http://www.flickr.com/photos/restlessglobetrotter

Discovering Data on the Web

UNIVERSITY OF Southampton ODI

Targets

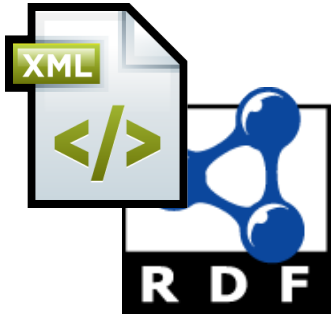
The most important outcomes for this session are to remove the barrier to using data in different **formats** and understand the importance of **identifiers** and **links**.

This session will also introduce some of the history of several fields of study and explain how they are all converging on the web.

Data Formats

Tabular Data

Hierarchical / Tree Data



Data Exchange Formats

Data Formats

Human Consumption

Machine Consumption

Characteristics

Tabular Data

Record based

Best suited to statistical and flat data

Easy to process

Tabular Data

ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA
22498	08/01/2012		The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	WC2H 7JY
23877	24/11/2012		The Duke of Cambridge, Vice Patron, Welsh Rugby Union, and The Duchess of Cambridge ...	DOC,DSSOC	CF10 1NS
23963	13/12/2012		The Queen and The Duke of Edinburgh this morning visited the Bank of England and ...	Q,DE,POW	EC2R 8AH
23963	13/12/2012	BUCKINGHAM PALACE	The Queen and The Duke of Edinburgh this morning visited the Bank of England and ...	Q,DE,POW	SW1A 1AA
24062	29/01/2013		The Duke of York this afternoon visited Northern Ireland Science Park, Queen's ...	DOY	BT26 6AG
24062	29/01/2013		The Duke of York this afternoon visited Northern Ireland Science Park, Queen's ...	DOY	BT3 9DT
24085	06/02/2013		The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	W1D 7LF
24085	06/02/2013		The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	W1S 4BS
24085	06/02/2013	BUCKINGHAM PALACE	The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	SW1A 1AA

data.gov.uk

In your groups, find two interesting datasets on the data.gov.uk website in different formats

What is the dataset about?

What format is the data in?

Is it clear how to interpret and use the data?

What questions would you like to answer using the data?

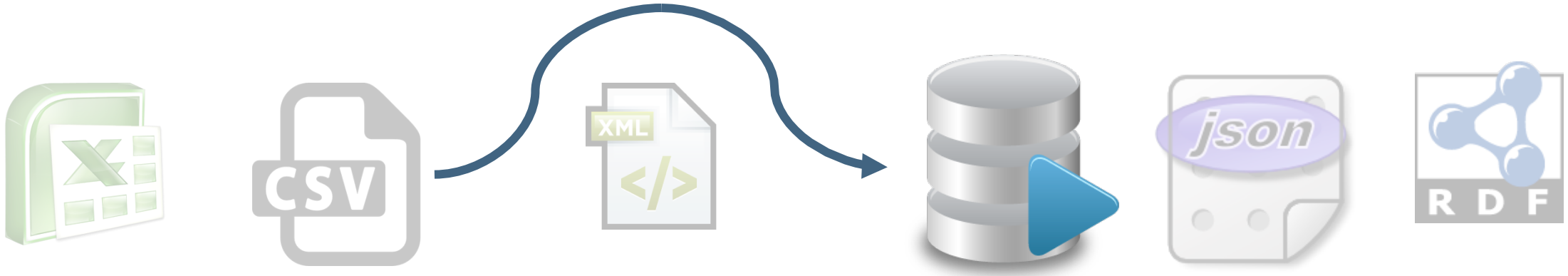
How would you go about answering these questions?

What did we find?

Human Consumption

Machine Consumption

Next Step for Data



Human Consumption

Machine Consumption

Data…bases

Added ability to perform powerful queries on tabular data.

Added **relationships** and a data model.

Tabular Data

ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA
22498	08/01/2012		The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	WC2H 7JY
23877	24/11/2012		The Duke of Cambridge, Vice Patron, Welsh Rugby Union, and The Duchess of Cambridge ...	DOC,DSSOC	CF10 1NS
23963	13/12/2012		The Queen and The Duke of Edinburgh this morning visited the Bank of England and ...	Q,DE,POW	EC2R 8AH
23963	13/12/2012	BUCKINGHAM PALACE	The Queen and The Duke of Edinburgh this morning visited the Bank of England and ...	Q,DE,POW	SW1A 1AA
24062	29/01/2013		The Duke of York this afternoon visited Northern Ireland Science Park, Queen's ...	DOY	BT26 6AG
24062	29/01/2013		The Duke of York this afternoon visited Northern Ireland Science Park, Queen's ...	DOY	BT3 9DT
24085	06/02/2013		The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	W1D 7LF
24085	06/02/2013		The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	W1S 4BS
24085	06/02/2013	BUCKINGHAM PALACE	The Duke of York, Patron, this morning visited Code Club at Soho Parish Church of ...	DOY	SW1A 1AA

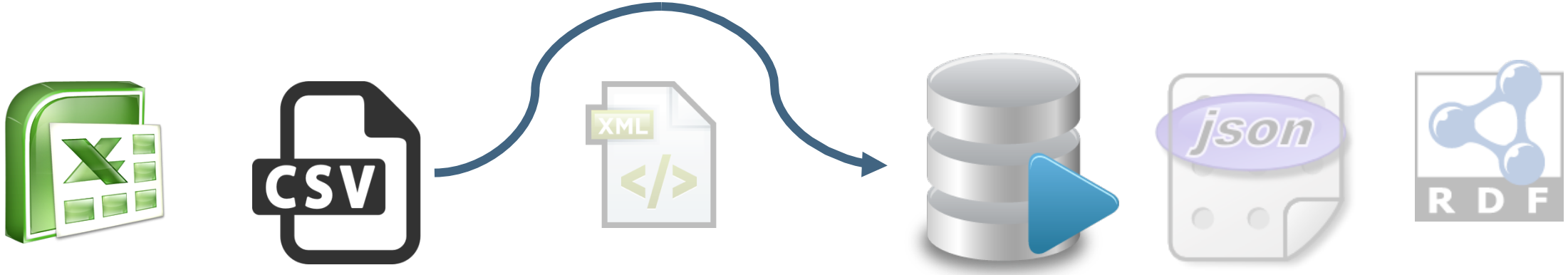
Normalisation

ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA

MRF	Name	Position
DOC	William Arthur Philip Louis Windsor	Duke of Cambridge
DSSOC	Catherine Elizabeth Mountbatten-Windsor	Duchess of Chambridge

