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

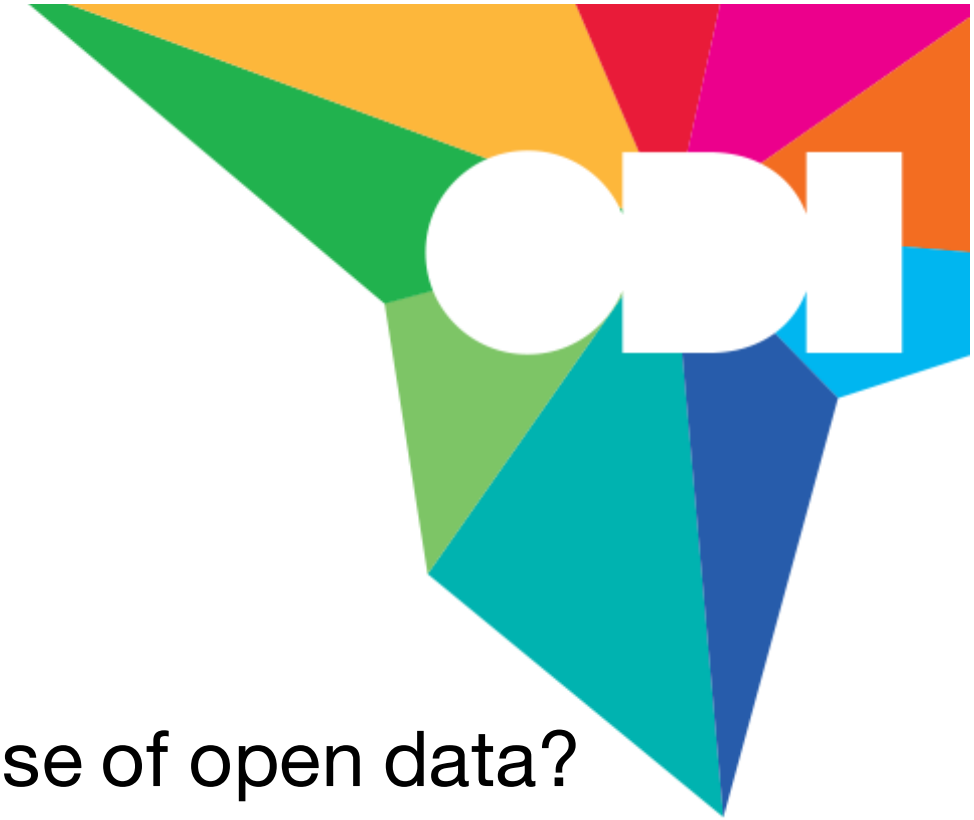
Open Data Science

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?

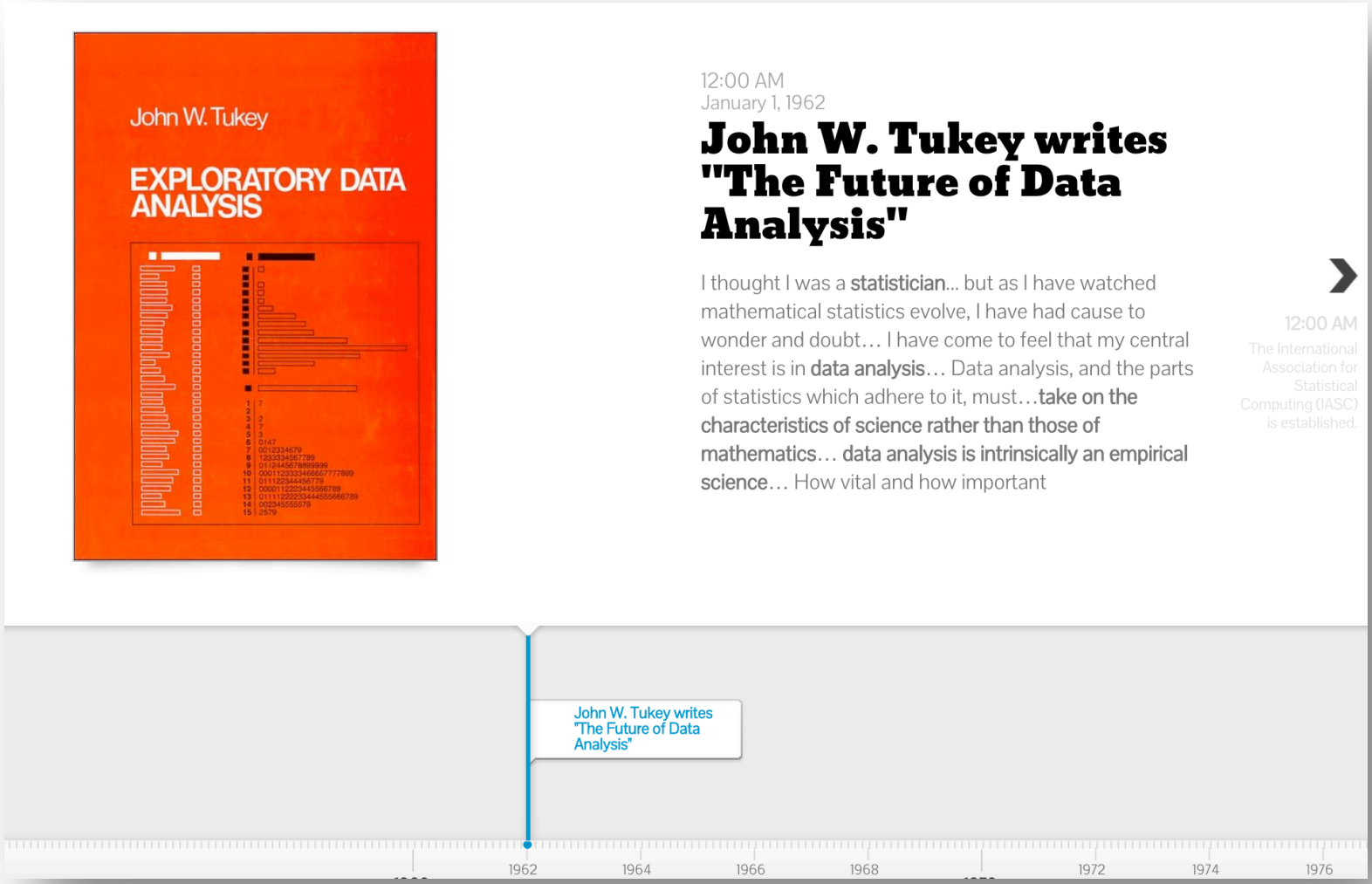
What do you want to do differently after the course?

Course aim

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

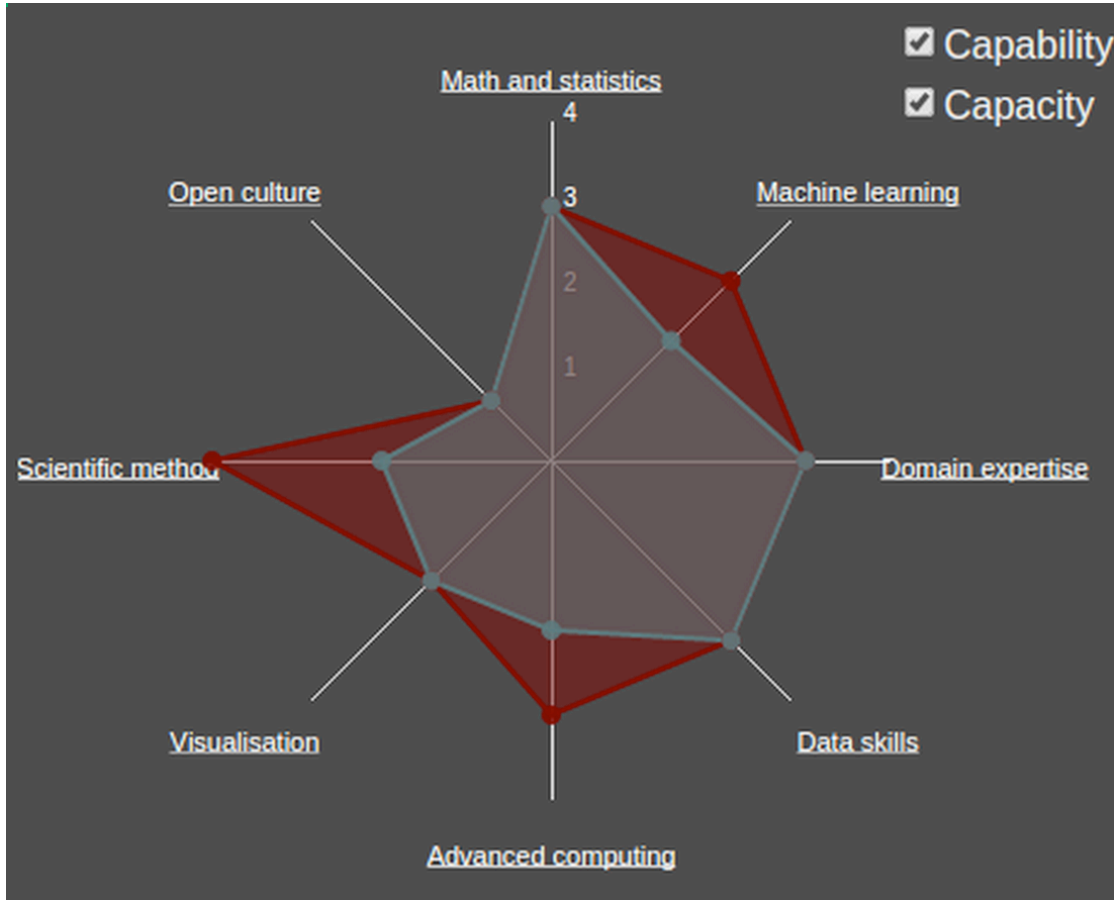
Question

What is data science?



History of data science

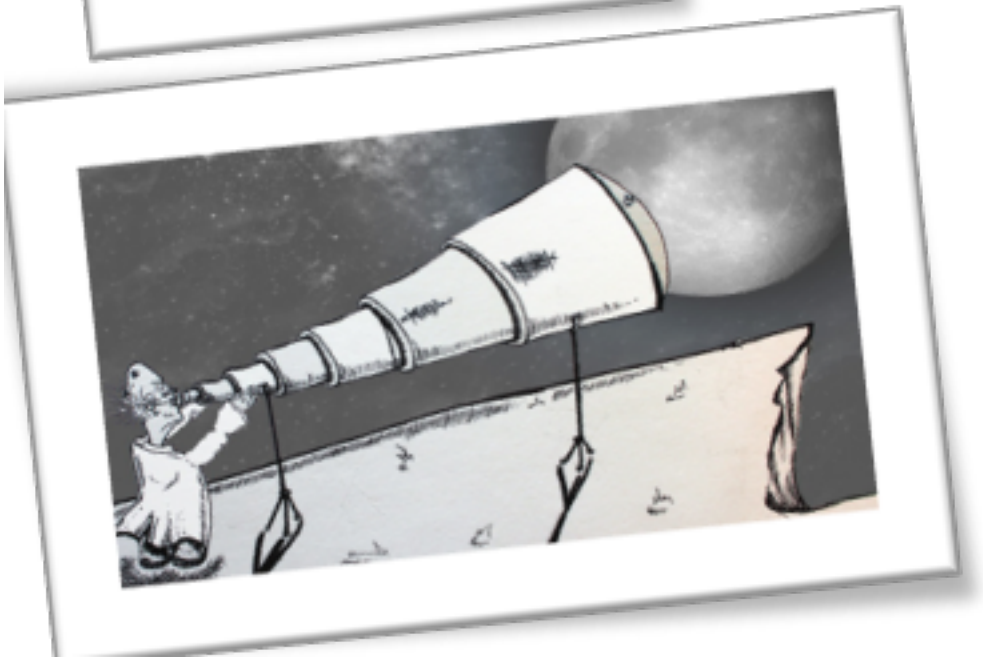
Data science



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

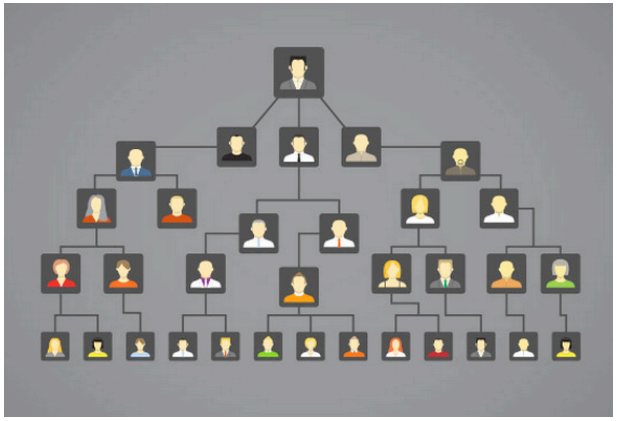
Exploring data formats

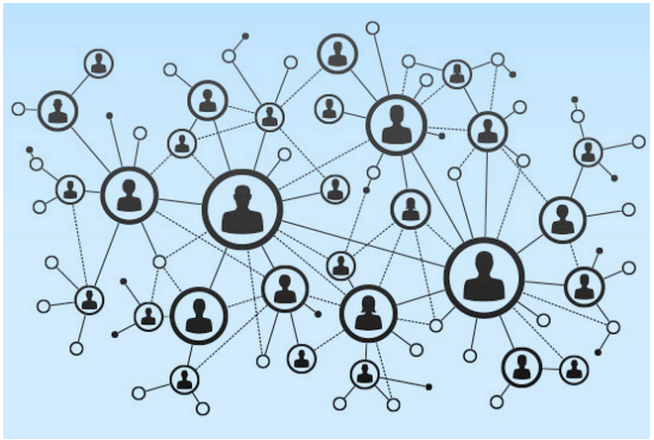
http://training.theodi.org/UnlockingData/

Structures

Tabular

Hierarchical



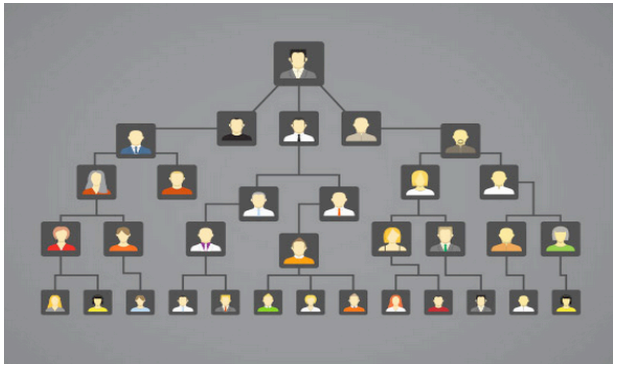


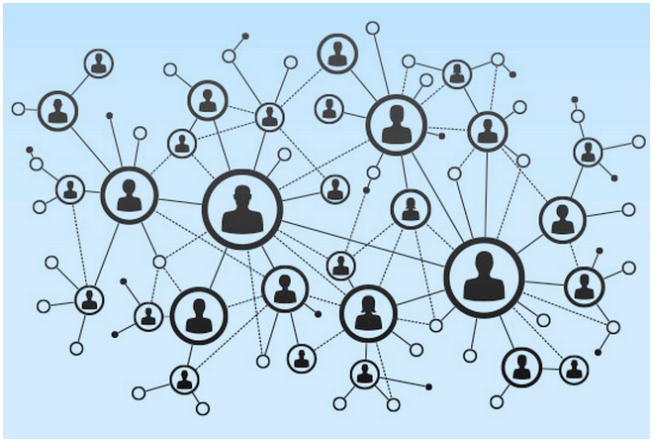
Network

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 worsheets

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

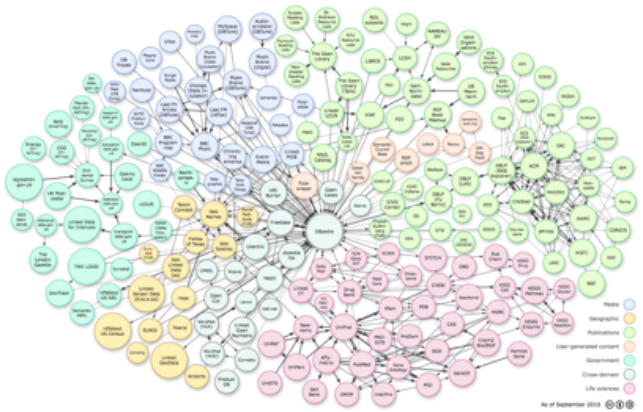
Analyse websites to identify sources of data

