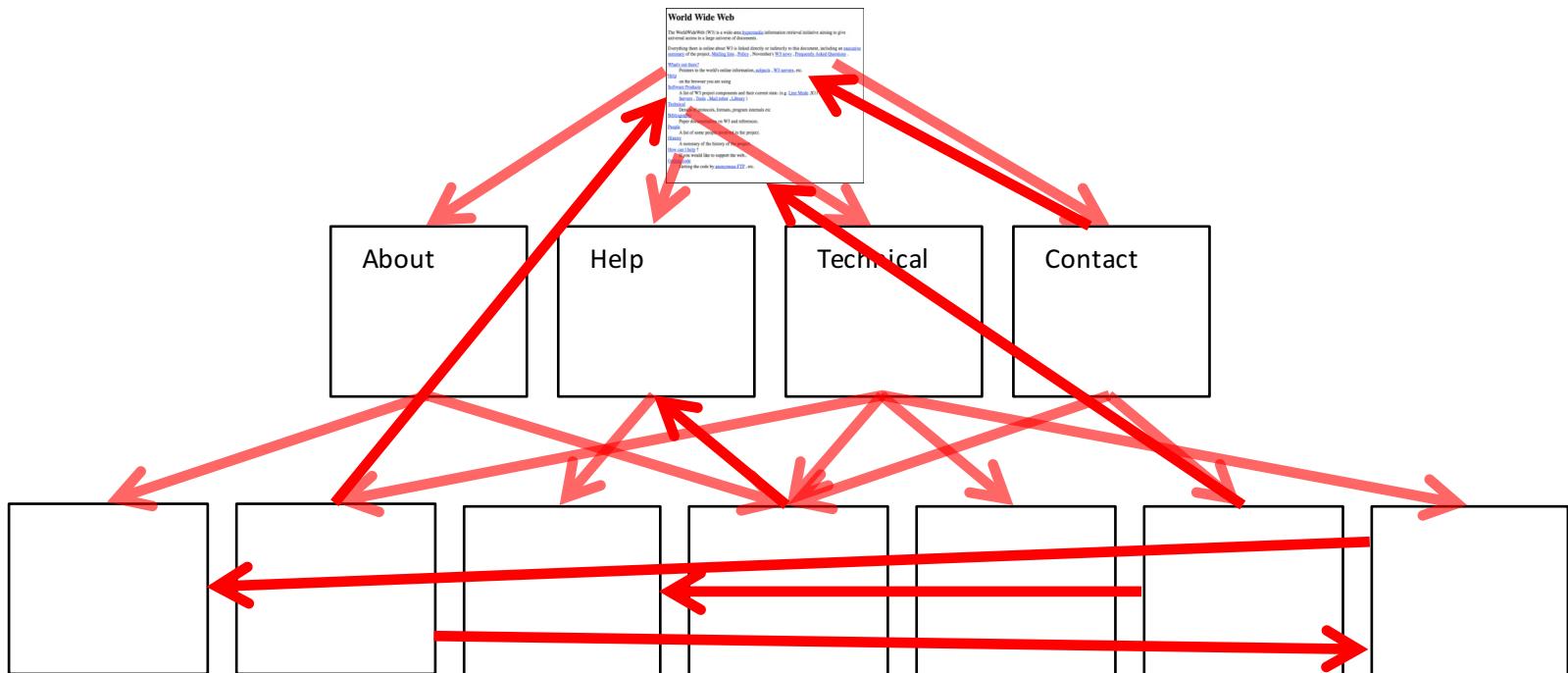
Nodes and Links

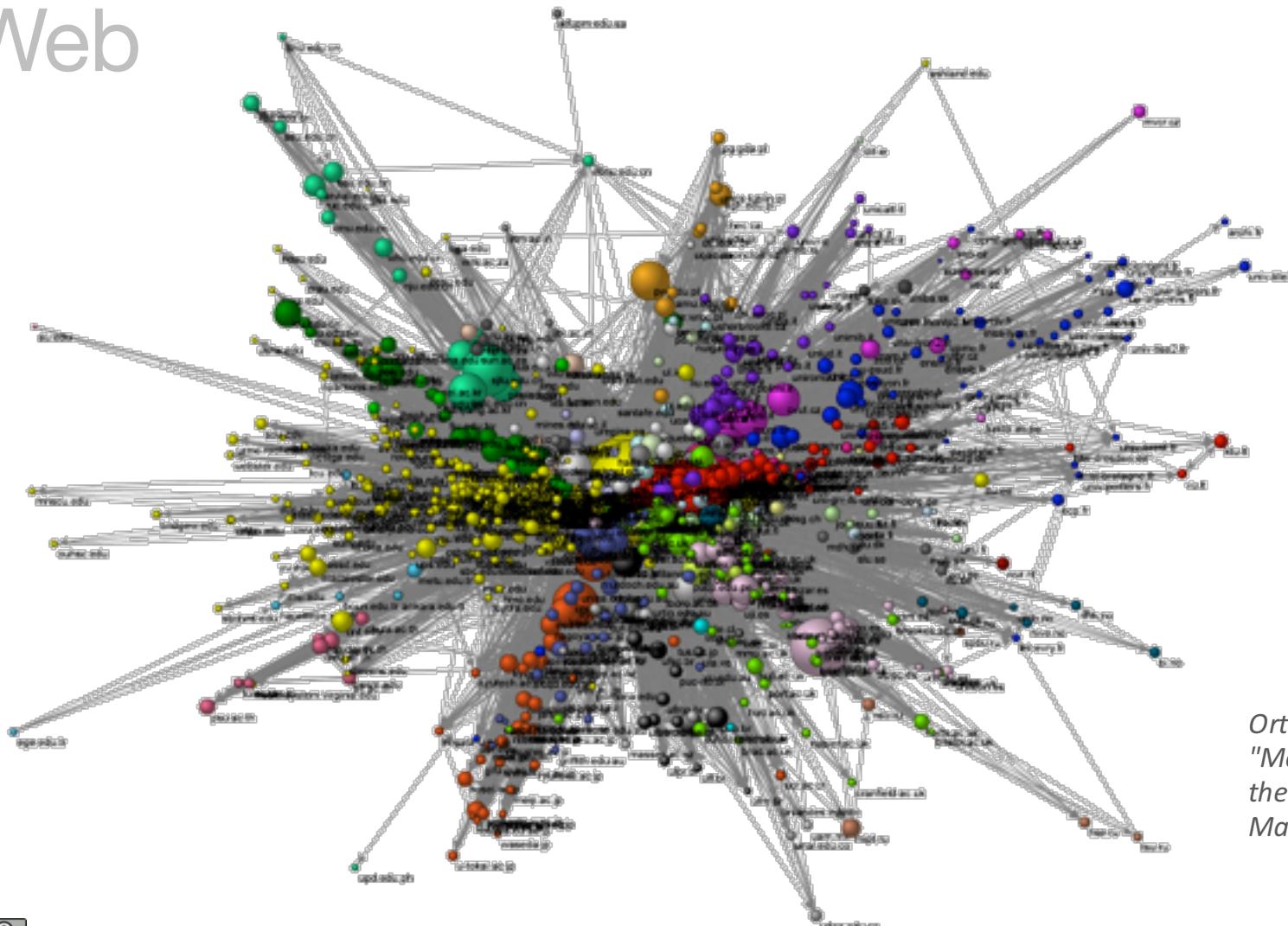


Nodes and Links



Nodes and Links





Ortega, Jose Luis, and Isidro F. Aguillo. "Mapping world-class universities on the web." *Information Processing & Management* 45.2 (2009): 272-279.



HTML

HyperText (Links) Markup Language

Here is some **really important** text!

Remember that in the morning we start at **<time>9:30am</time>**

<blink>

This text is likely to annoy you.

</blink>



HTML

```
<h>World Wide Web </h>
```

The WorldWideWeb (W3) is a wide-area [hypermedia](#) information retrieval initiative aiming to give universal access to a large universe of documents.

```
<h1>This is heading 1</h1>
```

```
<h2>This is heading 2</h2>
```

```
<h3>This is heading 3</h3>
```

```
<h4>This is heading 4</h4>
```

```
<h5>This is heading 5</h5>
```

```
<h6>This is heading 6</h6>
```



A Link

`Link to BBC News`



HTML 5

<menu>

<summary>

<figure>

<details>

<nav>

<legend>

<input>

<label>

<header>

<blink>

<section>

<marquee>

<footer>

<title>



<option>

<a>

HTML5



HTML 5

html																					col	table	
head	span																				div	fieldset	form
title	a																				body	h1	section
meta	rt	dfn	em	i	small	ins	s	br	p	blockquote	legend	optgroup	address	h2	header	caption	td						
base	rp	abbr	time	b	strong	del	kbd	hr	ol	dl	label	option	datalist	h3	nav	menu	th						
link	noscript	q	var	sub	mark	bdi	wbr	figcaption	ul	dt	input	output	keygen	h4	article	command	tbody						
style	script	cite	samp	sup	ruby	bdo	code	figure	li	dd	textarea	button	progress	h5	footer	summary	thead						
																		img	area	map	embed	object	
																		param	source	iframe	canvas	track	
																		audio	video				

HTML5



Why Markup?