Postcode	Place Name
SW1A 1AA	Buckingham Palace

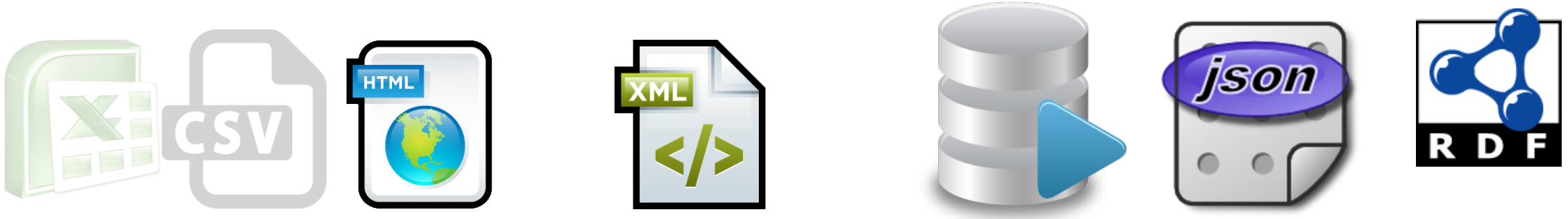
Without the Web



Human Consumption

Machine Consumption

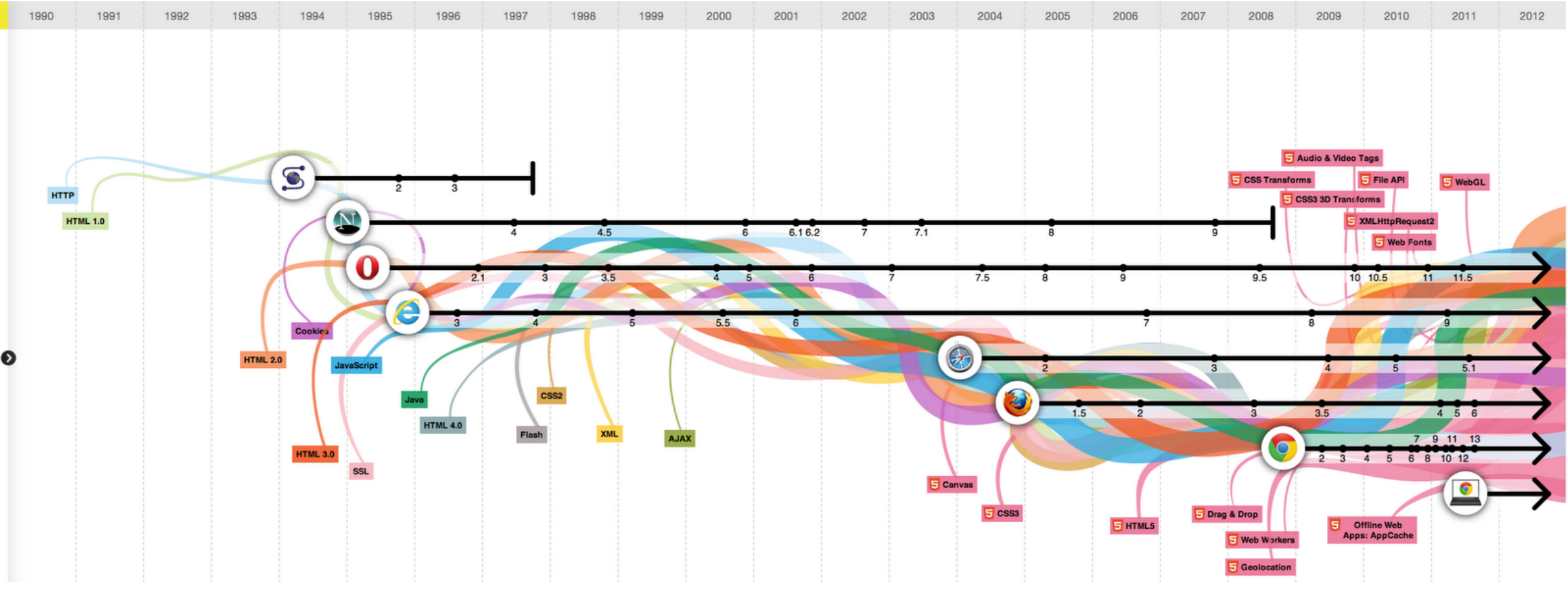
The Web



Human Consumption

Machine Consumption

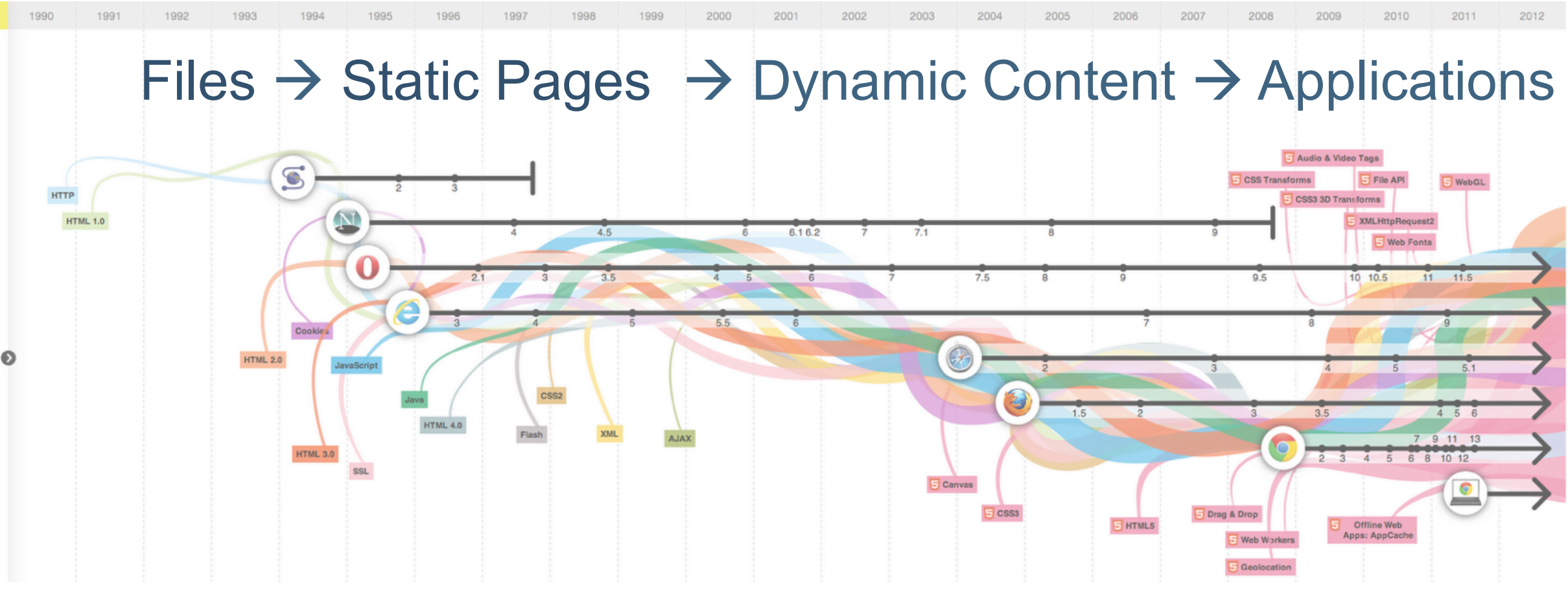
Evolution of the Web



http://www.evolutionoftheweb.com/