Use a number of tools to obtain data from the web

Approaches to publishing data

ON the web

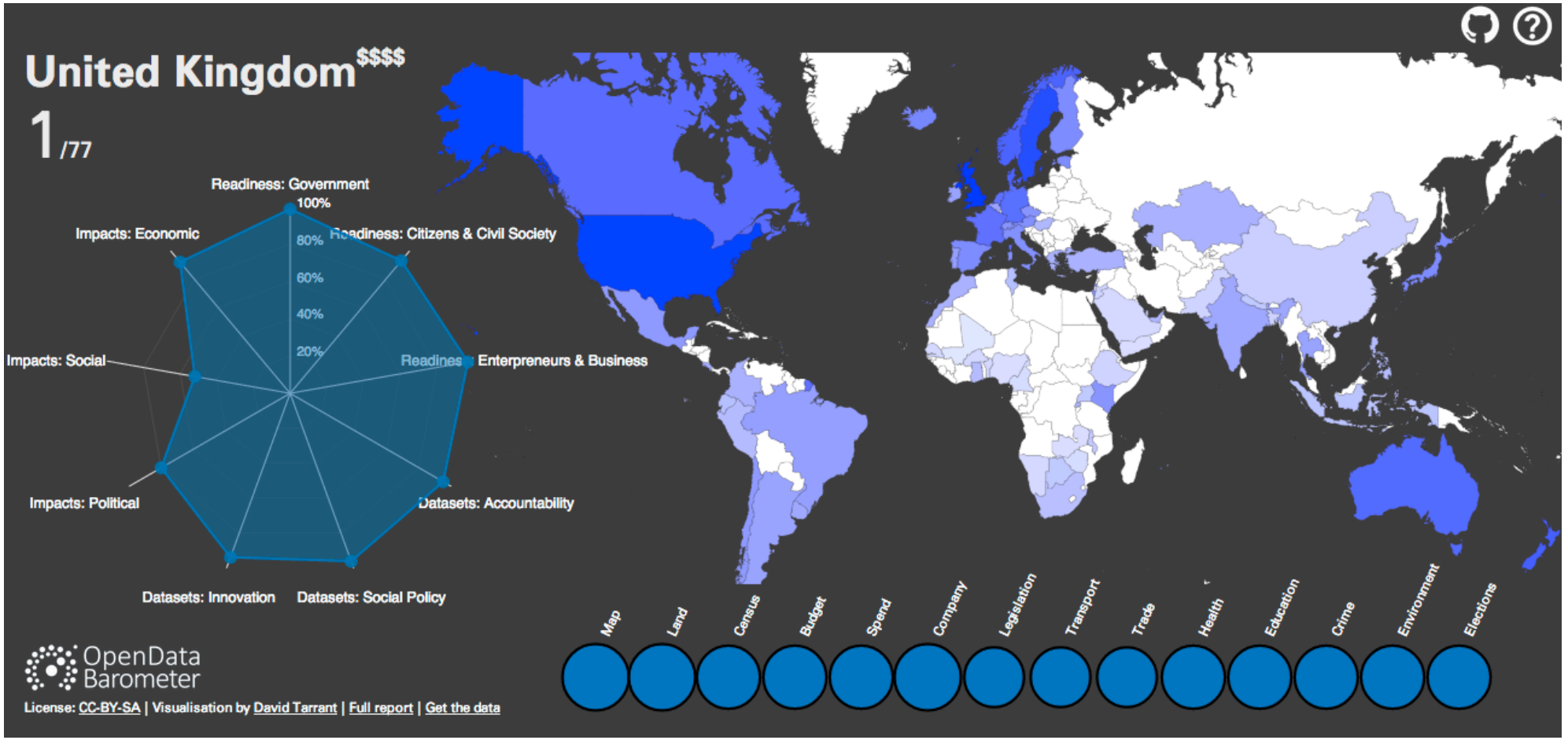
IN the web



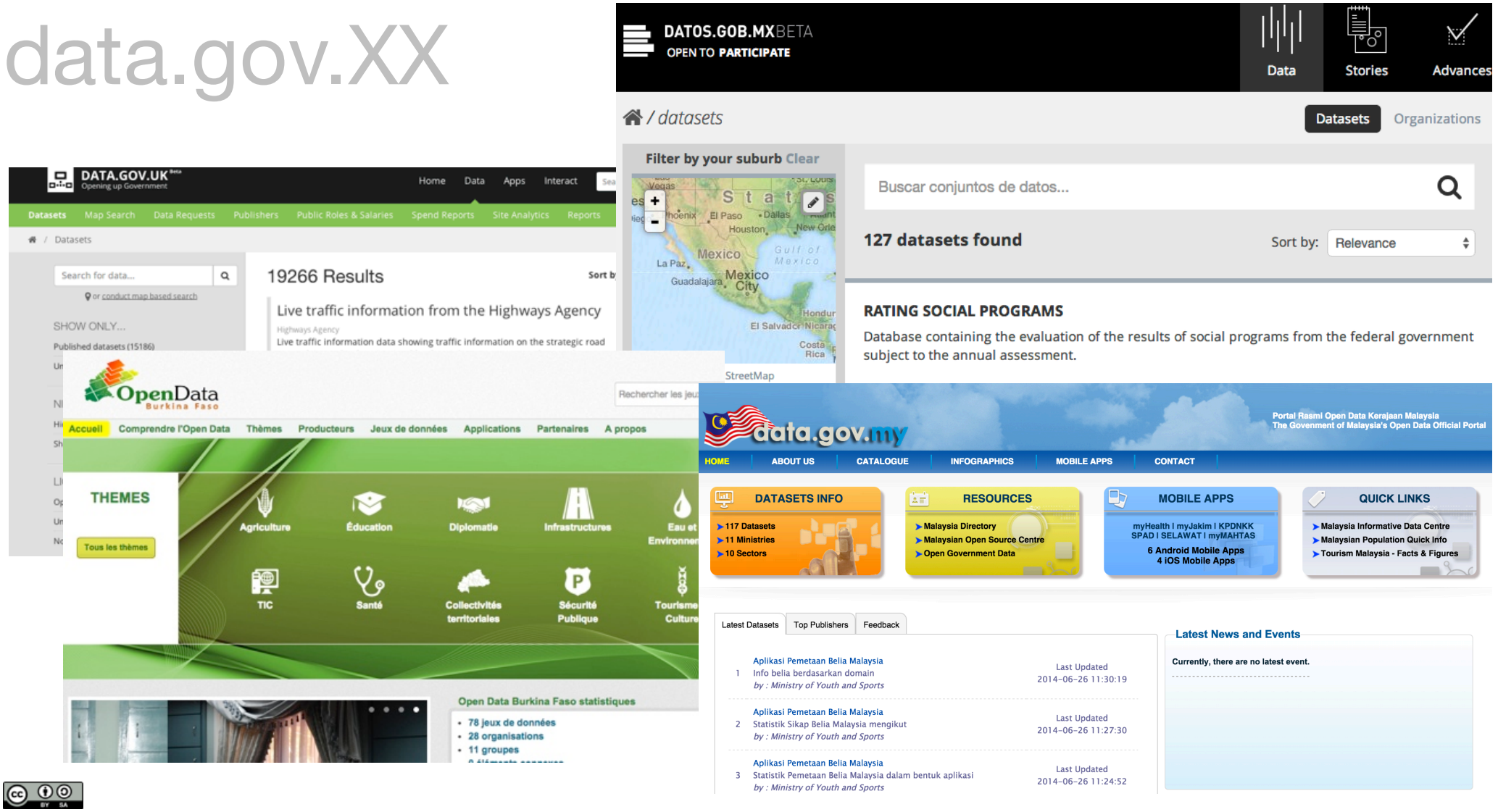
Finding data on the web **(of documents)**

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

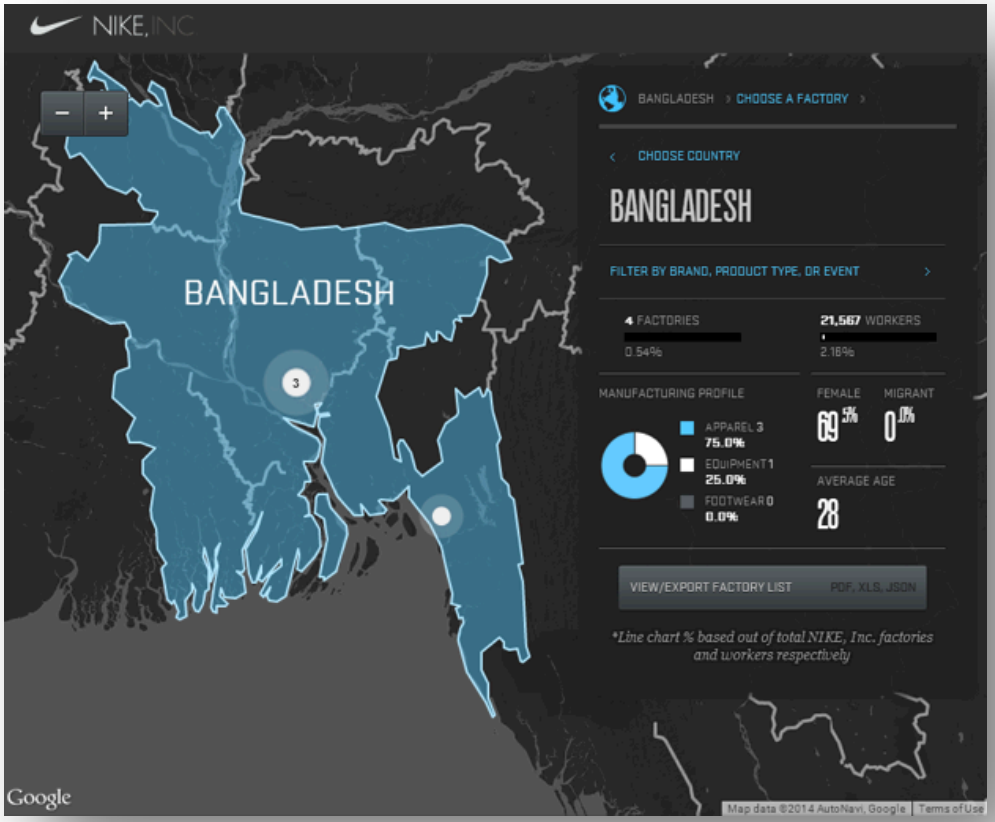
Government data



data.gov.XX

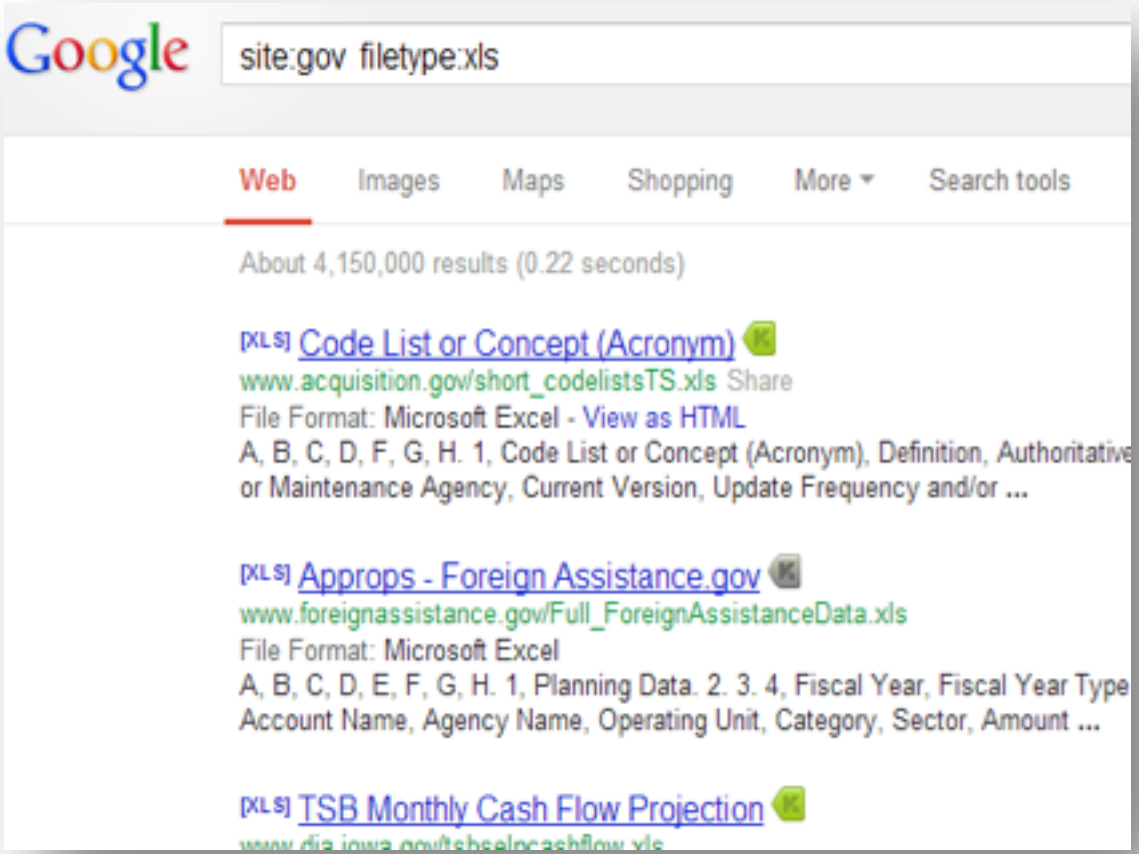

DATOS.GOB.MXBETA
OPEN TO PARTICIPATE
Data
Stories
Advances
/ datasets
Datasets
Organizations
Filter by your suburb Clear
Buscar conjuntos de datos...
127 datasets found
Sort by: Relevance
RATING SOCIAL PROGRAMS
Database containing the evaluation of the results of social programs from the federal government subject to the annual assessment.
DATA.GOV.UK
Opening up Government
Home
Data
Apps
Interact
Datasets
Map Search
Data Requests
Publishers
Public Roles & Salaries
Spend Reports
Site Analytics
Reports
19266 Results
Live traffic information from the Highways Agency
Live traffic information data showing traffic information on the strategic road
SHOW ONLY...
Published datasets (15186)
OpenData
Burkina Faso
Accueil
Comprendre l'Open Data
Thèmes
Producteurs
Jeux de données
Applications
Partenaires
A propos
THEMES
Tous les thèmes
Agriculture
Éducation
Diplomatie
Infrastructures
TIC
Santé
Collectivités territoriales
Sécurité Publique
Open Data Burkina Faso statistiques
78 jeux de données
28 organisations
11 groupes
data.gov.my
Portal Rasmi Open Data Kerajaan Malaysia
The Govenment of Malaysia's Open Data Official Portal
HOME
ABOUT US
CATALOGUE
INFOGRAPHICS
MOBILE APPS
CONTACT
DATASETS INFO
117 Datasets
11 Ministries
10 Sectors
RESOURCES
Malaysia Directory
Malaysian Open Source Centre
Open Government Data
MOBILE APPS
myHealth I myJakim I KPDNKK
SPAD I SELAWAT I myMAHTAS
6 Android Mobile Apps
4 iOS Mobile Apps
QUICK LINKS
Malaysia Informative Data Centre
Malaysian Population Quick Info
Tourism Malaysia - Facts & Figures
Latest Datasets
Top Publishers
Feedback
1 Aplikasi Pemetaan Belia Malaysia
Info belia berdasarkan domain
by : Ministry of Youth and Sports
Last Updated 2014-06-26 11:30:19
2 Aplikasi Pemetaan Belia Malaysia
Statistik Sikap Belia Malaysia mengikut
by : Ministry of Youth and Sports
Last Updated 2014-06-26 11:27:30
3 Aplikasi Pemetaan Belia Malaysia
Statistik Pemetaan Belia Malaysia dalam bentuk aplikasi
by : Ministry of Youth and Sports
Last Updated 2014-06-26 11:24:52
Latest News and Events
Currently, there are no latest event.

Suppliers



http://manufacturingmap.nikeinc.com/#

Google advanced



site: Get results only from certain sites or domains

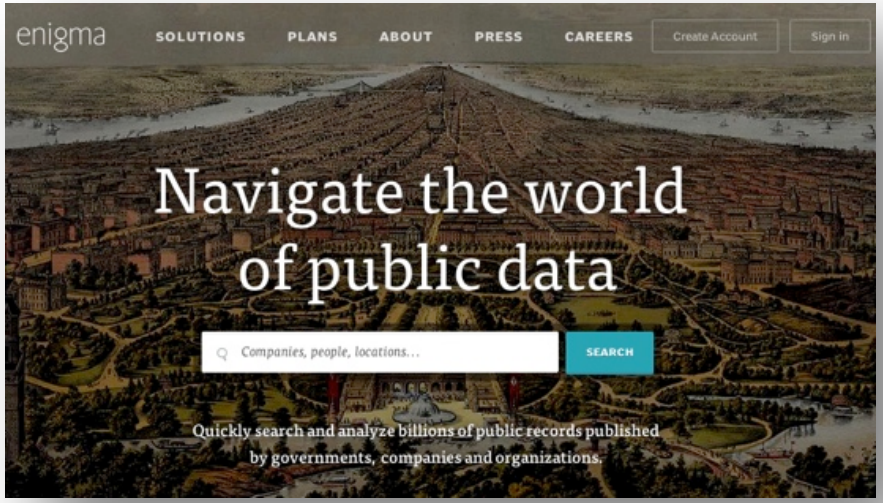
link: Find pages that link to a certain page

related: Find sites similar to one you already know

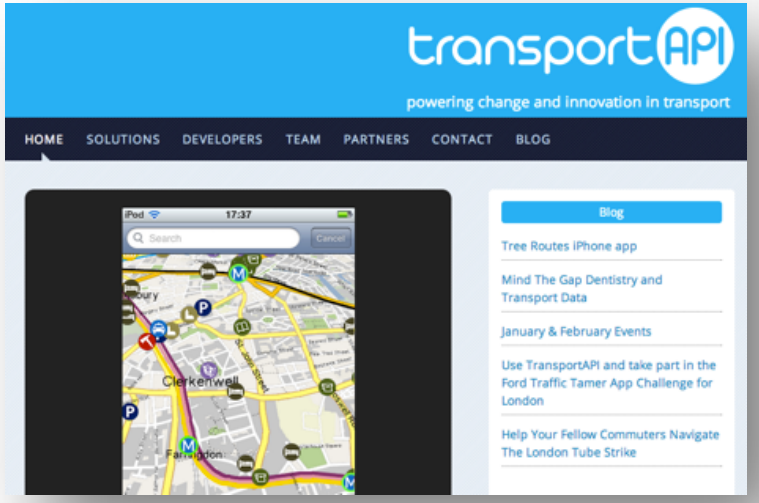
filetype: Find certain file types only

Aggregators and portals

Collect together data from across the web into one place.