Aids your browser to render the page

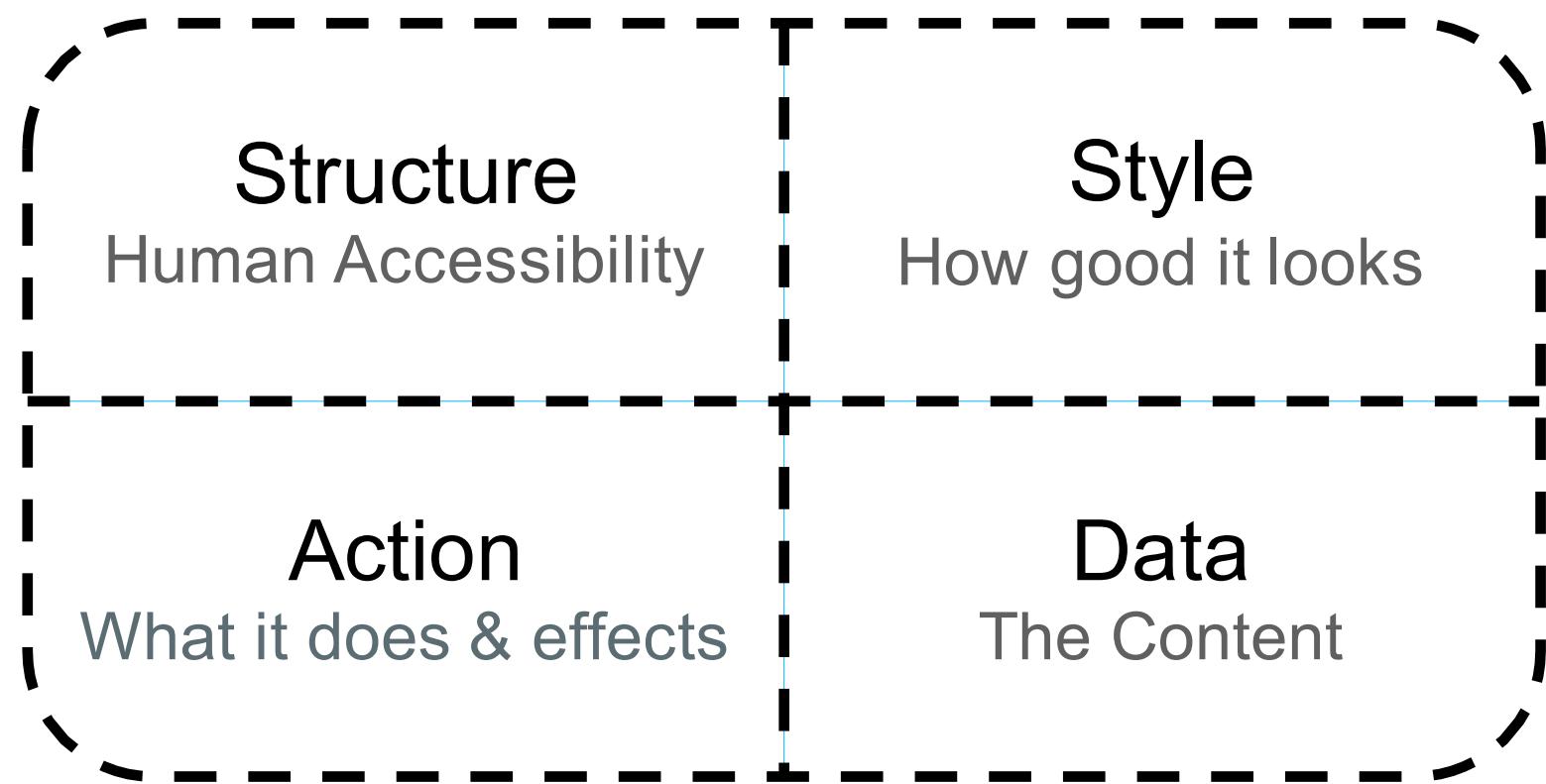
Critical for screen readers!