Evolution of the Web

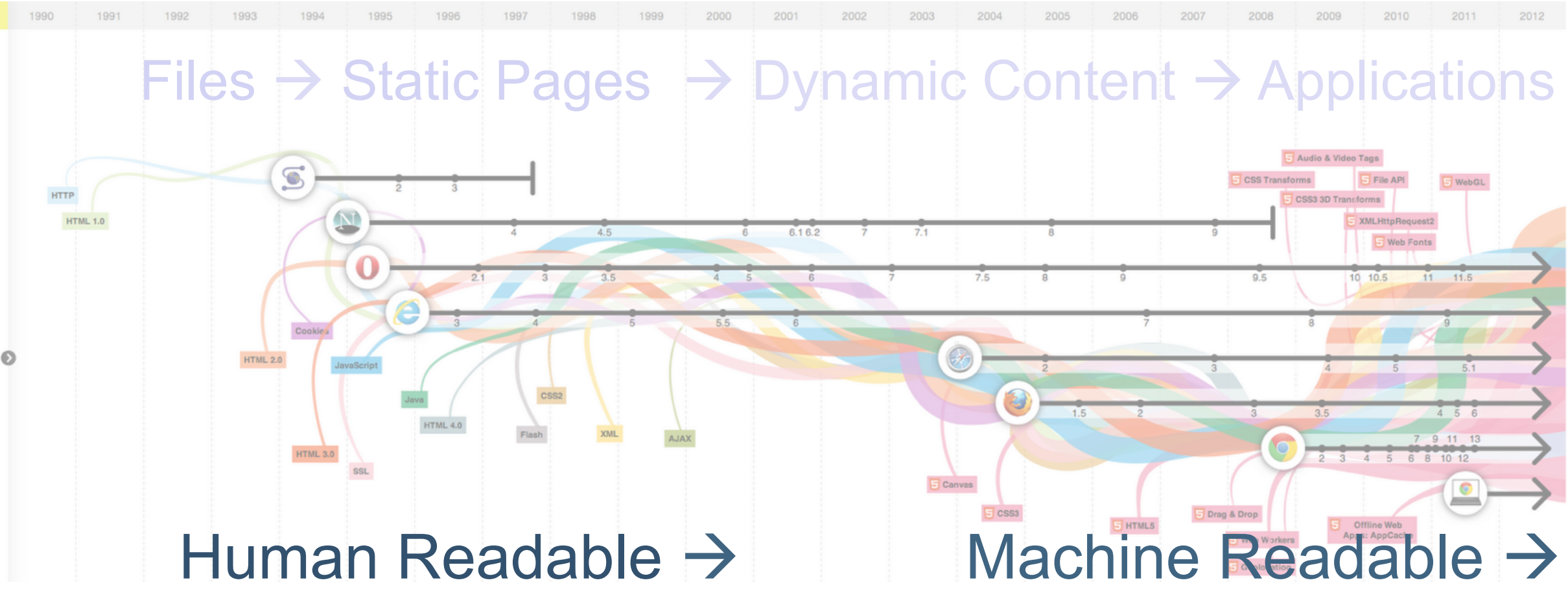
Files → Static Pages → Dynamic Content → Applications



http://www.evolutionoftheweb.com/

Evolution of the Web

Files → Static Pages → Dynamic Content → Applications



Human Readable → Machine Readable →

http://www.evolutionoftheweb.com/

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.