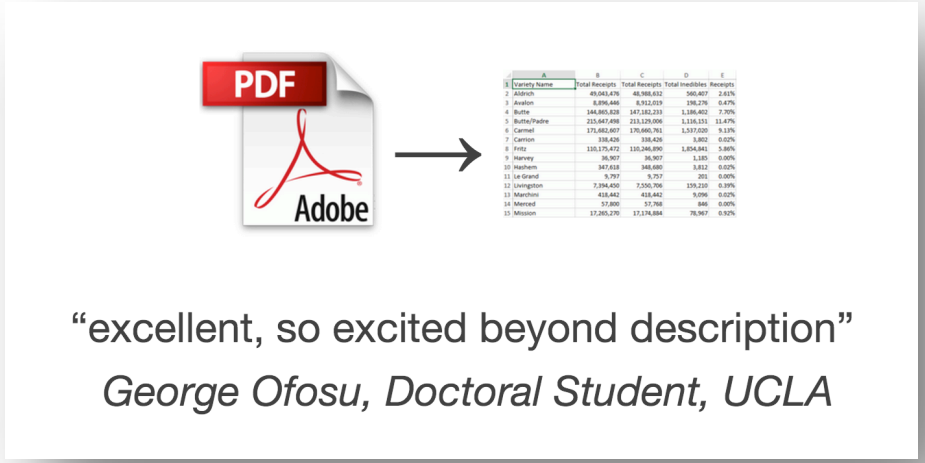
enigma.io



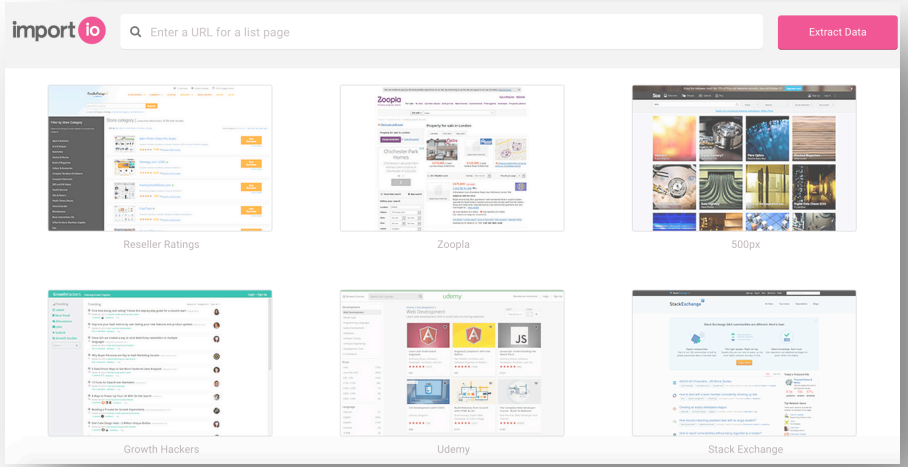
transportAPI

Scraping

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



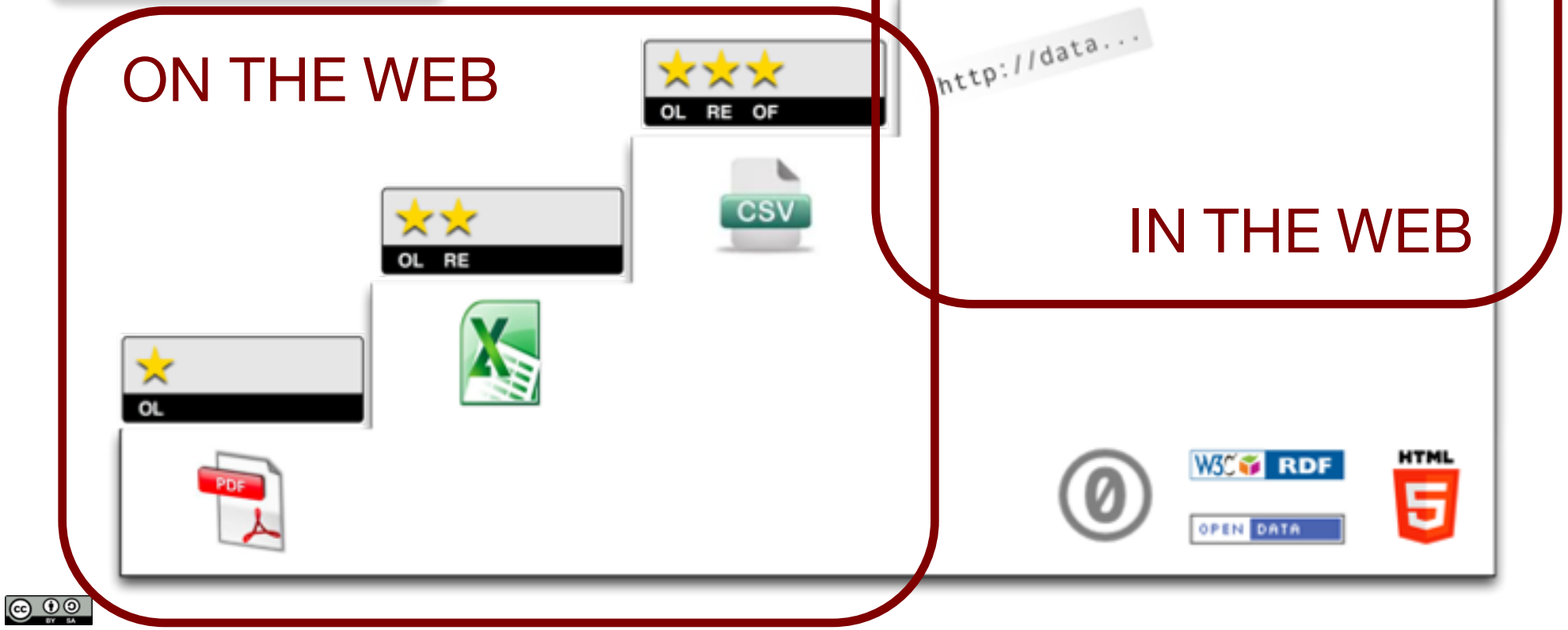
pdftables.com



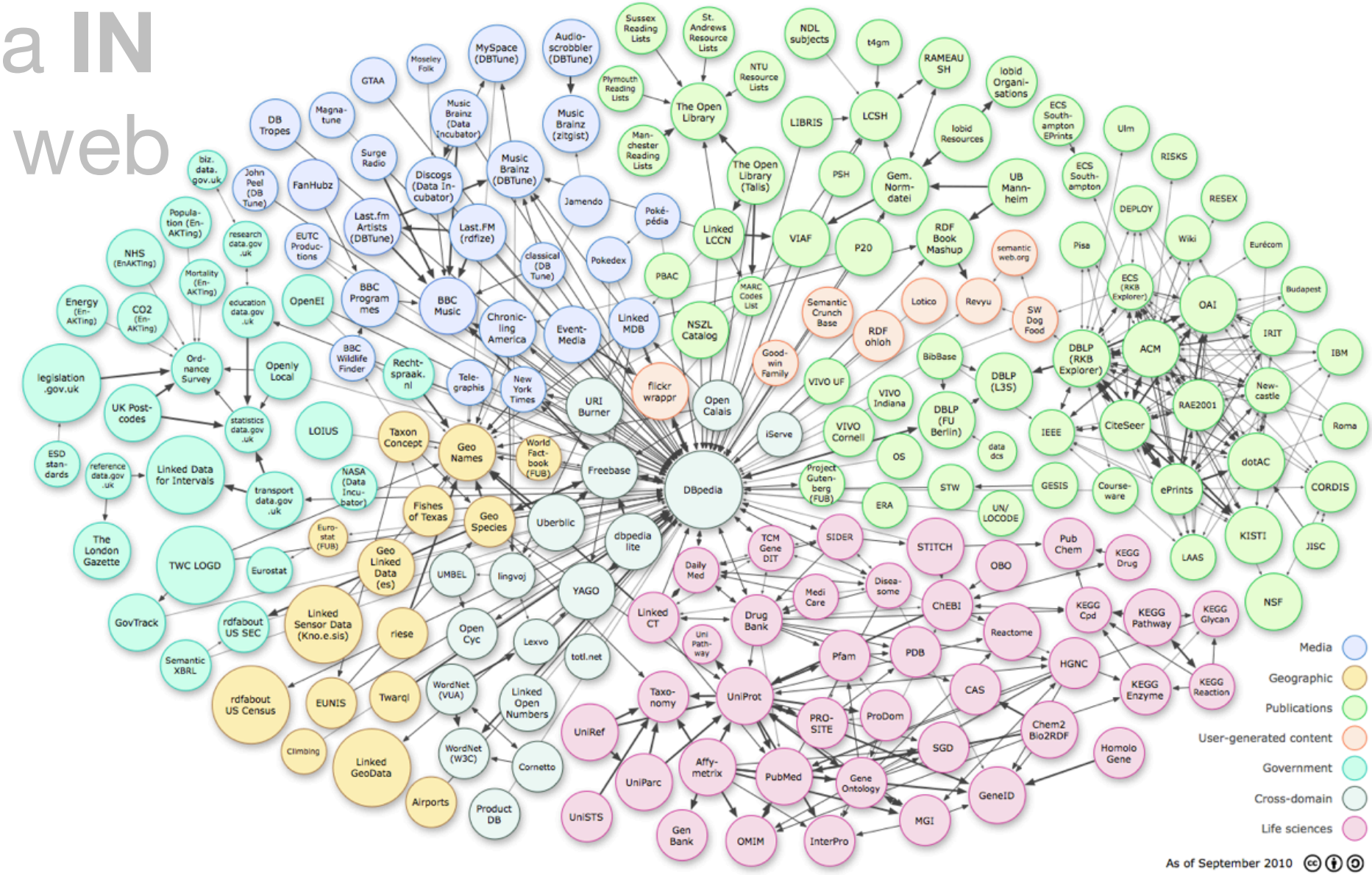
magic.import.io

5-Stars

http://5stardata.info/



Data **IN** the web



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)

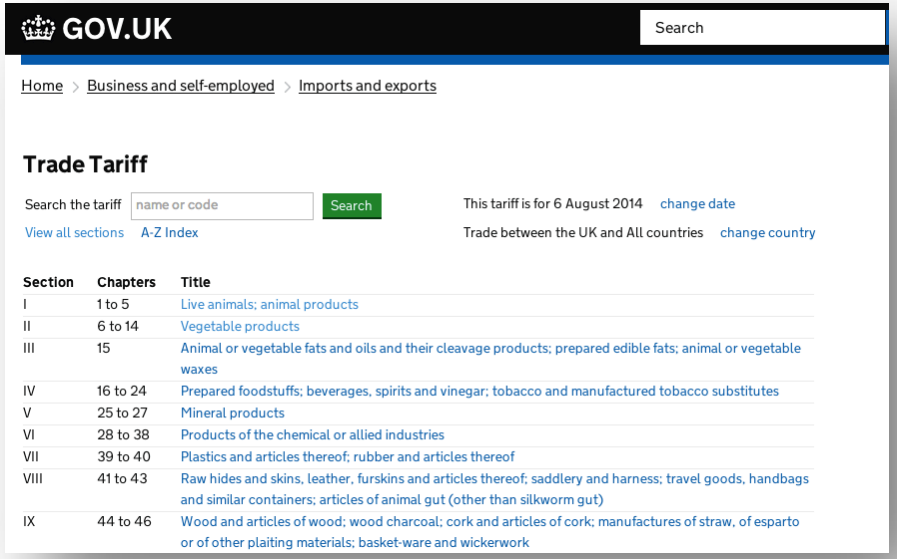
How the web should work, but people forgot that Tim put this in when he invented it!

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

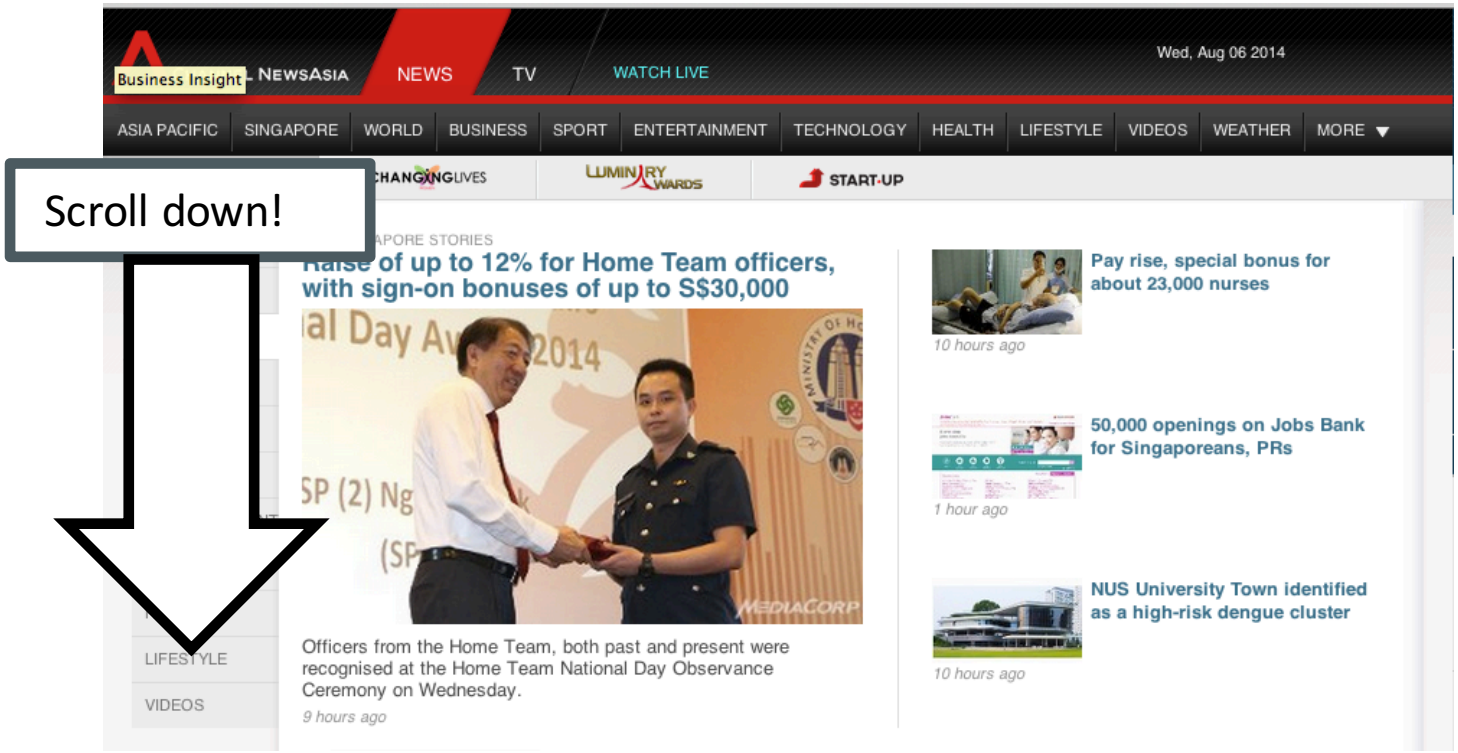


UK Trade Tariff

BBC Music and Programmes

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

2. Look for alternative links



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

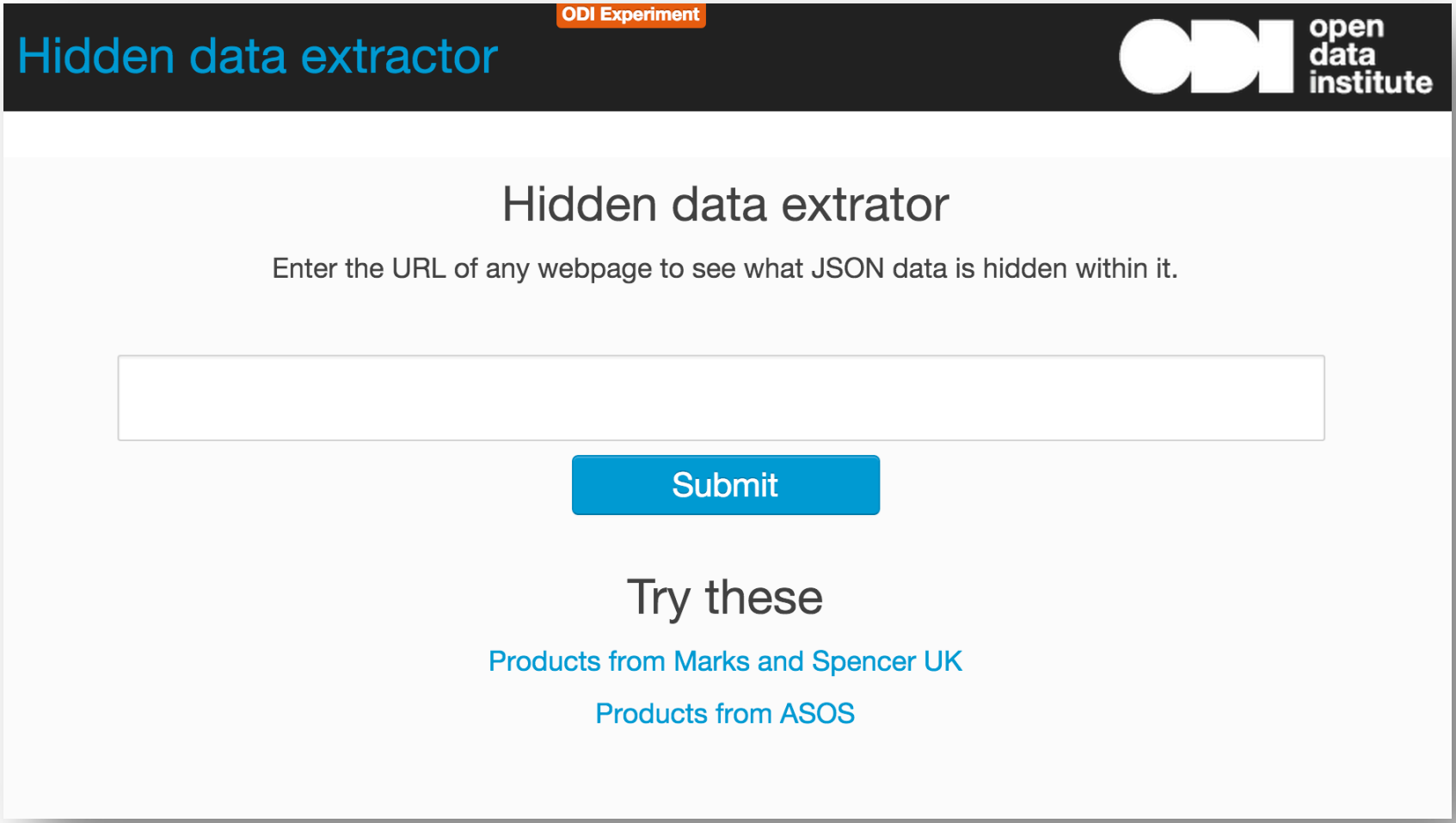
ADVERTISE WITH US
Online Advertising
Mobile Advertising
TV Advertising
Contact Sales

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



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

You've gone incognito. Pages you view in incognito tabs won't stick around in your browser's history, cookie store, or search history after you've closed **all** of your incognito tabs. Any files you download or bookmarks you create will be kept.

However, you aren't invisible. Going incognito doesn't hide your browsing from your employer, your internet service provider, or the websites you visit.

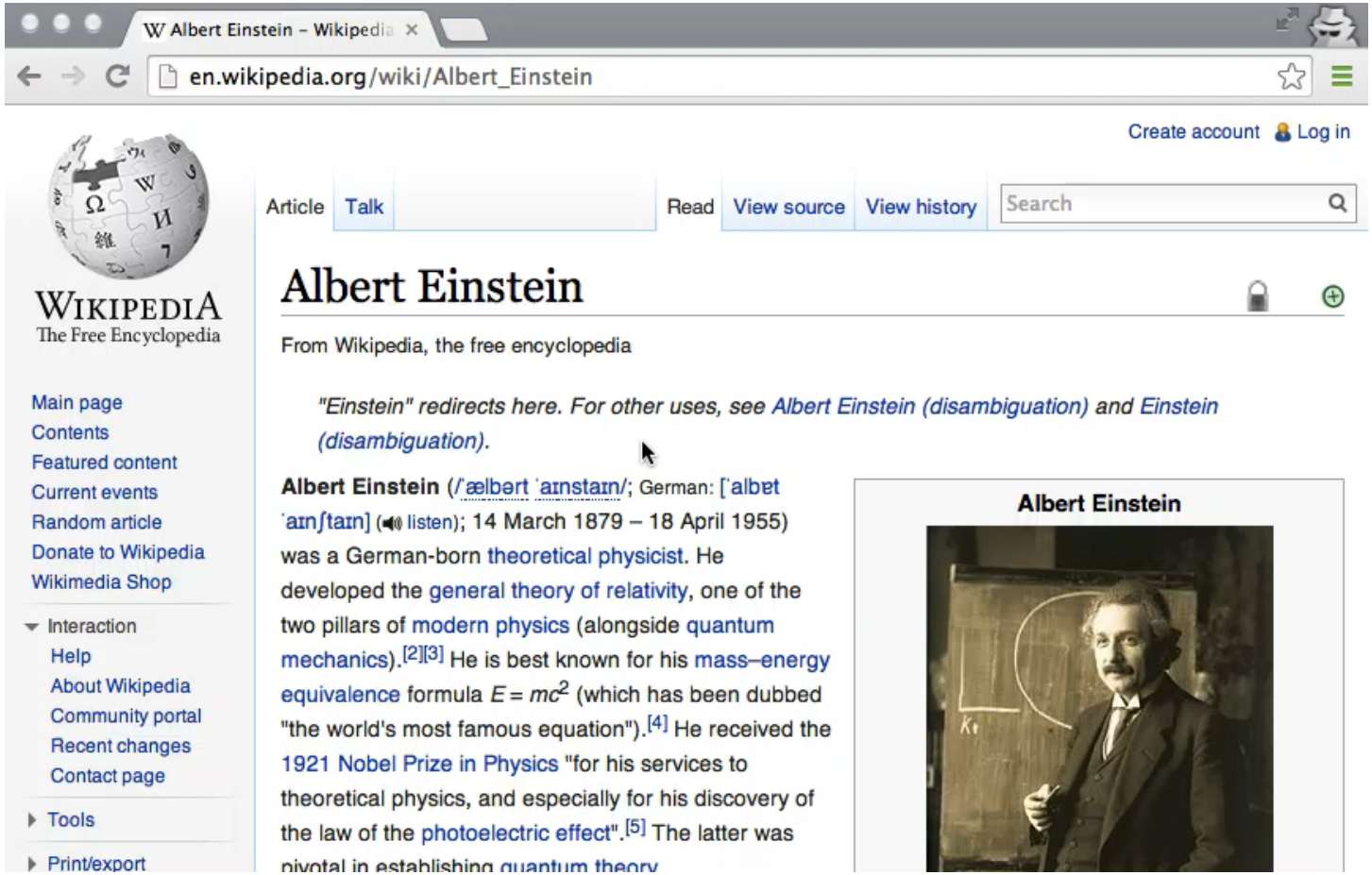
Learn more about incognito browsing.