Adds semantics about importance of elements.

Aids search engines



Building Blocks



The Language of the Web

HTML5



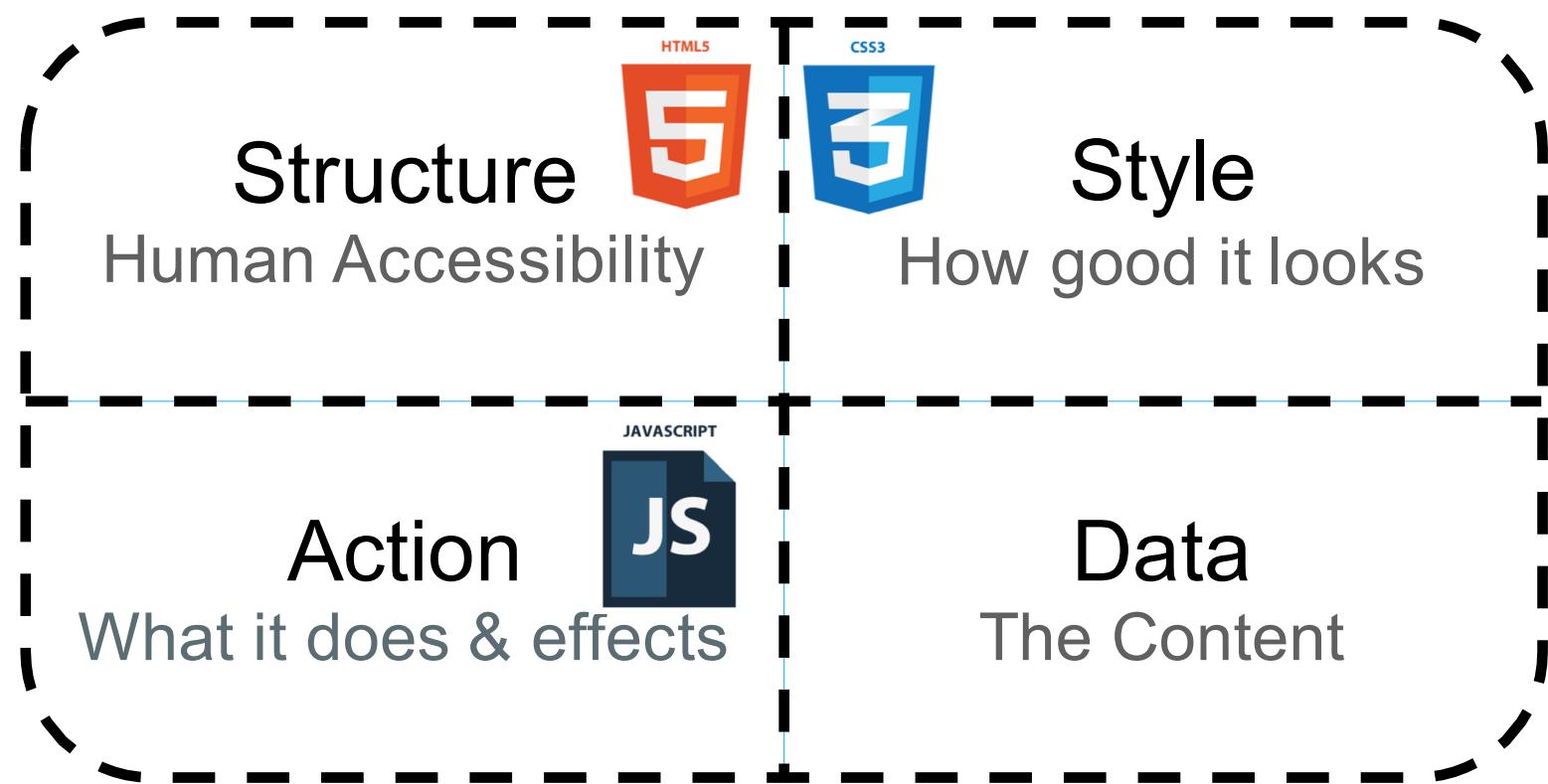
CSS3



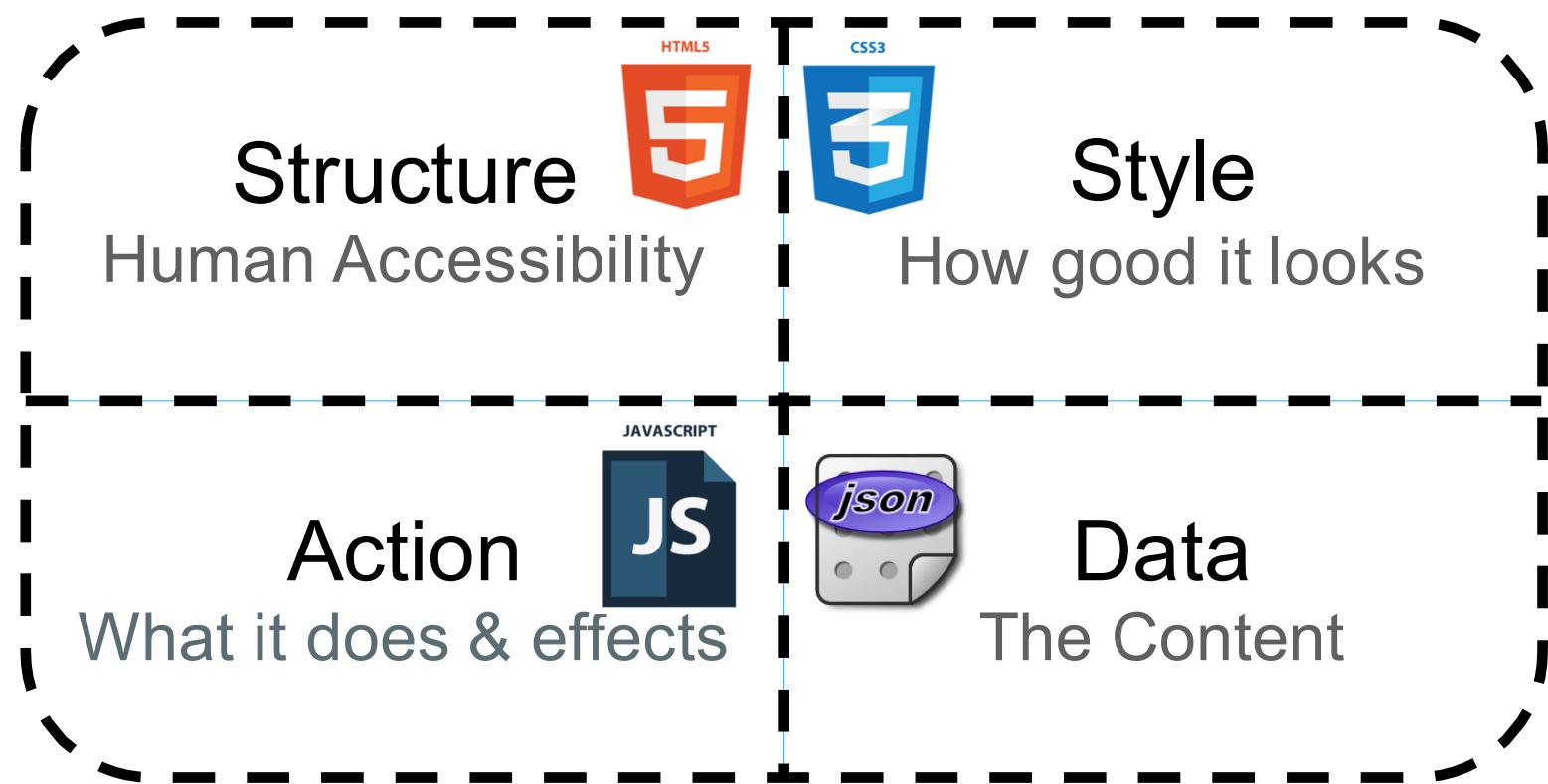
JAVASCRIPT



Building Blocks



JSON: A Natural Fit



Last exercise

Build a web page that shows the time
and platform for your train home*.

<http://training.theodi.org/UnlockingData/>



* I've made a massive assumption that someone catches a train home.



Session 1

Unlocking data from the web

Session 2

Processing data

Session 3

Publishing insight

Outcomes

List and identify the key structures and formats of data

Analyse websites to identify sources of data

Use a number of tools to obtain data from the web



Outcomes

Clean and enrich data

Create a number of data processing pipelines

Translate data between a number of formats



Outcome

Build simple web pages that bring together a number of datasets to reveal new insight





Thank-you