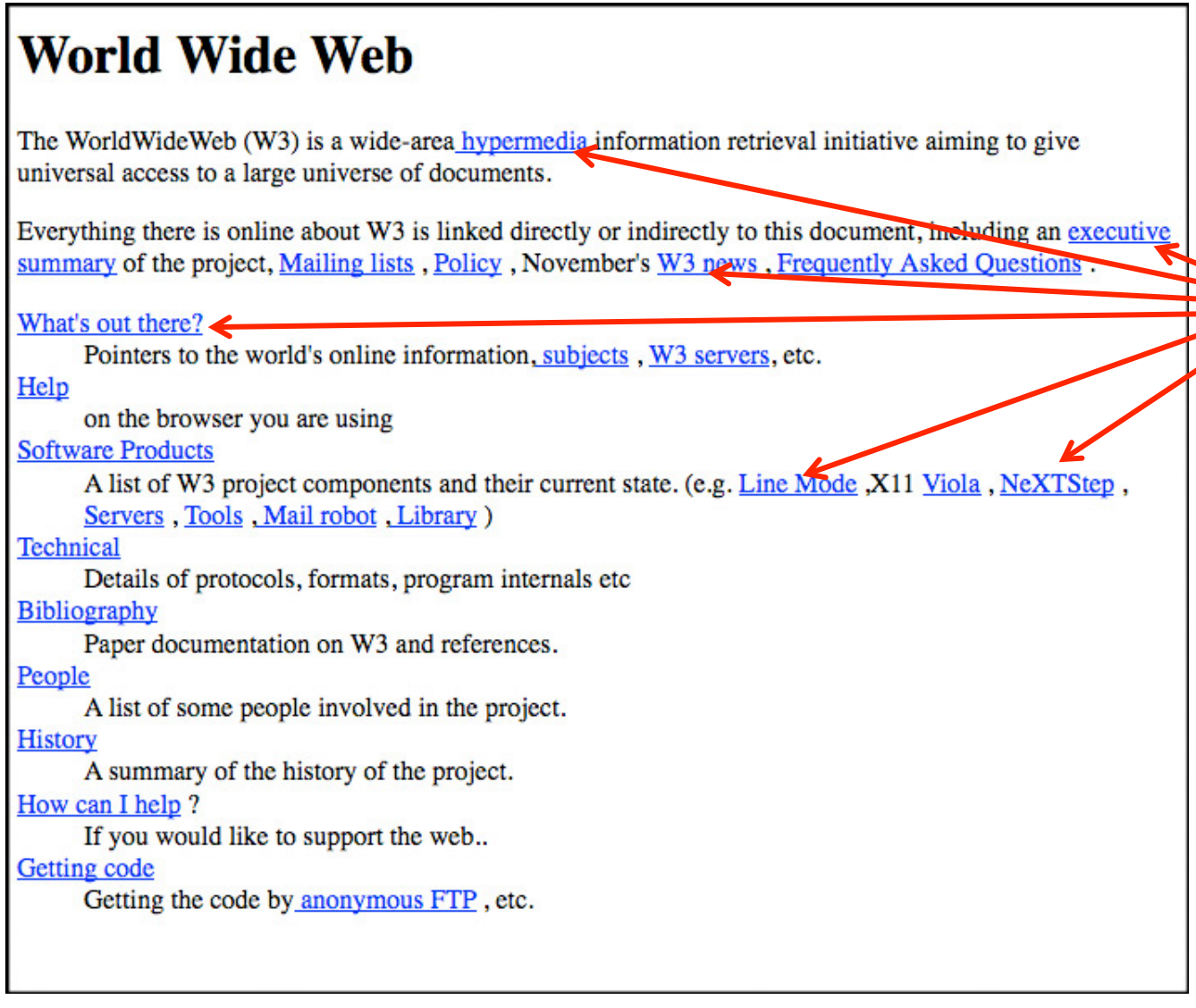
Human Readable Web

Question 1

What is significant about the first web page?

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>

 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.

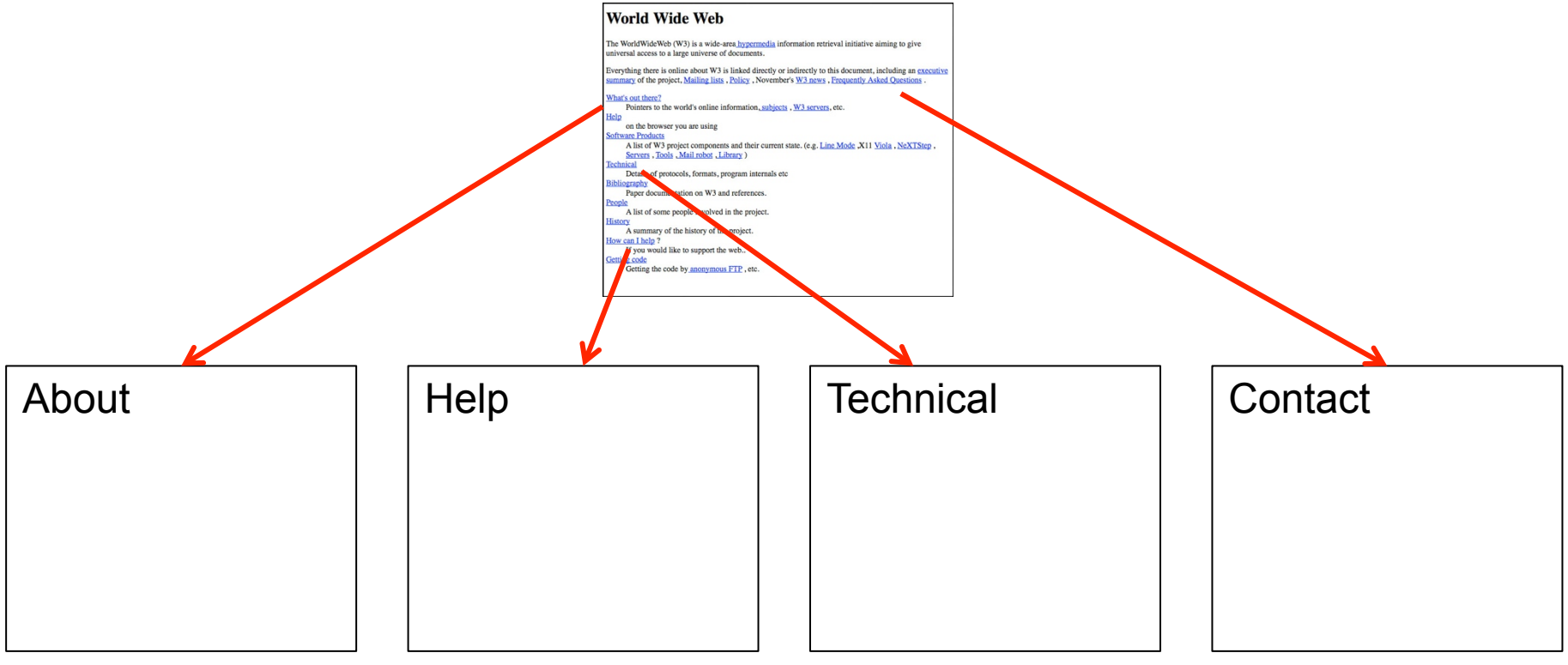
People
A list of some people involved in the project.