Because Google Chrome does not control how extensions handle your personal data, all extensions have been disabled for incognito windows. You can reenable them individually in the extensions manager.

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?

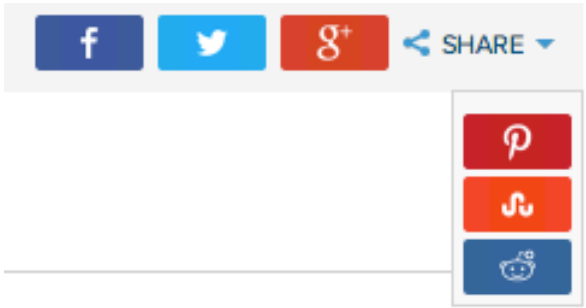

Albert Einstein - Wikipedia
en.wikipedia.org/wiki/Albert_Einstein
Create account
Log in
Article
Talk
Read
View source
View history
Search
WIKIPEDIA
The Free Encyclopedia
Albert Einstein
From Wikipedia, the free encyclopedia
"Einstein" redirects here. For other uses, see Albert Einstein (disambiguation) and Einstein (disambiguation).
Albert Einstein (/ˈælbərt ˈaɪnstaɪn/; German: [ˈalbɐt 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
Albert Einstein
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
Print/export

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

Common links

RSS Feed 

Social channels



Languages

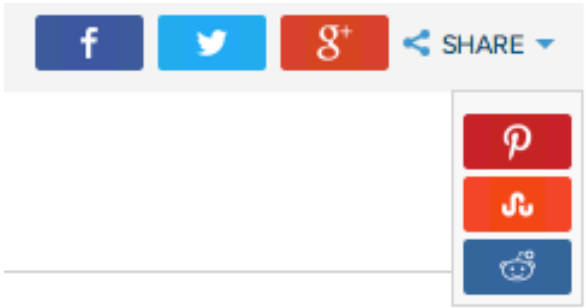
Printable version 

Print version

Common links

RSS Feed 

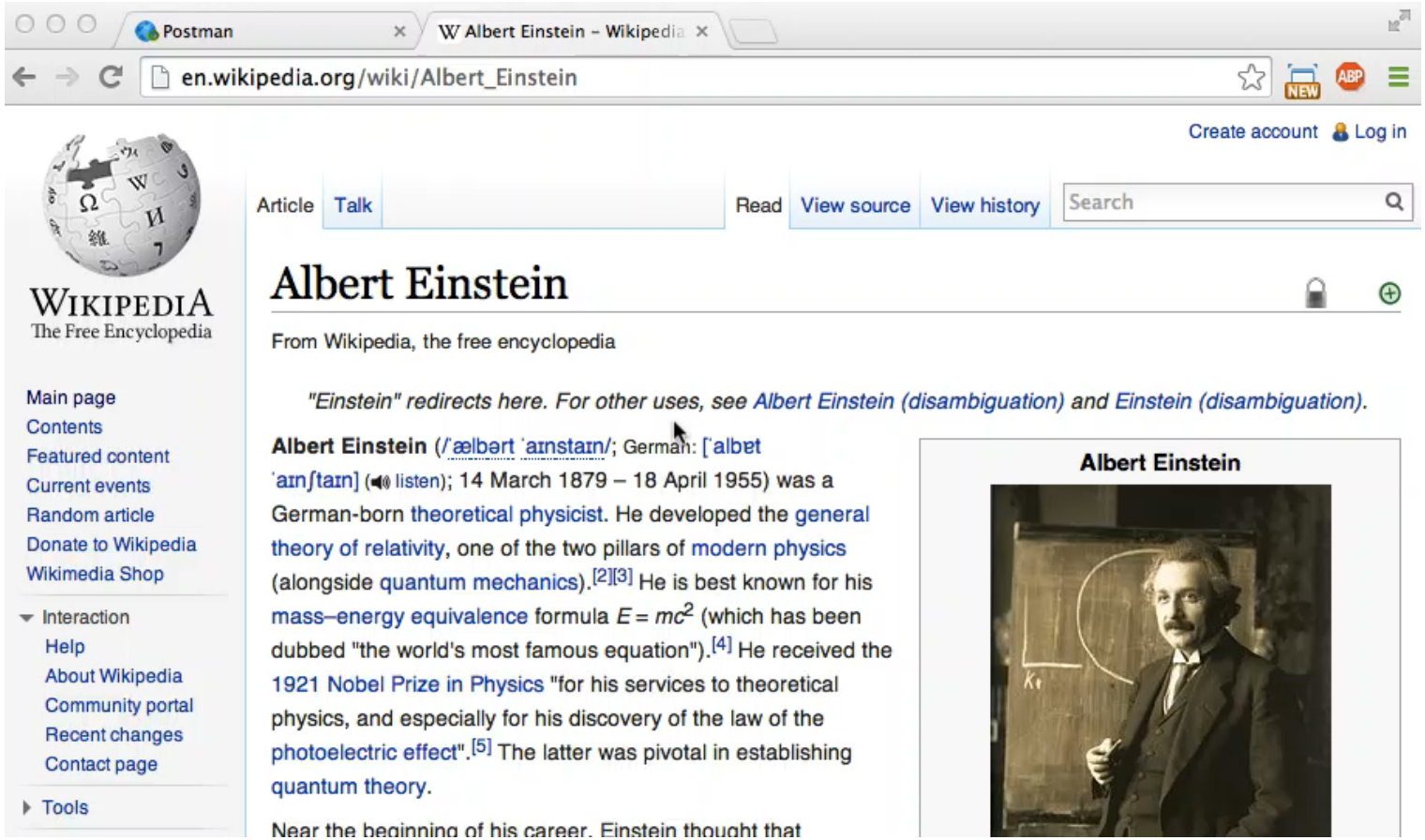
Social channels



Languages

Printable version

How a machine does it



A Web Request

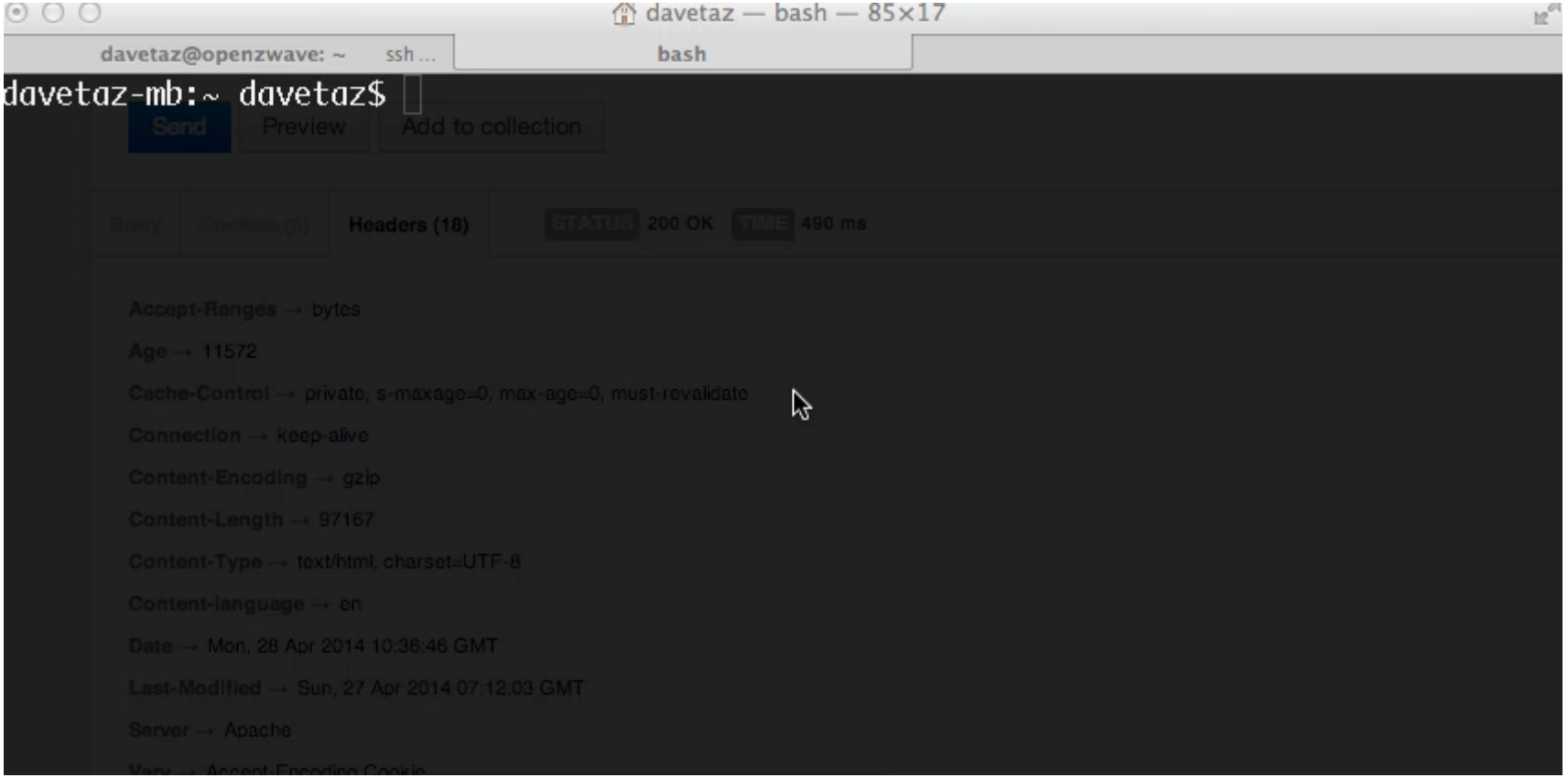
http://en.wikipedia.org/wiki/Albert_Einstein