History
A summary of the history of the project.

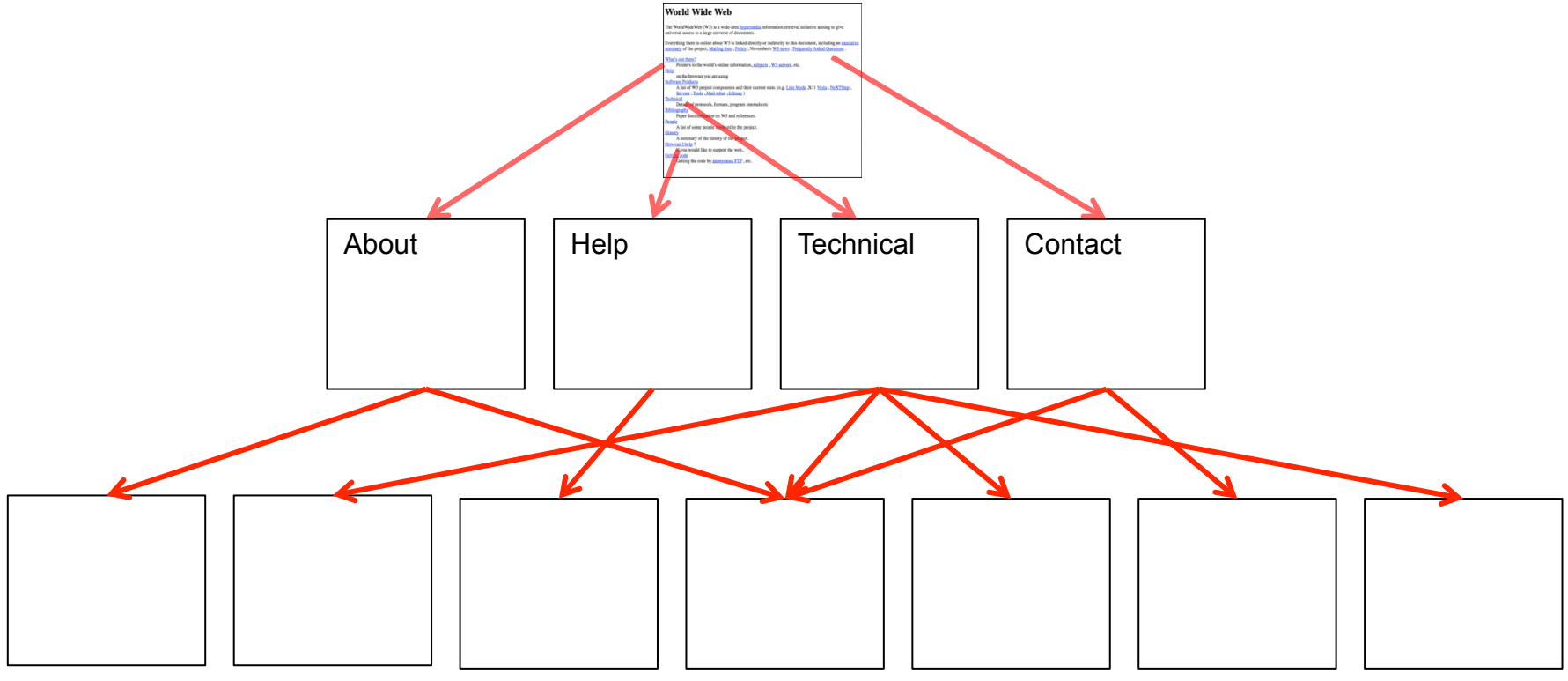
How can I help ?
If you would like to support the web..

Getting code
Getting the code by anonymous FTP , etc.
</list>

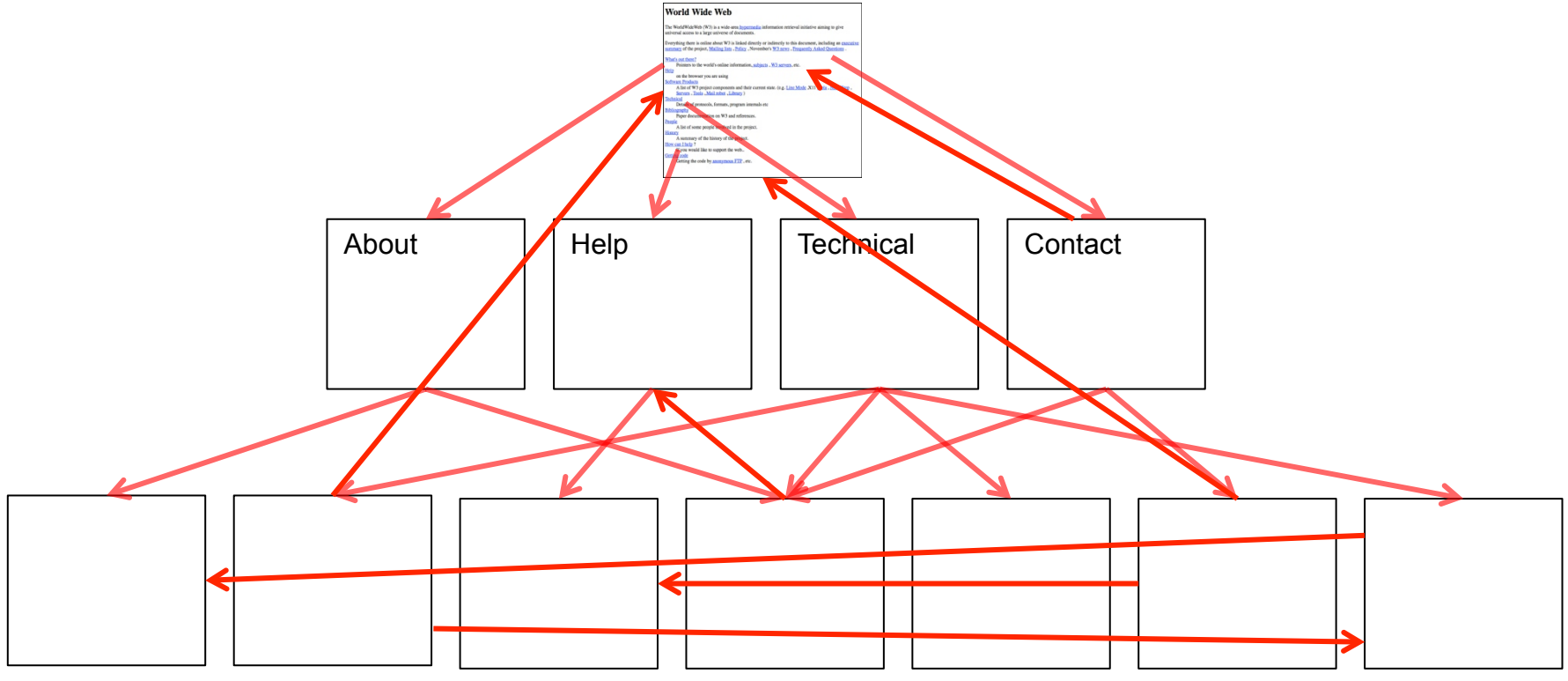
Nodes and Links



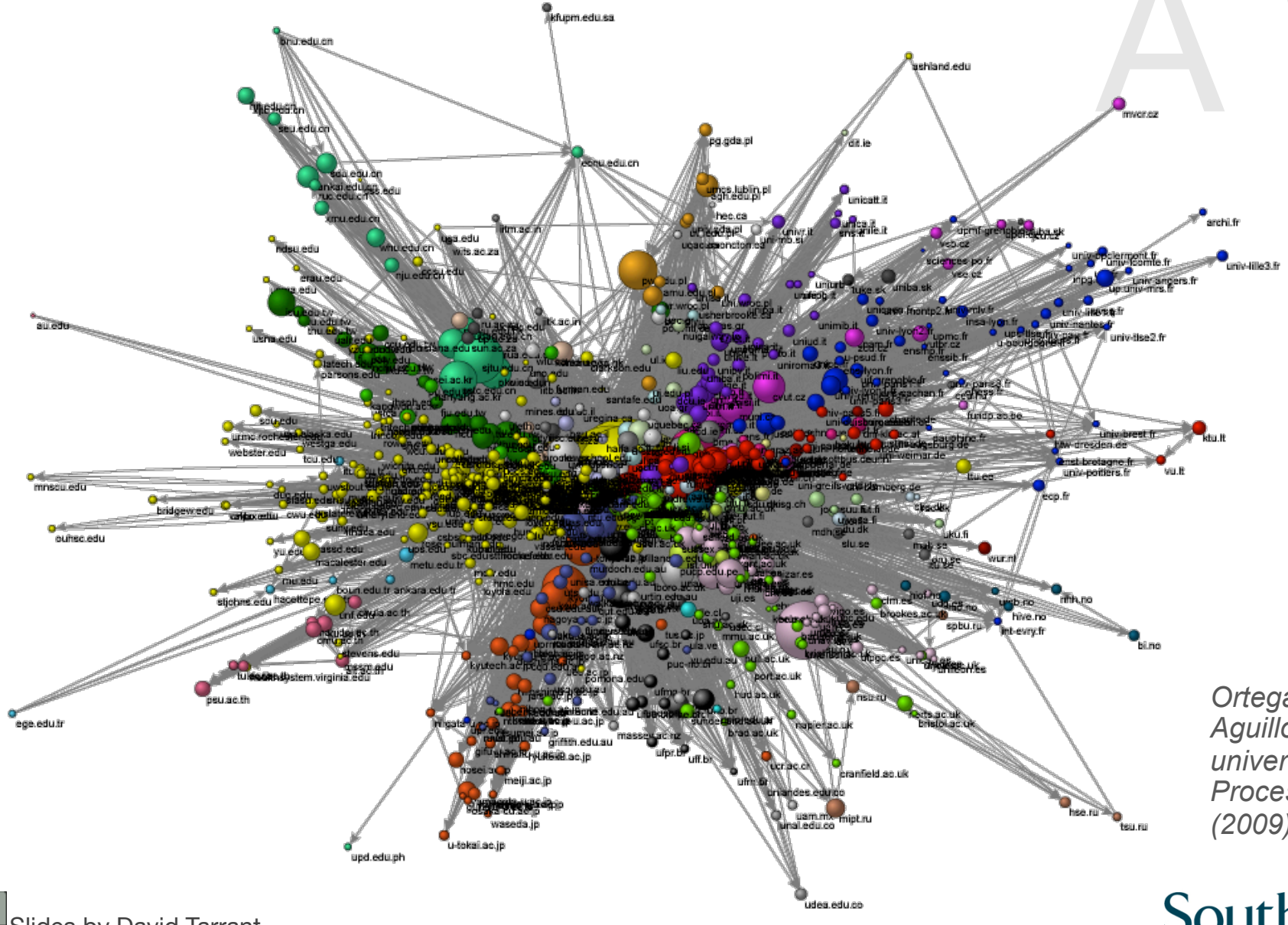
Nodes and Links



Nodes and Links



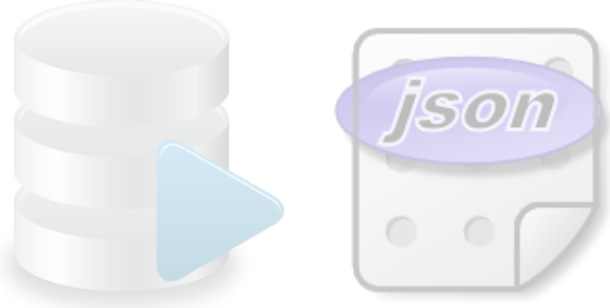
A Web



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

UNIVERSITY OF Southampton

The Web



Human Consumption

Machine Consumption

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>