- **GET** / HTTP/1.1

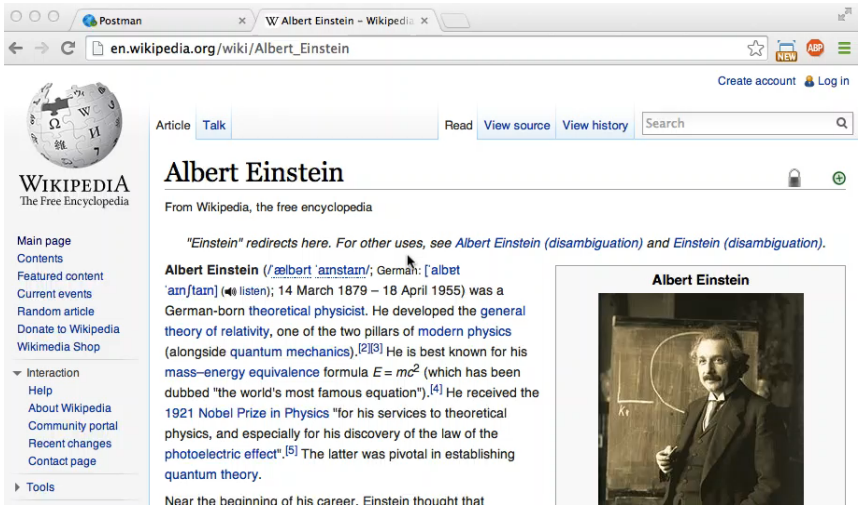
```
telnet en.wikipedia.org 80

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

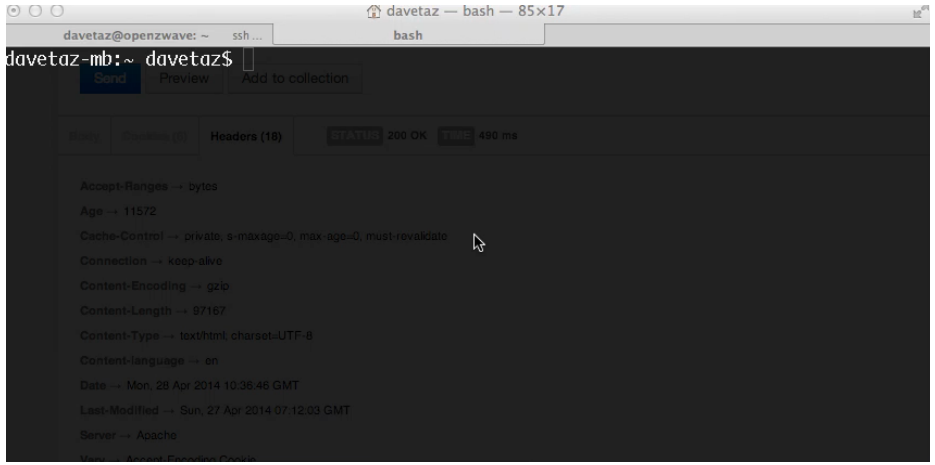
A Web Request



The same

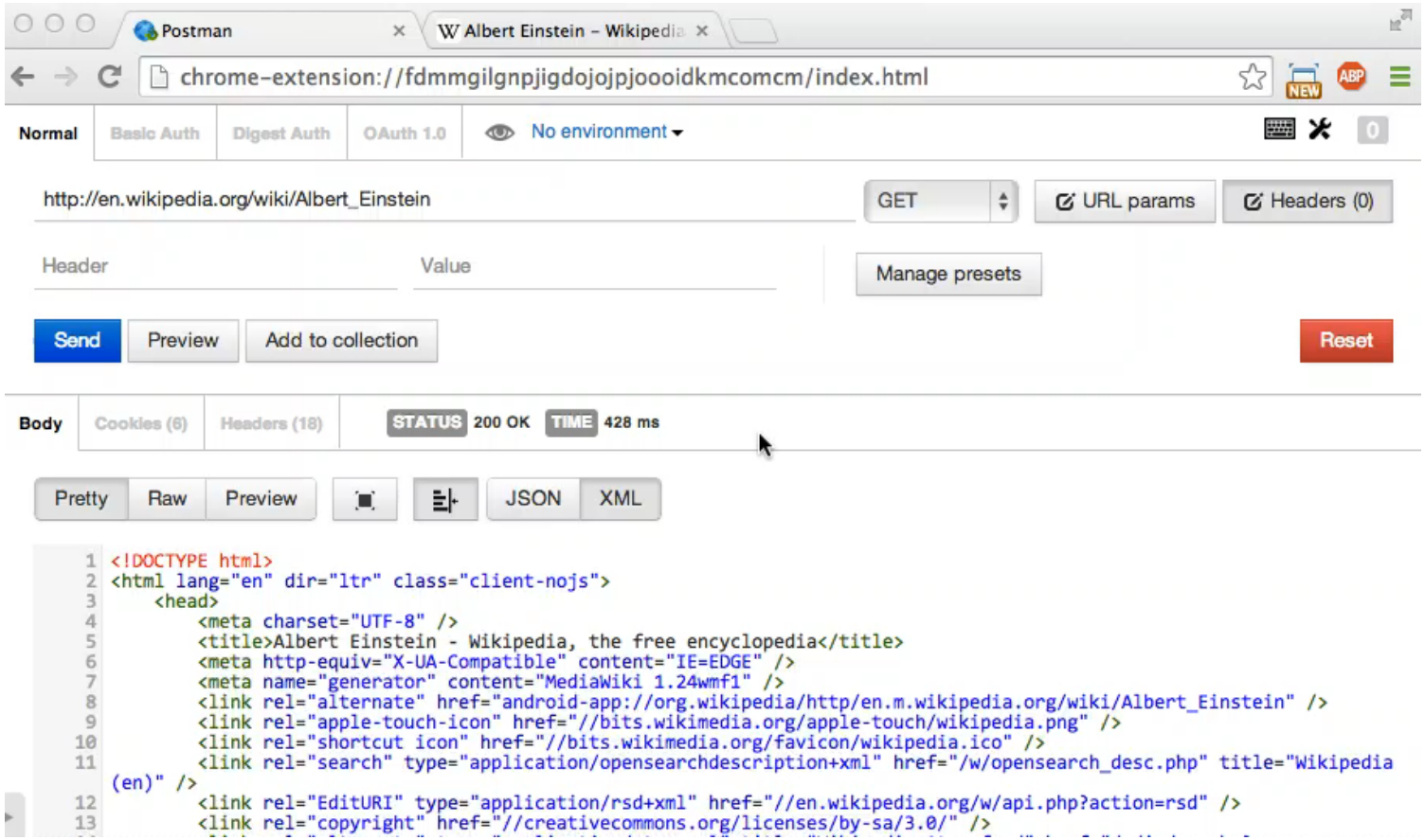


=



Headers

The headers define the properties of an HTTP transaction.



How a machine does it

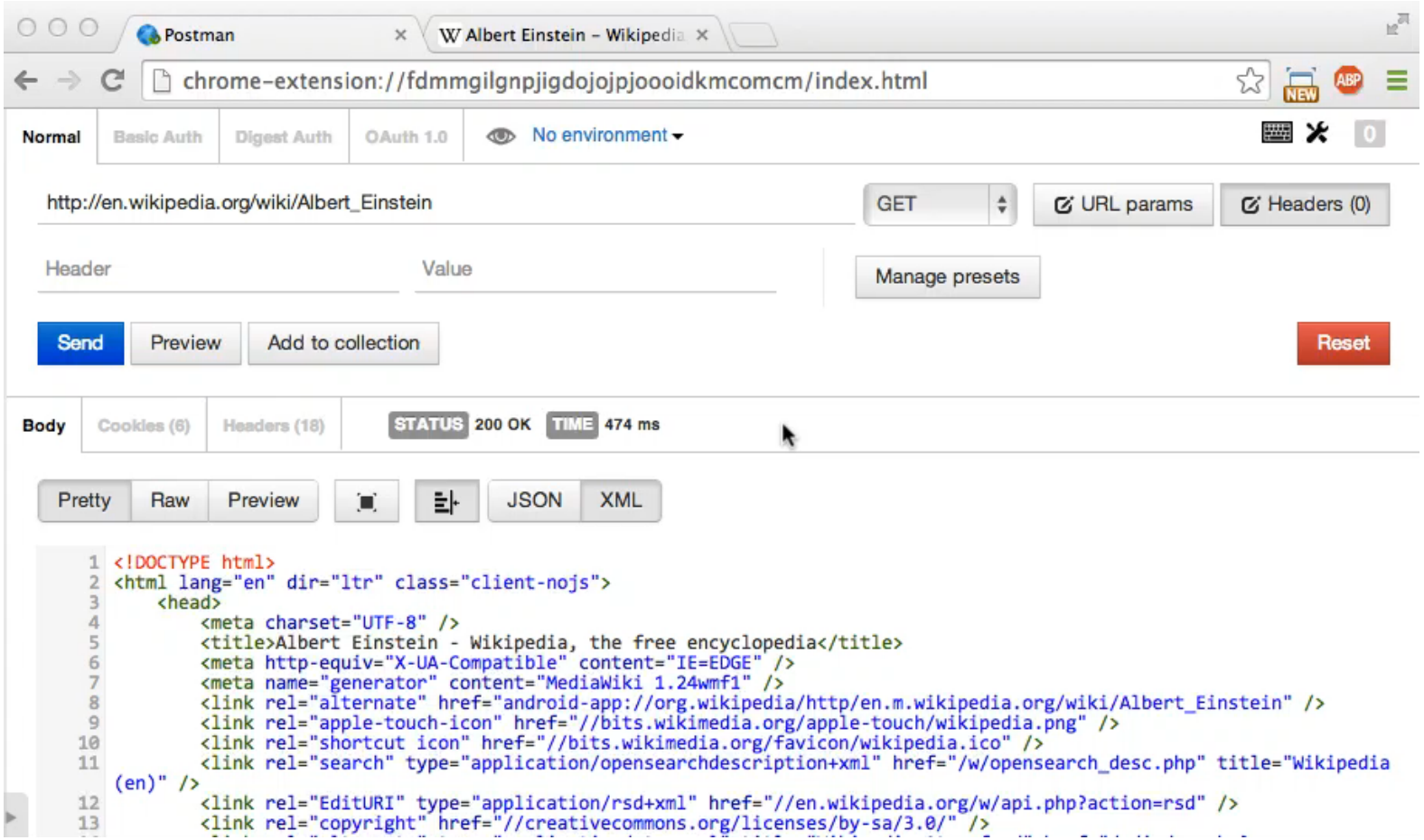
I have an address

I would like a different version:

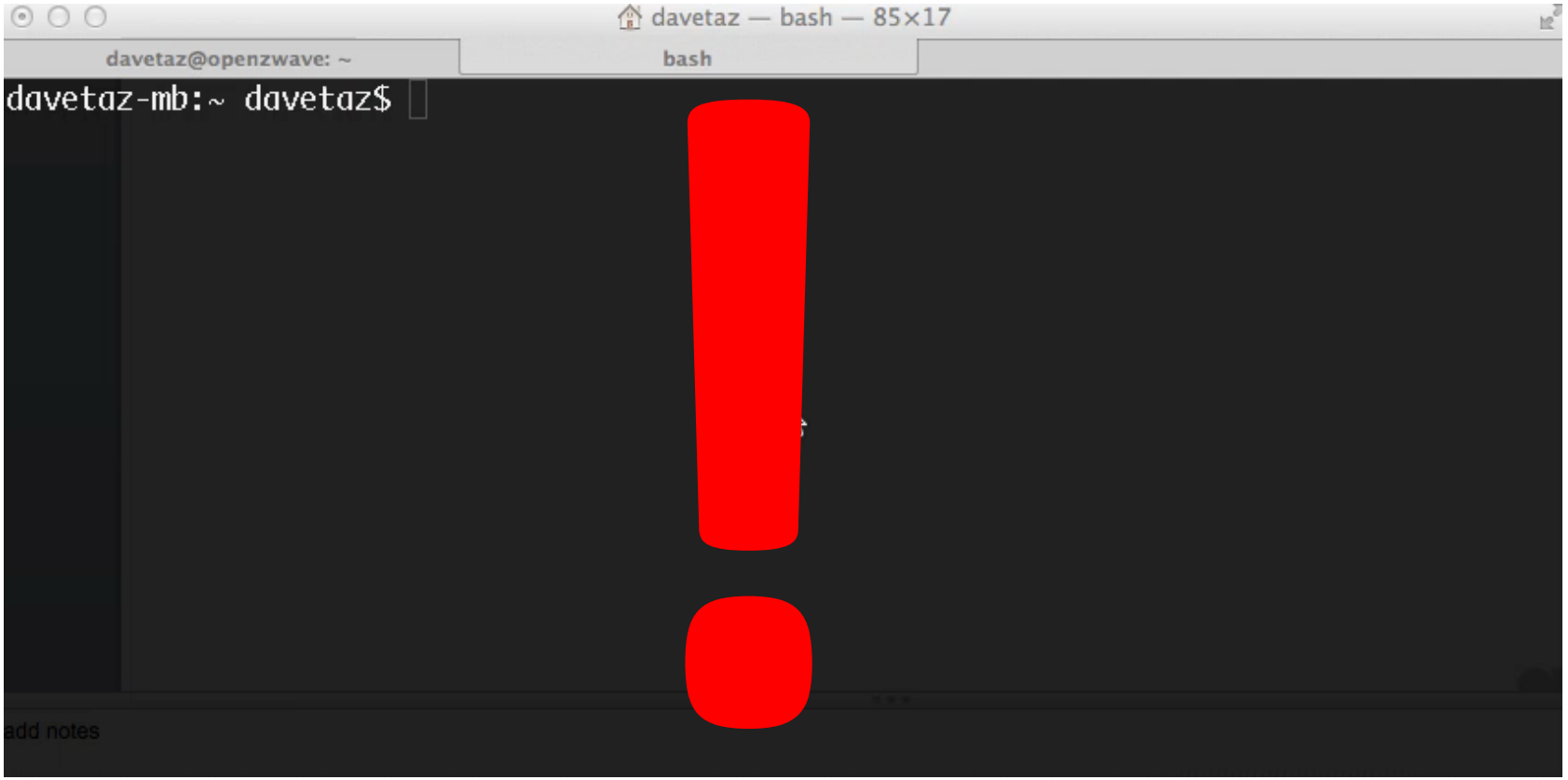
Accept:  application/rss+xml

Accept-Language:  es-es

Spanish please



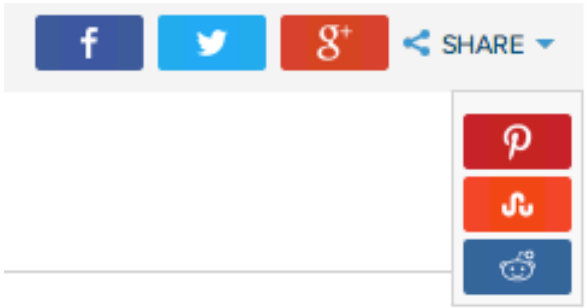
What happened?



Common links

RSS Feed 

Social channels



Languages

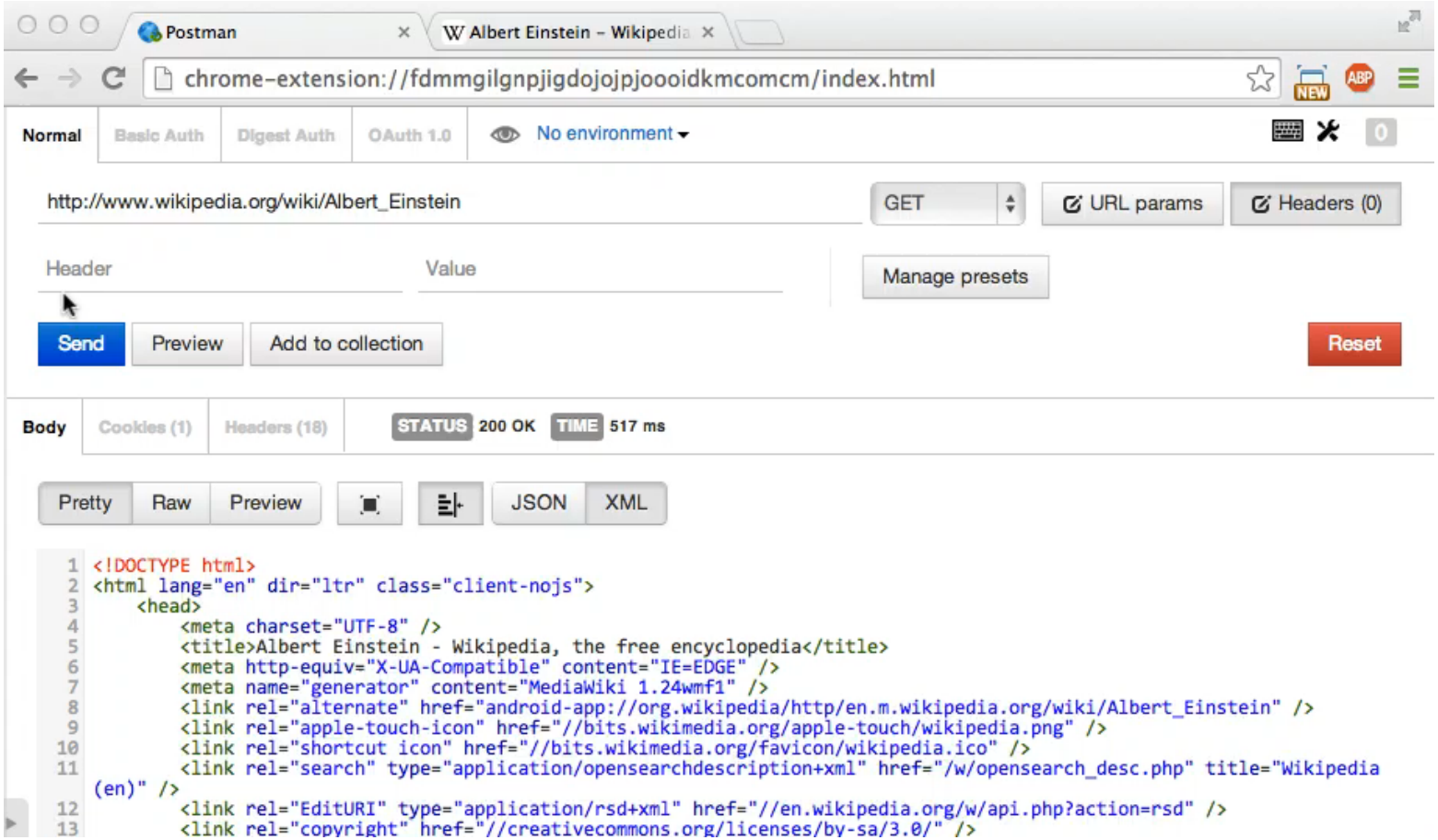
Printable version

Error Codes

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

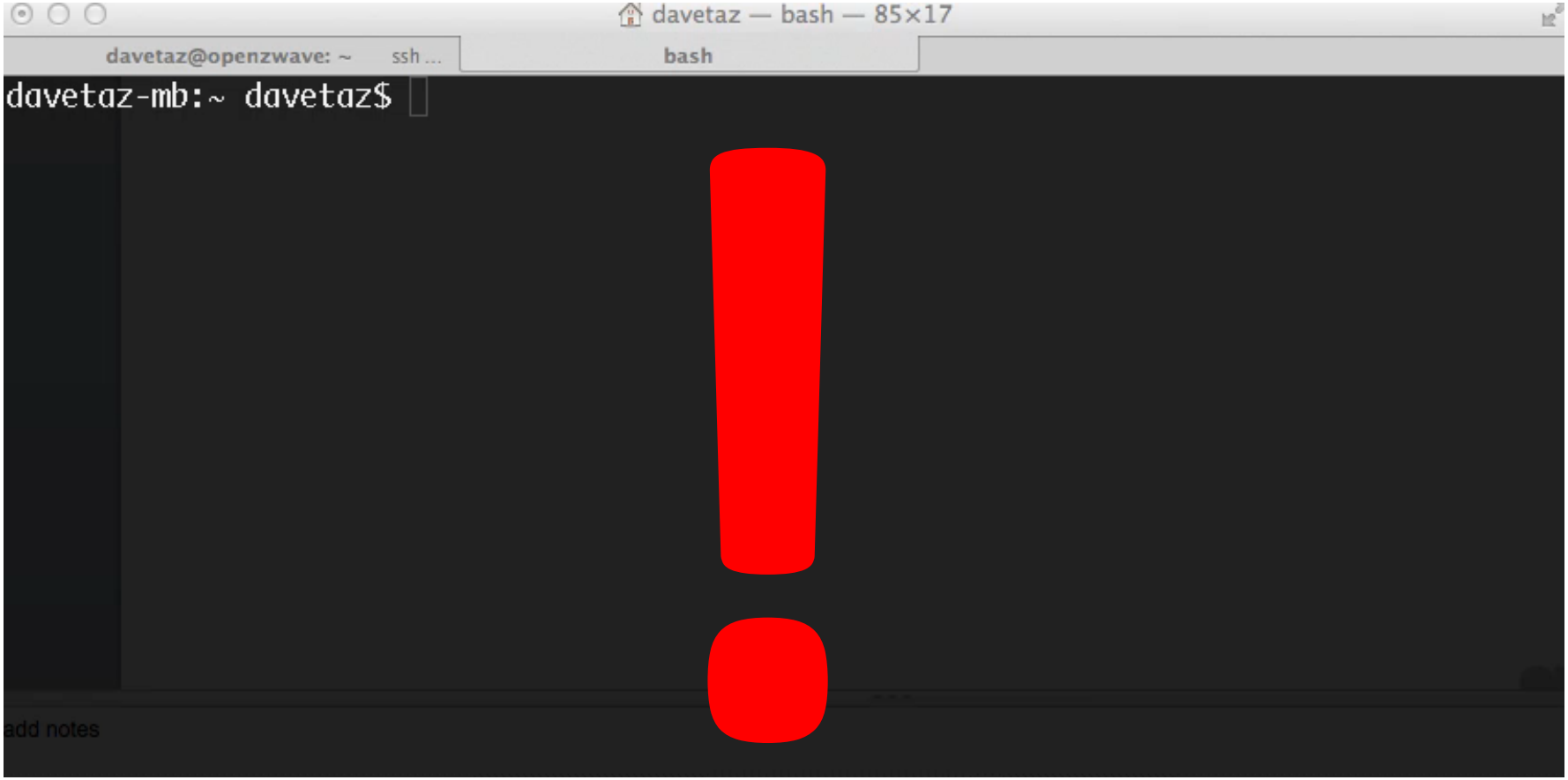
Try again

application/atom+xml



and again

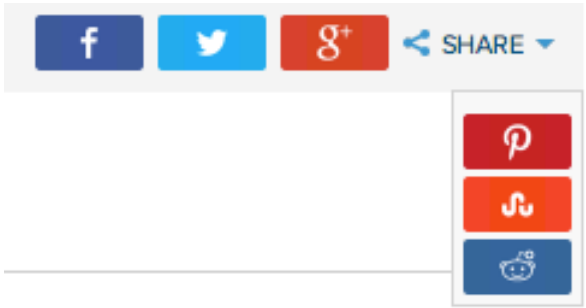
application/atom+xml

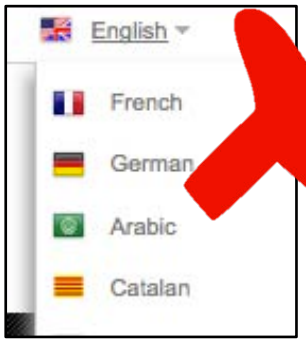


Common links

RSS Feed

Social channels

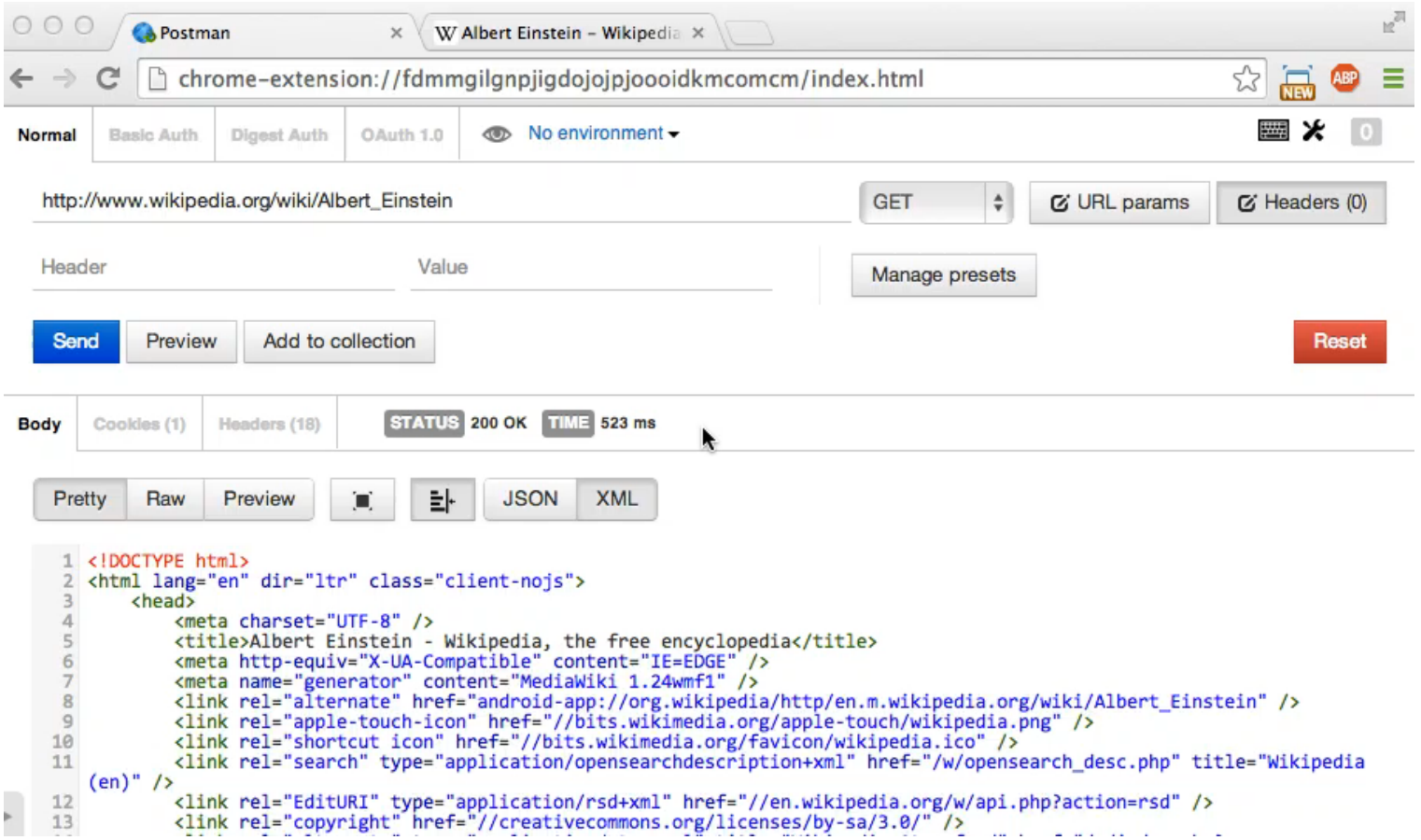




Languages

Printable version

So what do you have?

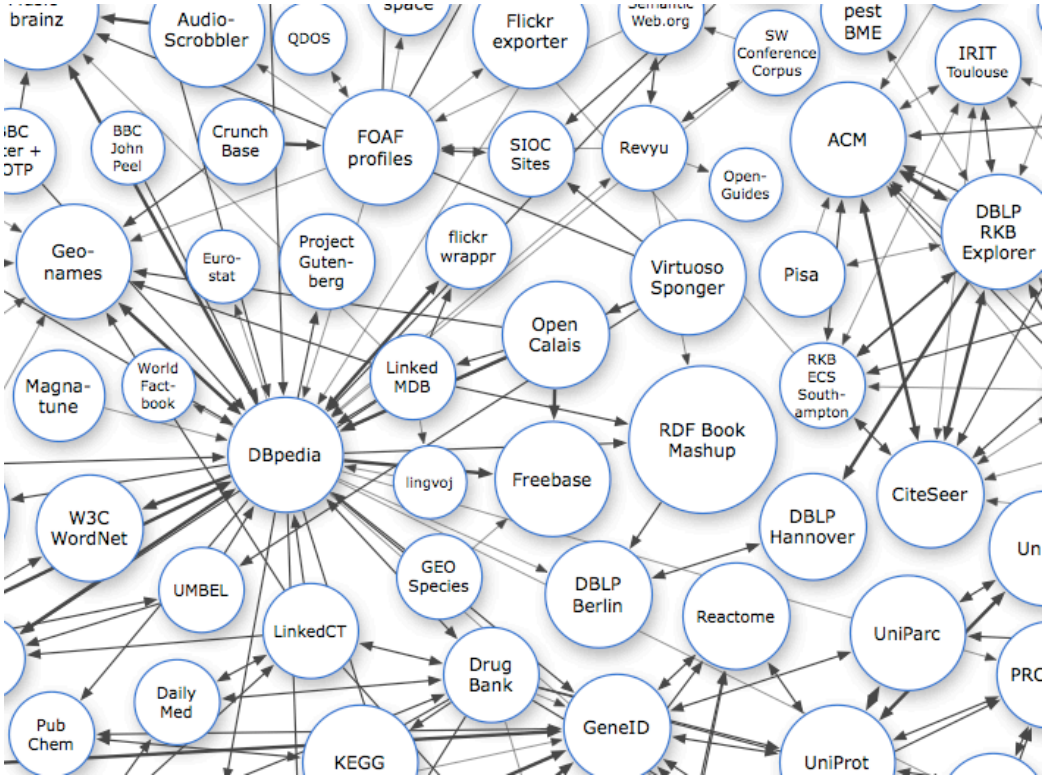


The <link rel=…> tag

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

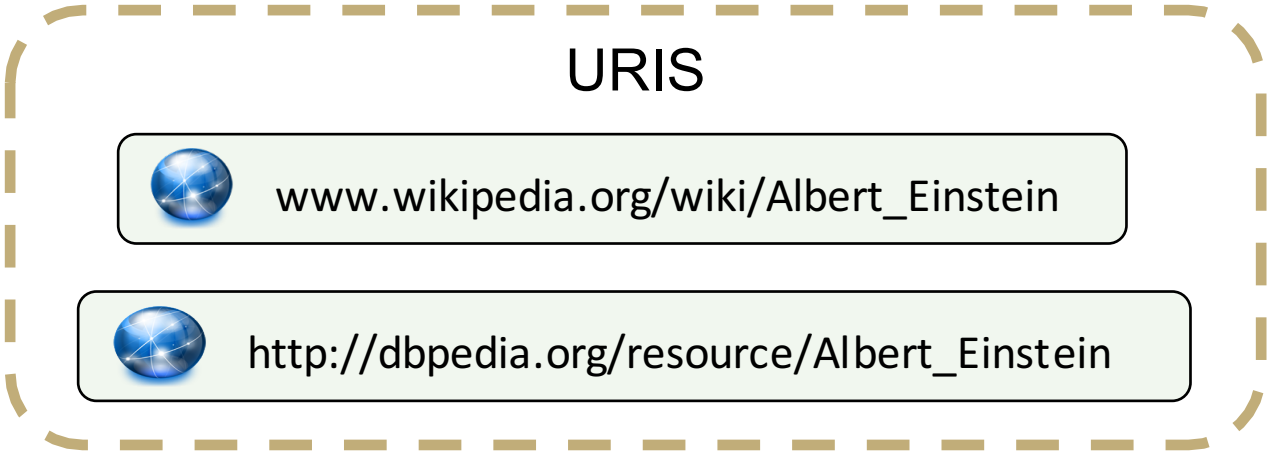
1. To stop the number of HTTP headers forever expanding?
2. To add custom headers? (Even though that is what X-* headers are)
3. Should machines have to search all the possible link-rel headers related to a resource and other links. (recurse, recurse, recurse)

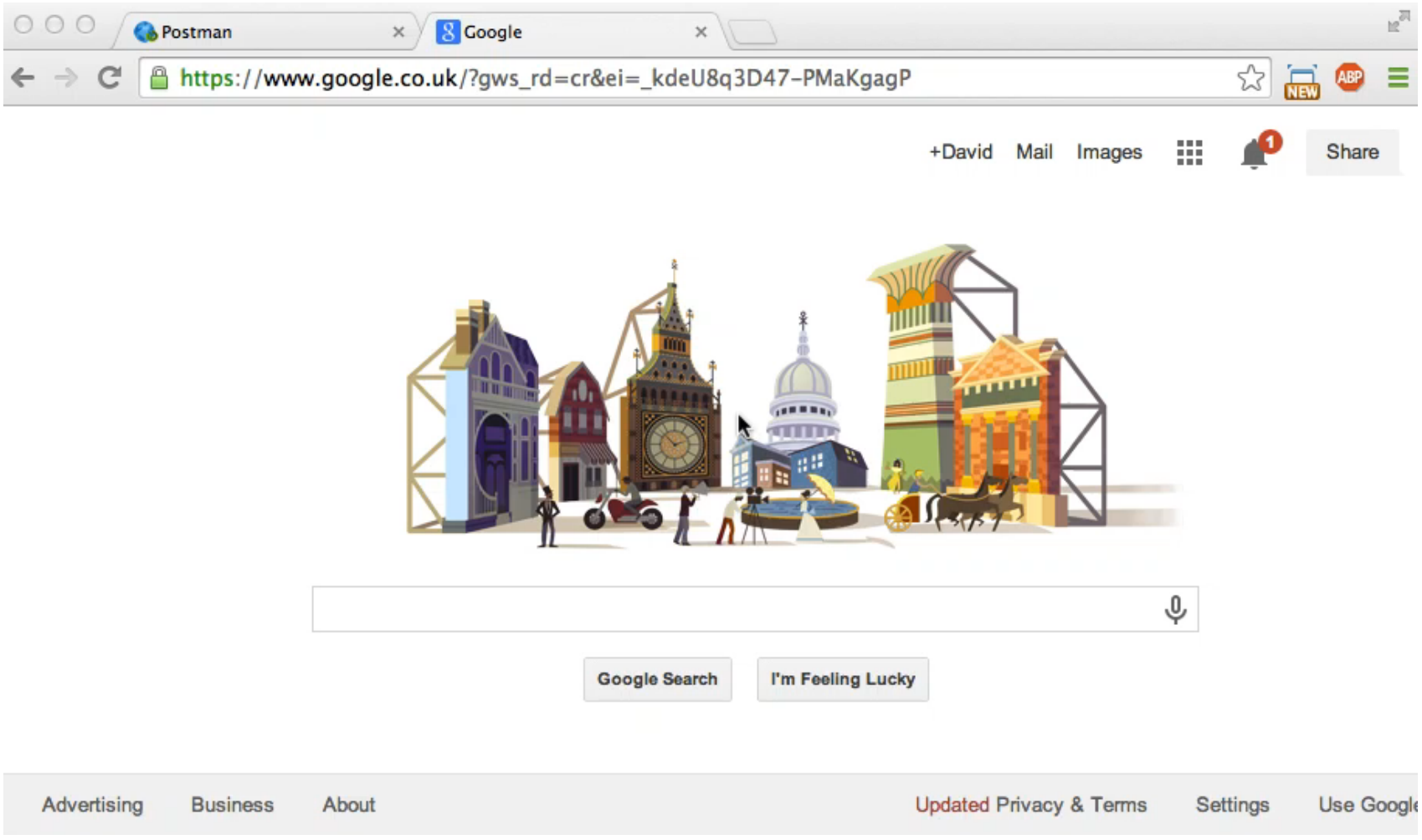
It’s not all doom and gloom

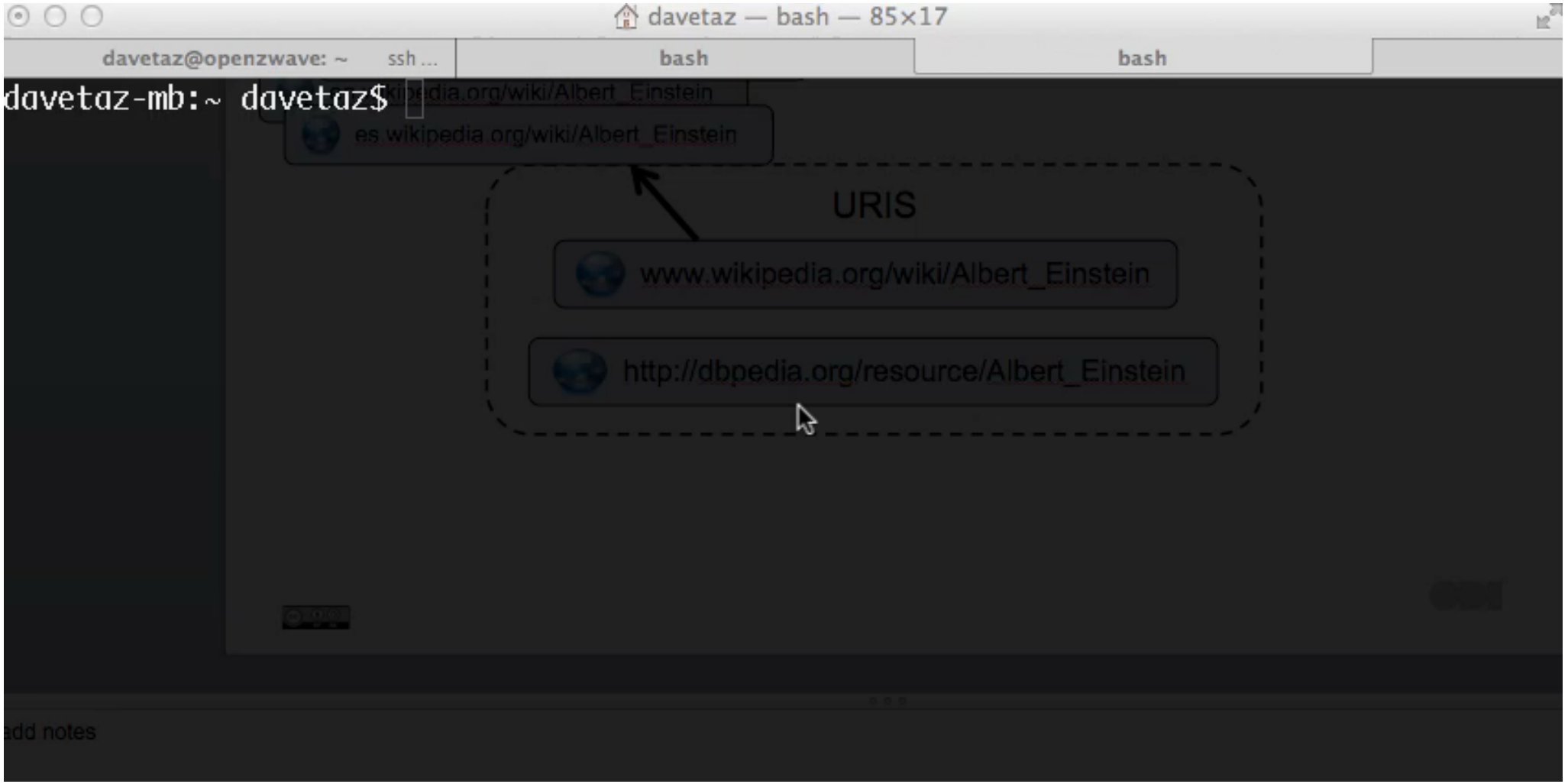


The data version of wikipedia, available in a machine friendly way.

Wikipedia & dbpedia




Postman
Google
https://www.google.co.uk/?gws_rd=cr&ei=_kdeU8q3D47-PMaKgagP
+David
Mail
Images
Share
Google Search
I'm Feeling Lucky
Advertising
Business
About
Updated Privacy & Terms
Settings
Use Google

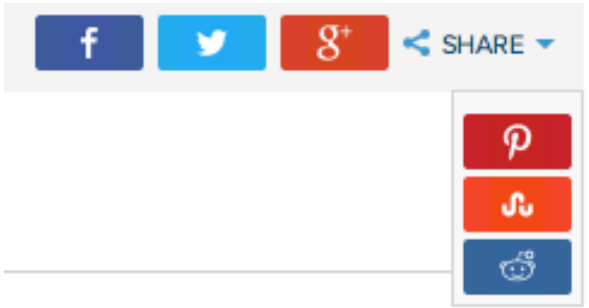