</marquee>

 Slides by David Tarrant

<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

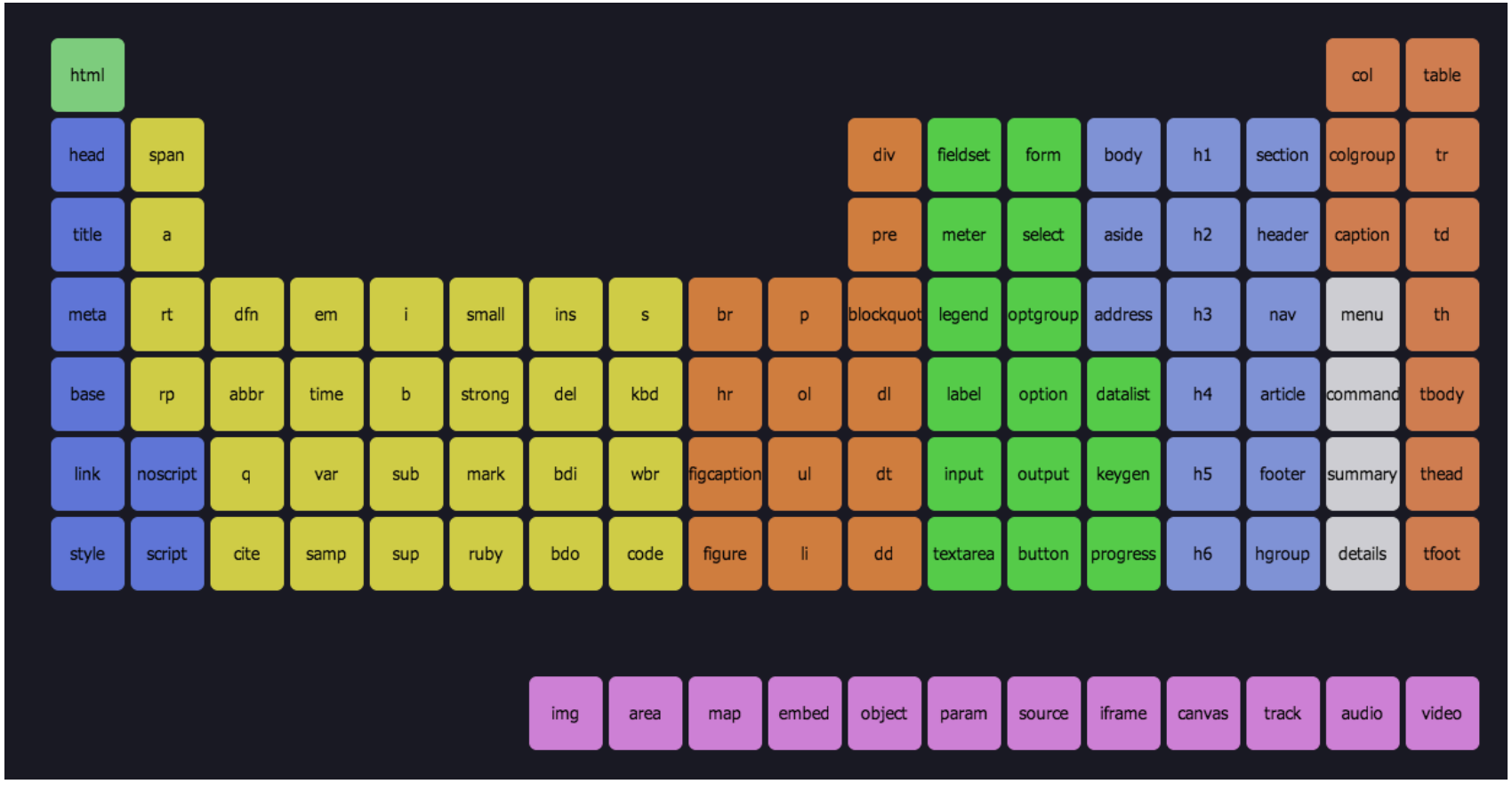
<menu>
<figure>
<nav>
<input>
<header>
<section>
<footer>

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

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

UNIVERSITY OF Southampton ODI

HTML 5

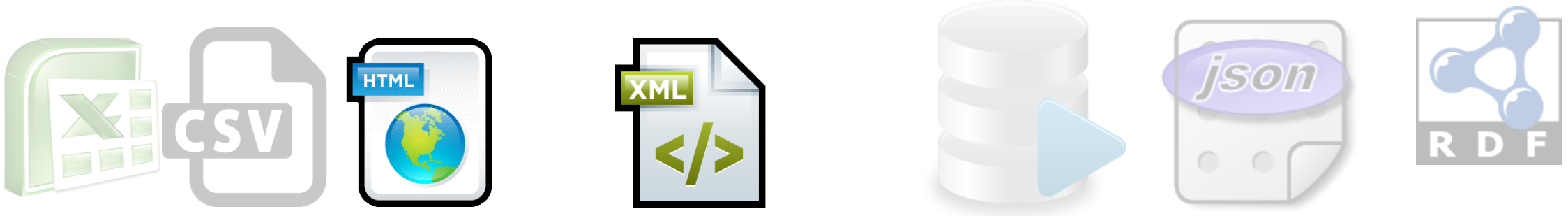


UNIVERSITY OF Southampton

Why Markup?