davetaz — bash — 85×17
davetaz@openzwave: ~
ssh ...
bash
bash
davetaz-mb:~ davetaz$
es wikipedia.org/wiki/Albert_Einstein
URIS
www.wikipedia.org/wiki/Albert_Einstein
http://dbpedia.org/resource/Albert_Einstein
edd notes

Common links

RSS Feed

Languages

Social channels



Printable version

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

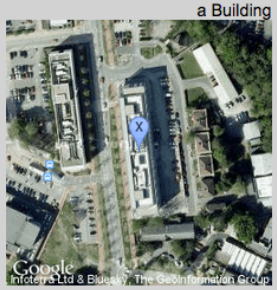
View Disability Report for this Building

©2010 Francois-Xavier Beckers (CC-BY)

Occupants

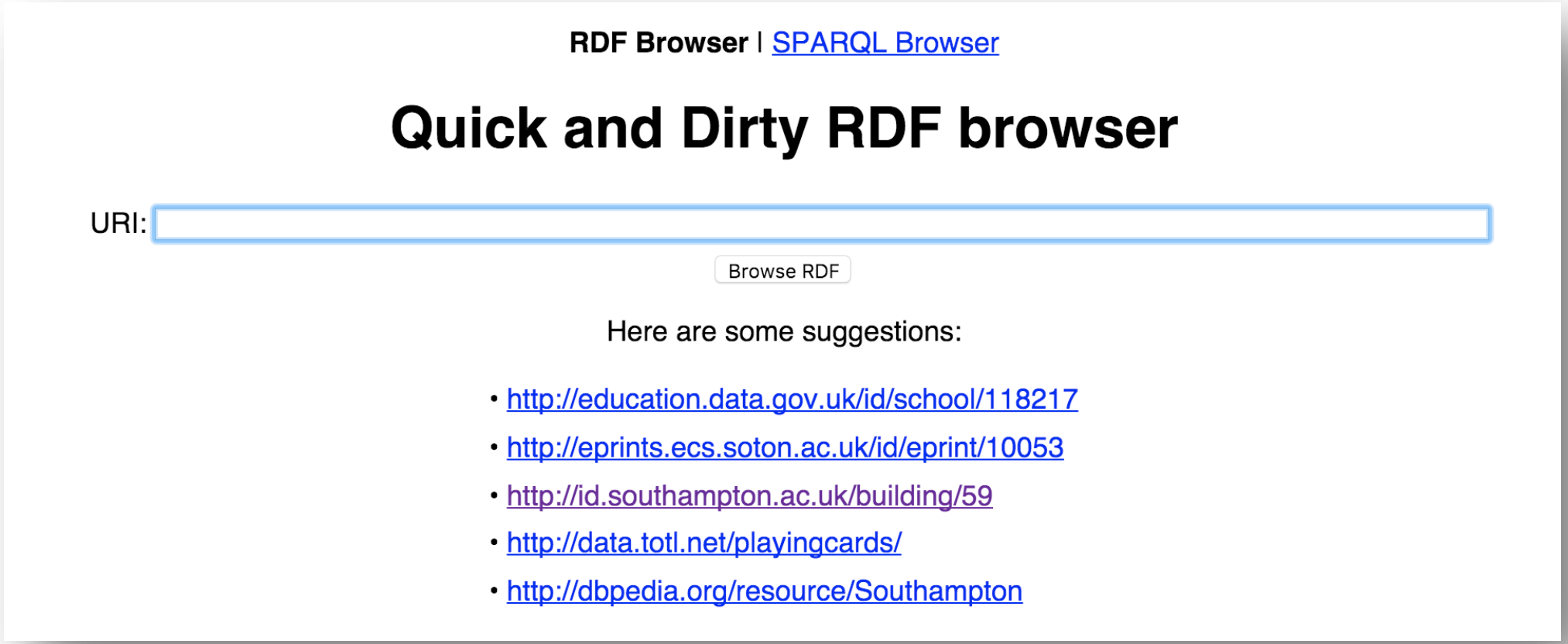
Electronics & Computer Science
Southampton Education School
Agents, Interactions & Complexity
Web & Internet Science
Leadership School Improve &Effectiveness
Lifelong & Work-Related Learning
Mathematics & Science Education
Social Justice & Inclusive Education
Teaching Only Staff
Deanery

a Building



ton.ac.uk/building/32
→ rooms:Building, http://id.southampton.ac.uk/ns/UoSBuilding
→ "EEE Building"
cupant → Electronics & Computer Science, Southampton Education School, Agents, Interactions
plexity, 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
outhampton.ac.uk/ns/disabledGoPage → http://www.disabledgo.com/en/access-guide/building-
ations:easting → "442544"^^xsd:integer
ations:northing → "115392"^^xsd:integer
Organization → University of Southampton
outhampton.ac.uk/ns/ombielName → "Bldg 32 (EEE)"
eature → Building 32 is non-residential
outhampton.ac.uk/ns/buildingDate → "2006"
outhampton.ac.uk/ns/buildingArchitect → John McAslan & Partners
patial → "POLYGON((-1.3961073411331264 50.93683868764933,-1.3958347895092957 50.9368567227702,-1.3956958407975968
606065737417,-1.3959558923017397 50.93603859197583,-1.3961073411331264 50.93683868764933))"
outhampton.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 ← RAeS Solent Branch Christmas Special Lecture - The Red Arrows

Browsing the web of data



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
 }
 moanDaintily();
 sleep();
}
```

```
makeOut(10, ["neck", "ear", "mouth"]);
```


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

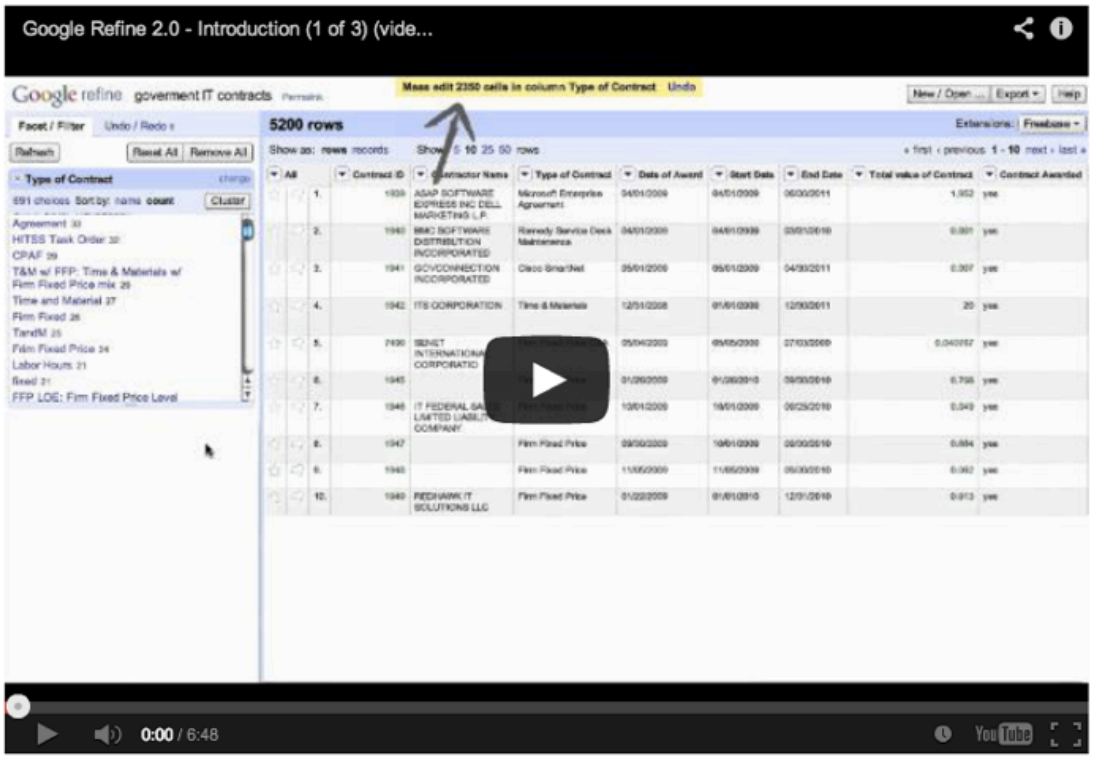
Outcomes

Clean and enrich data

Create a number of data processing pipelines

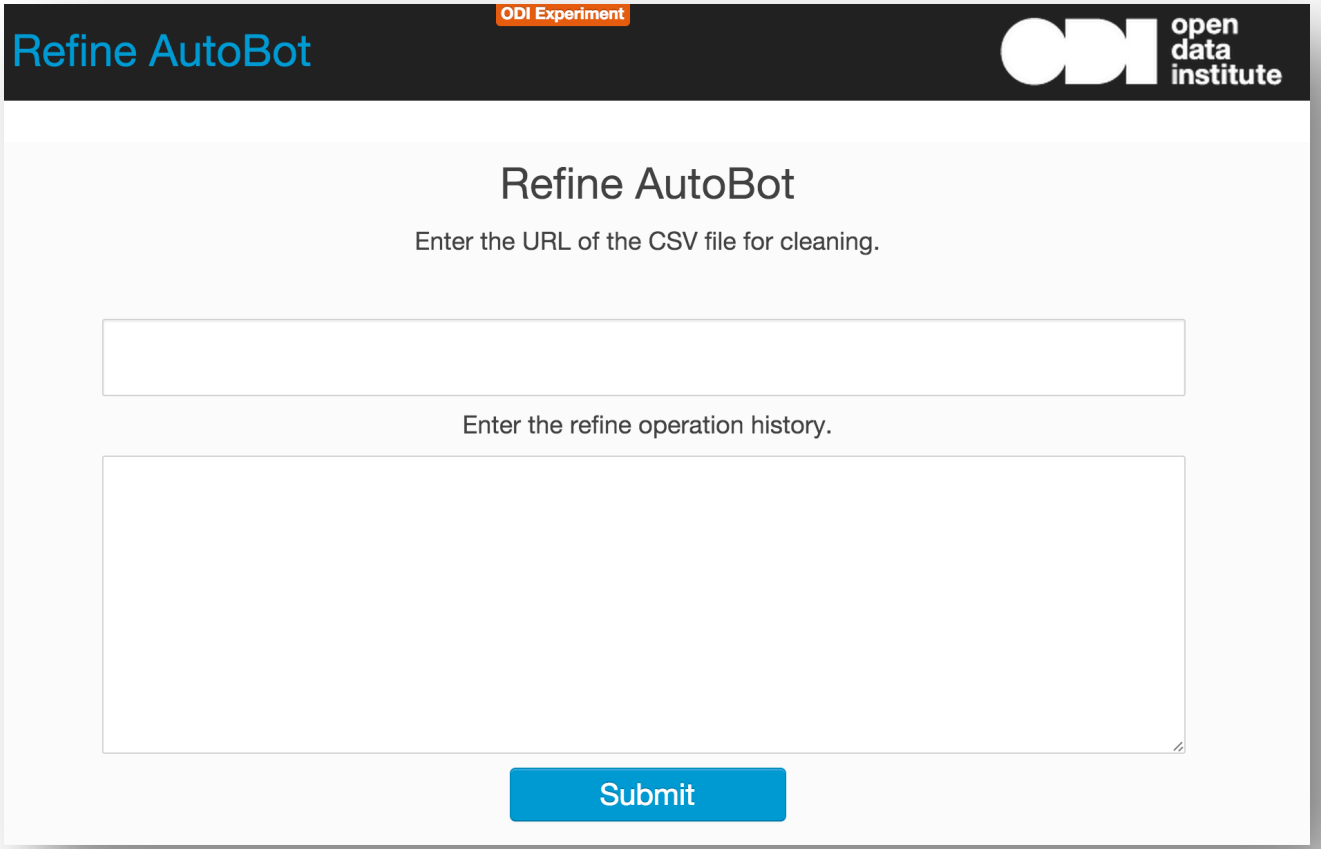
Translate data between a number of formats

Introducing Open Refine



http://openrefine.org

Data processing pipelines



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

opencorporates

The Open Database Of The Corporate World

Aggregator/Enabler

We have information on 70,597,888 companies

SEARCH

search companies search officers

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

Just released: OpenCorporates API v0.3

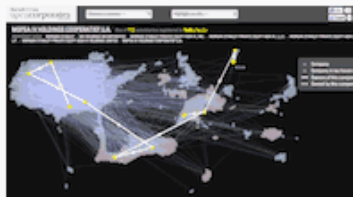
Corporate network data, financial accounts, complex filters, and more. Read more

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

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

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.

Human Readable Web

Question

What is significant about the first web page?

What actually did Tim invent?

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>

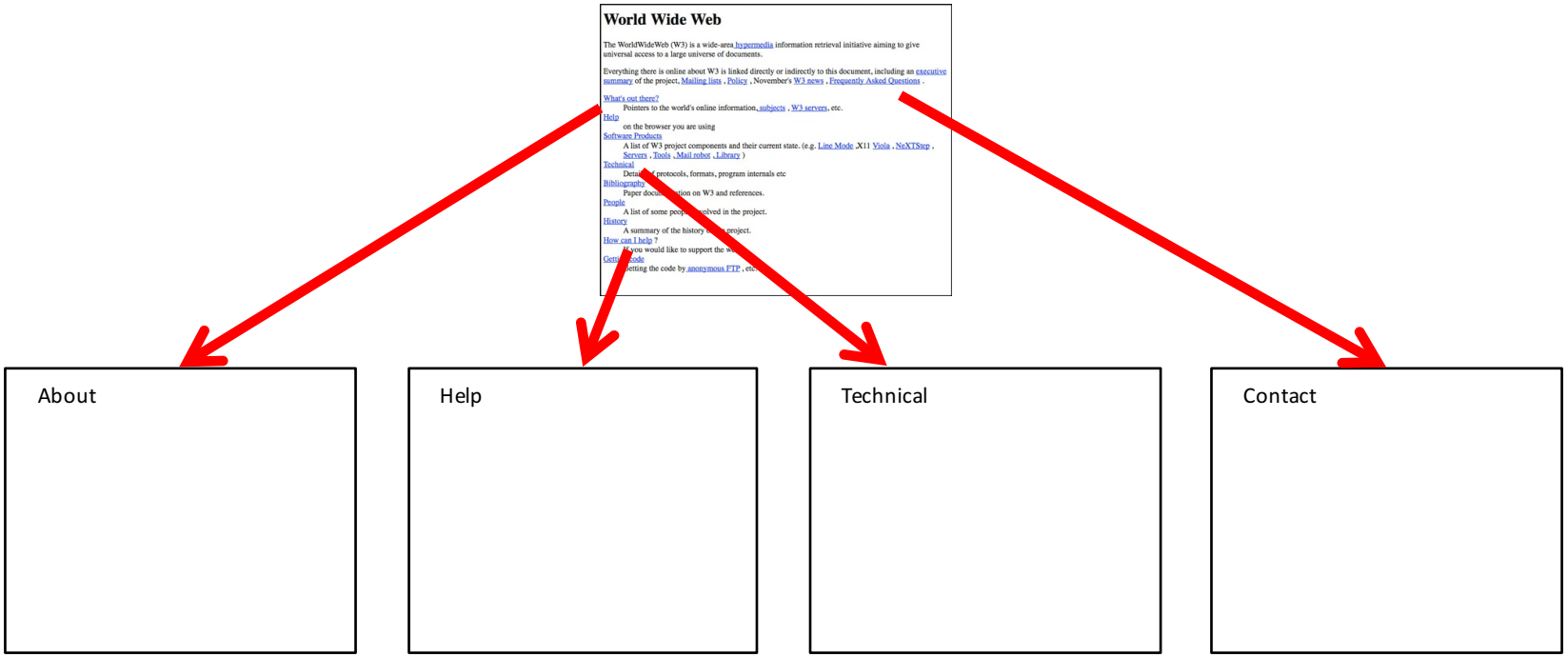
The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

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 .

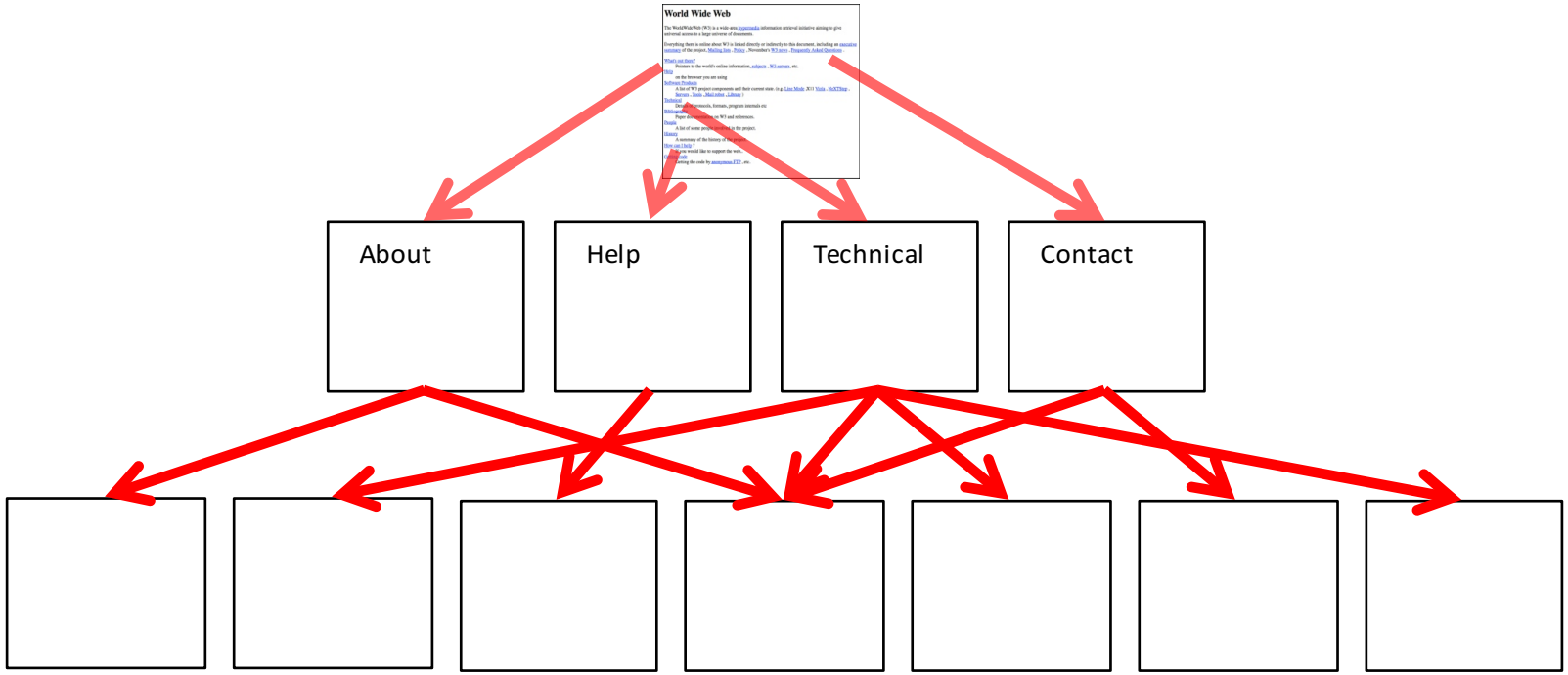
<list>
What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details 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>

Markup Links & Tags

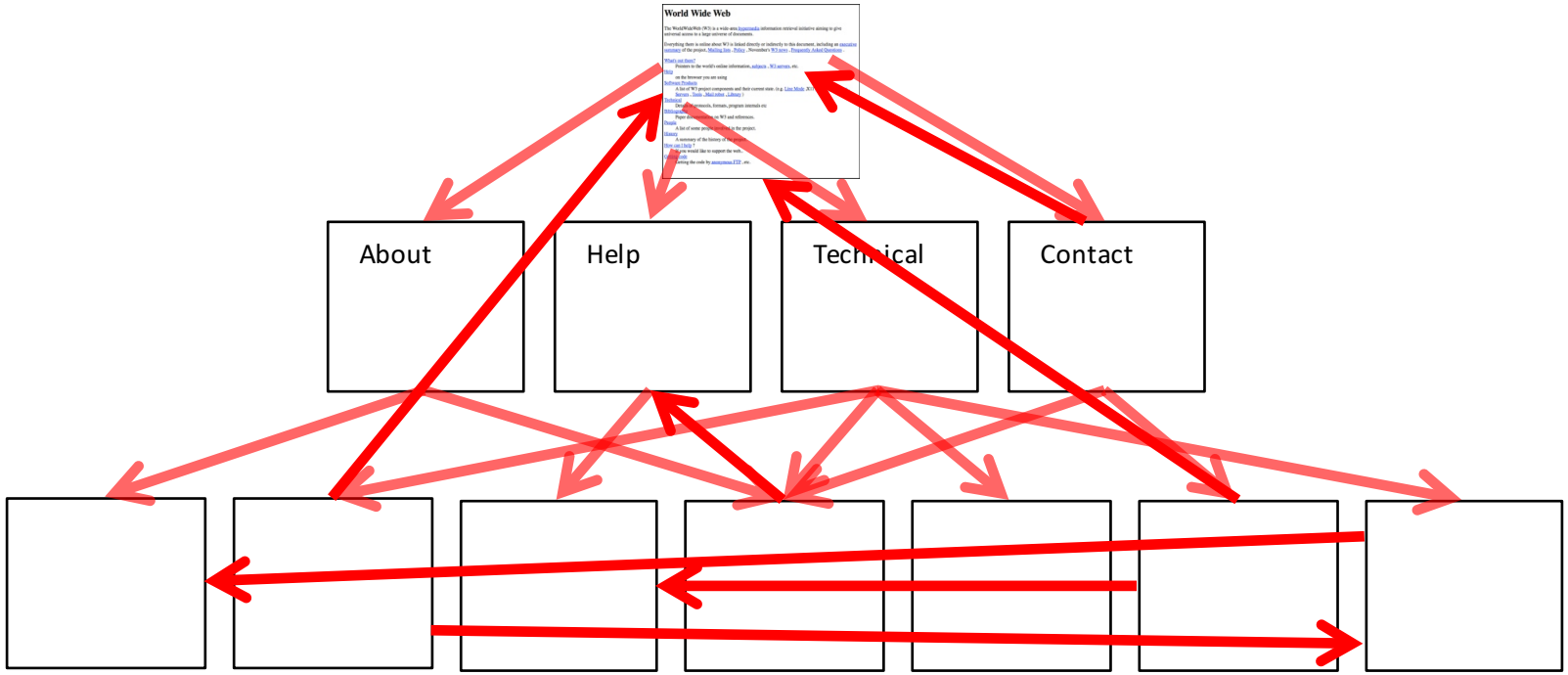
Nodes and Links



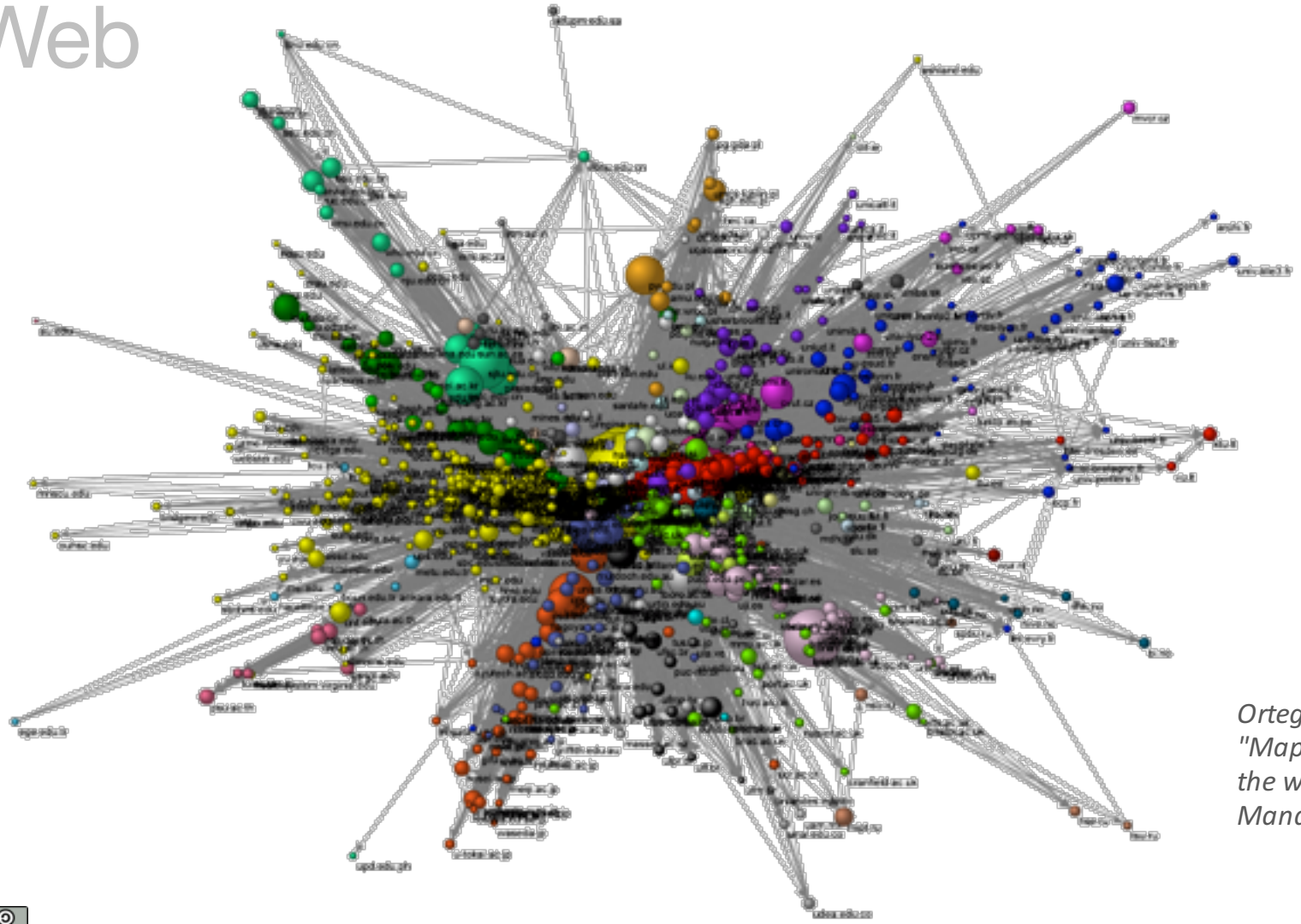
Nodes and Links



Nodes and Links



A Web



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>

<marquee>

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

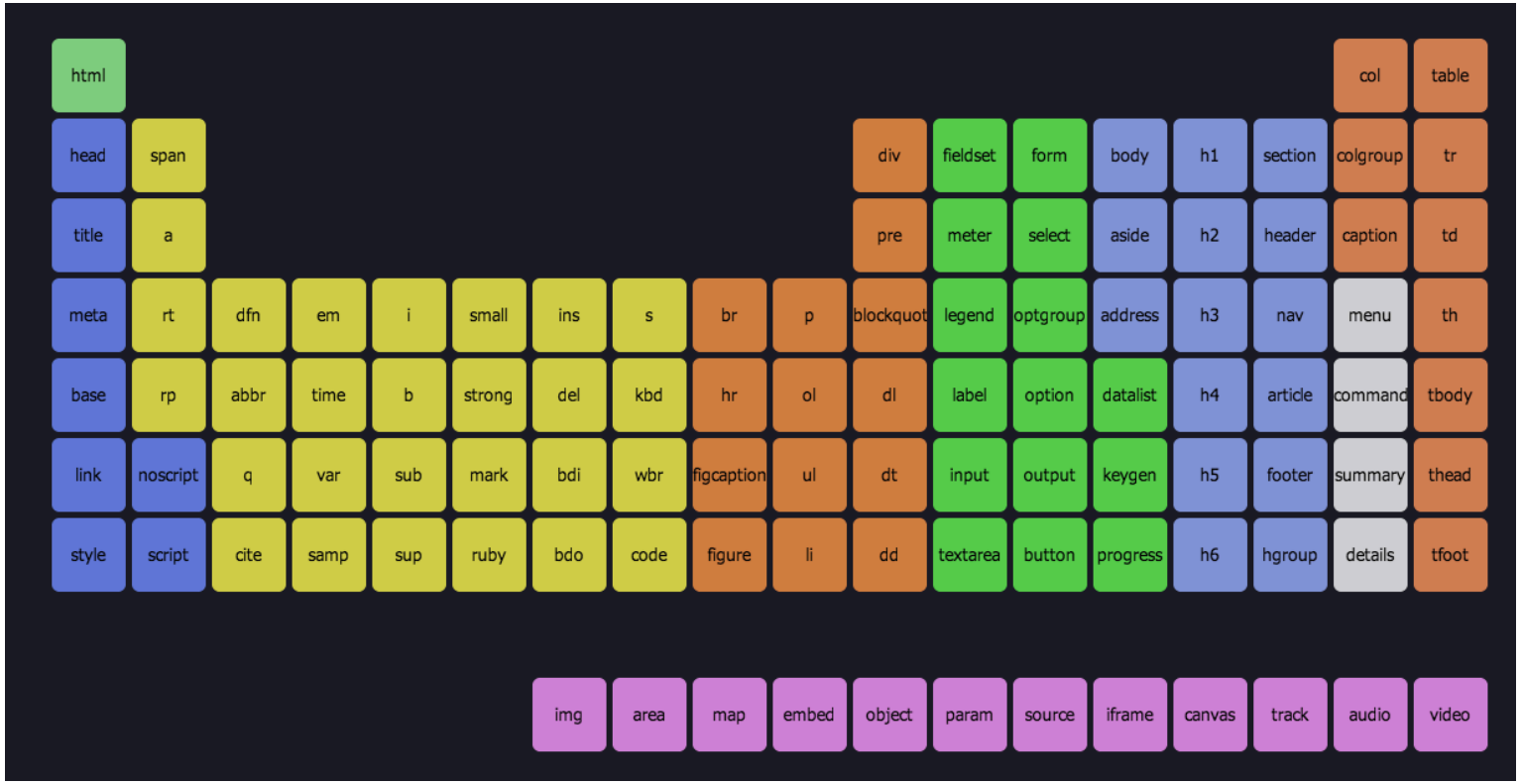
HTML 5

<menu>
<figure>
<nav>
<input>
<header>
<section>
<footer>

<summary>
<details>
<legend>
<label>
<title>
<option>
<a>

~~<blink>~~
~~<marquee>~~

HTML 5



Why Markup?

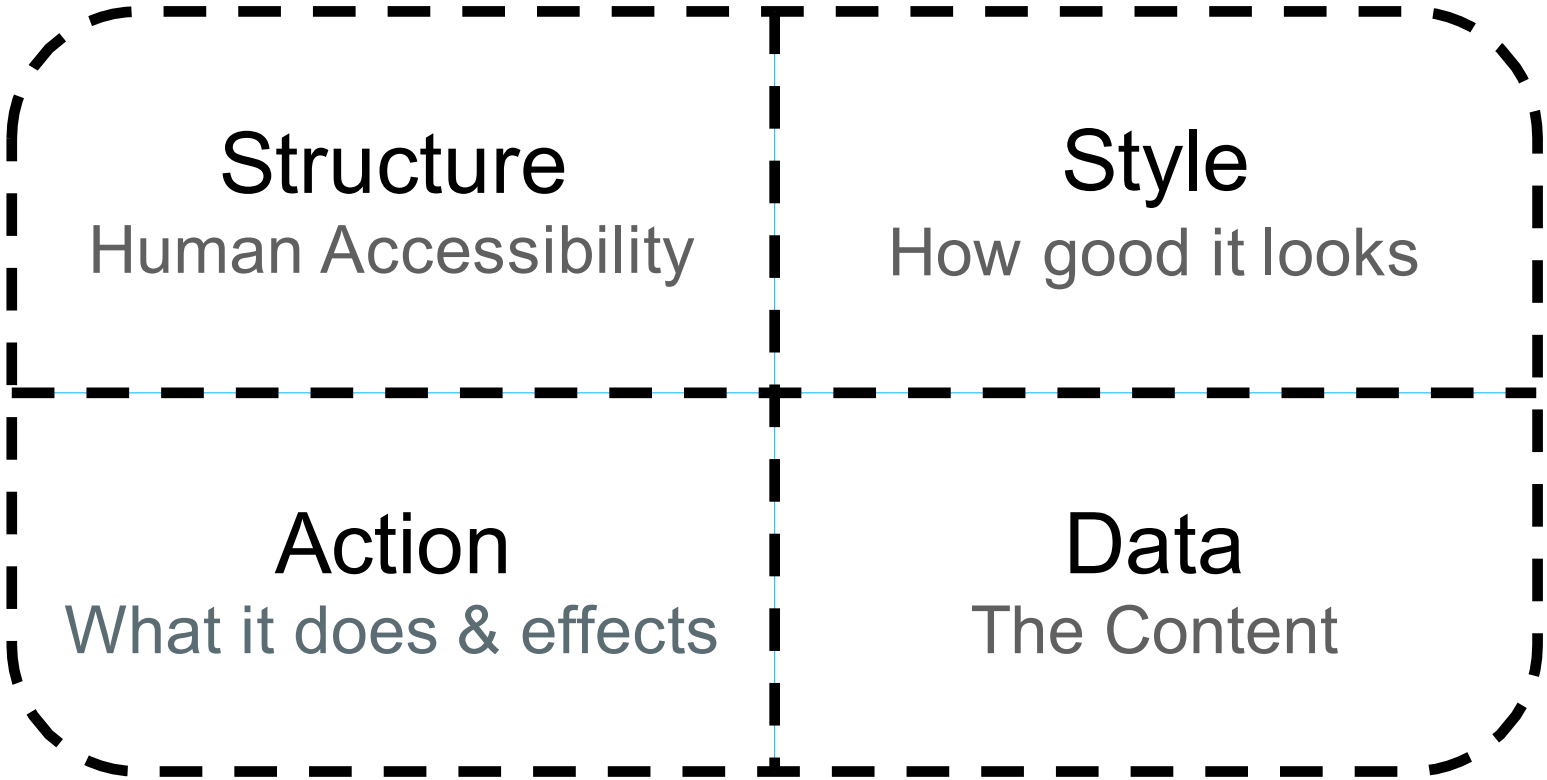
Aids your browser to render the page

Critical for screen readers!

Adds semantics about importance of elements.

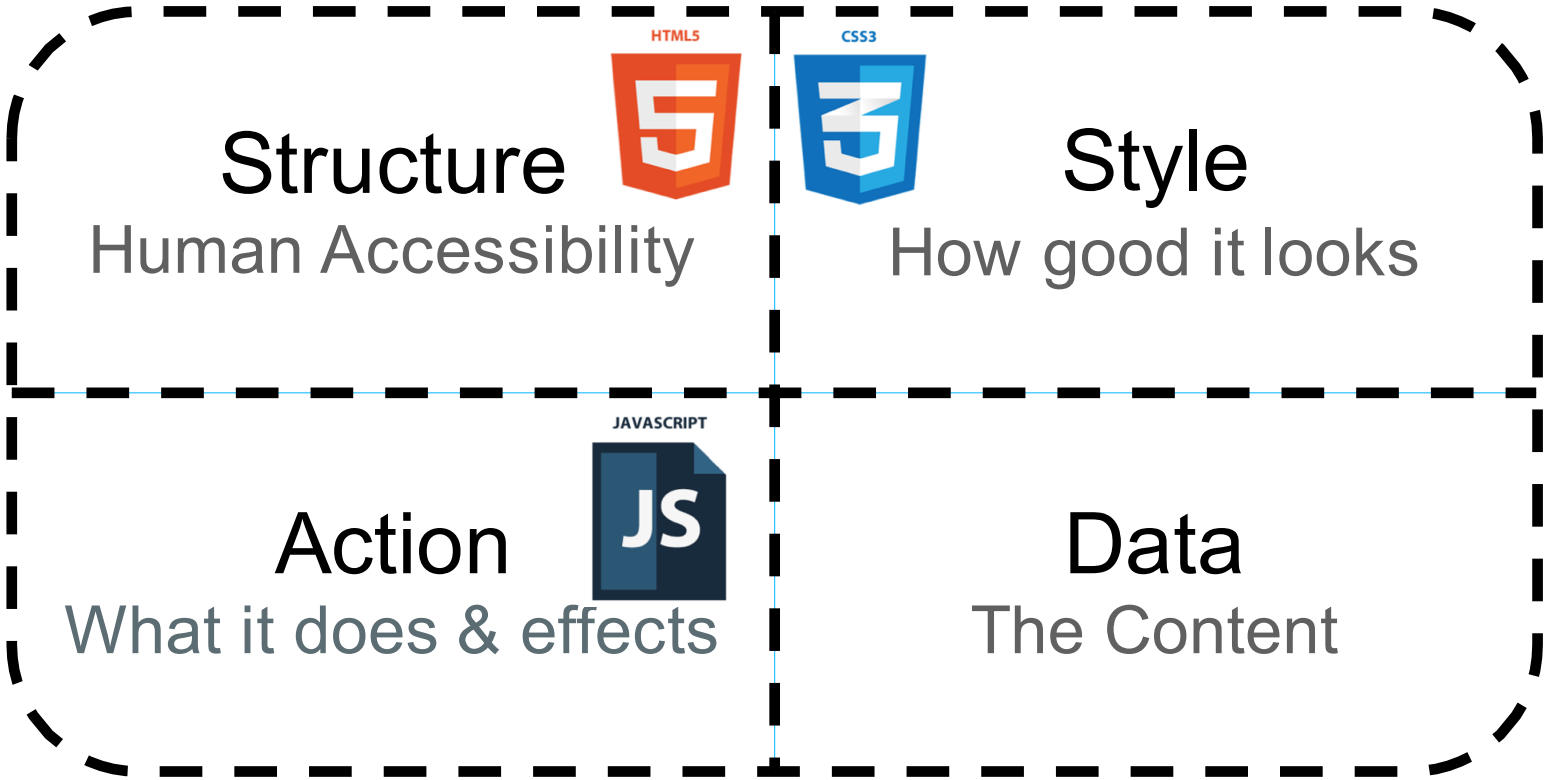
Aids search engines

Building Blocks

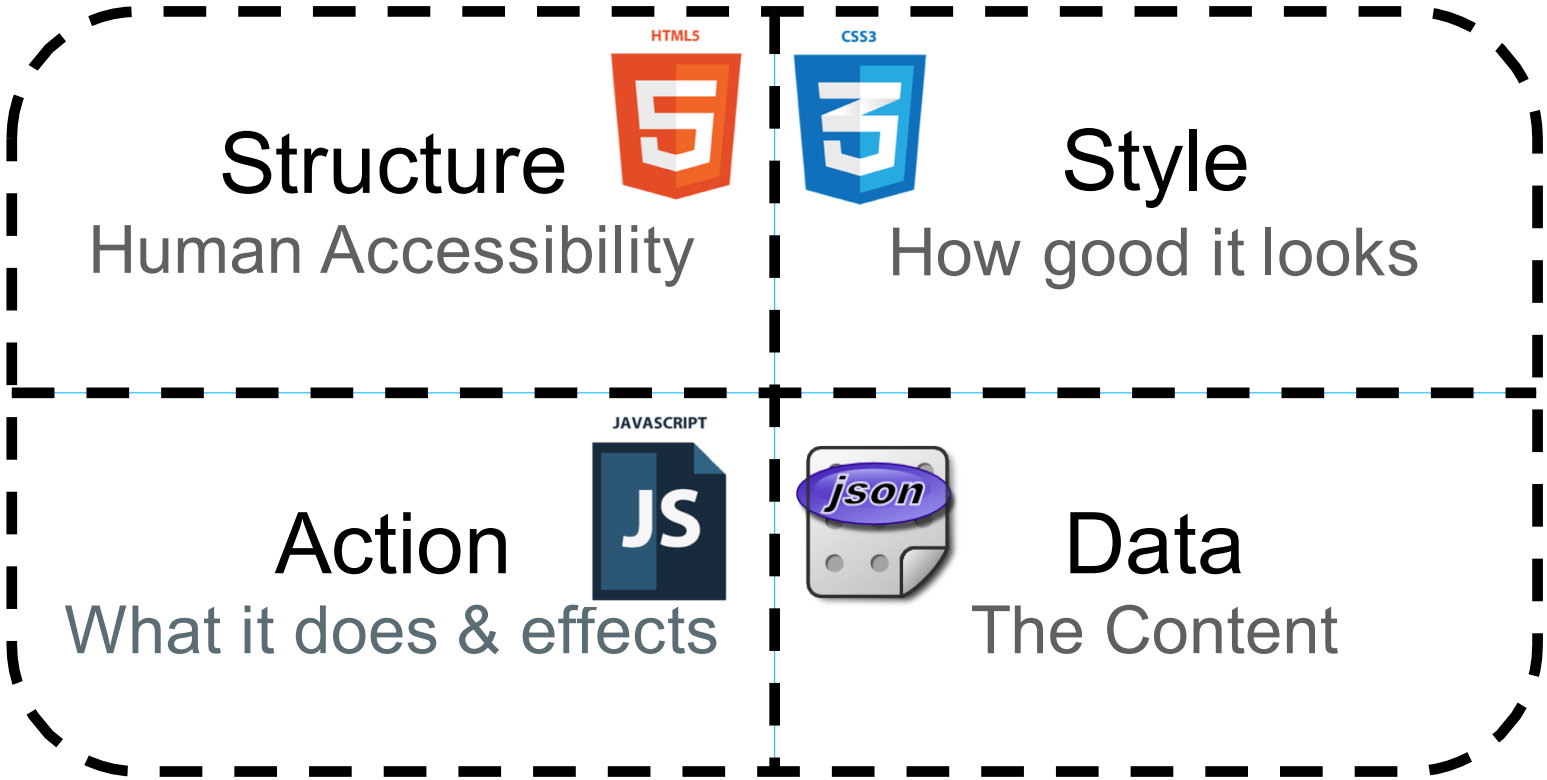


The Language of the Web

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