- Aids your browser to render the page
- Critical for screen readers!
- Adds semantics about importance of elements.
- Aids search engines

The Web



Human Consumption

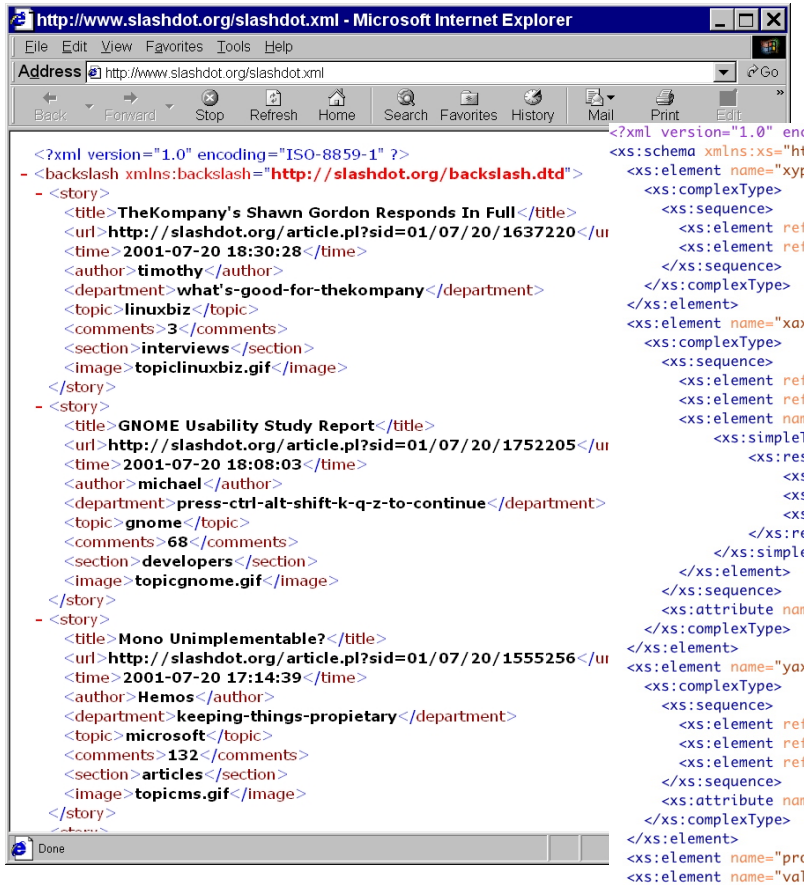
Machine Consumption

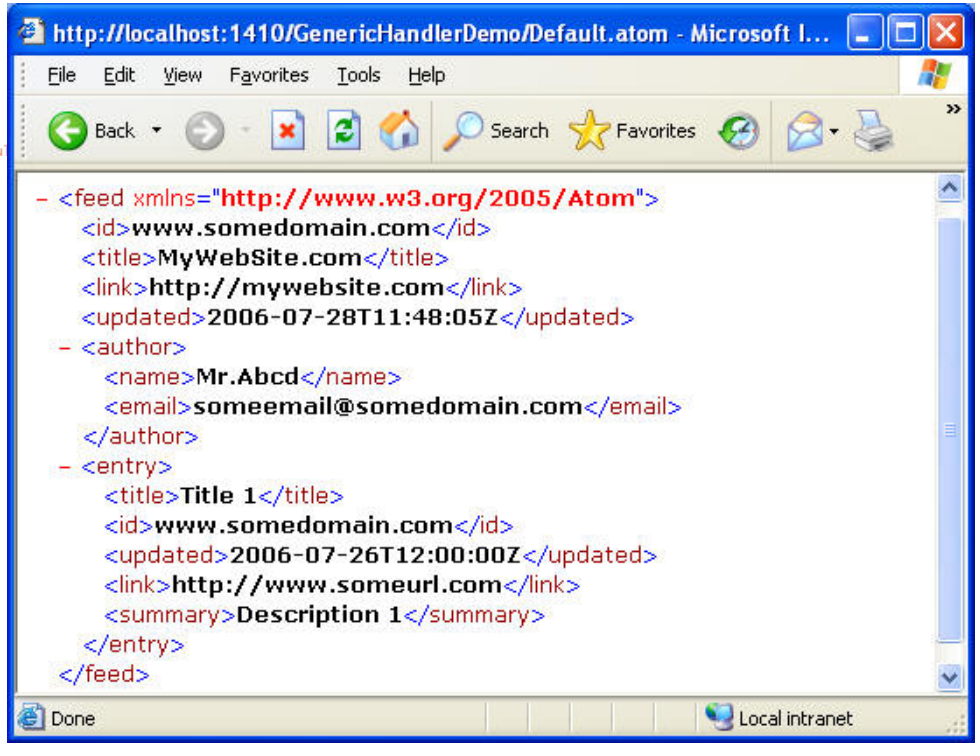
Data on the Web

more markup…

```
<book>
  <title>Winnie the Pooh</title>
  <author>A. A. Milne</author>
</book>
```

Markup for anything





Slides by David Tarrant

UNIVERSITY OF Southampton

Data Exchange

eXtensible Markup Language

Forms the basis for hundreds of XML formats






Resource Description Framework

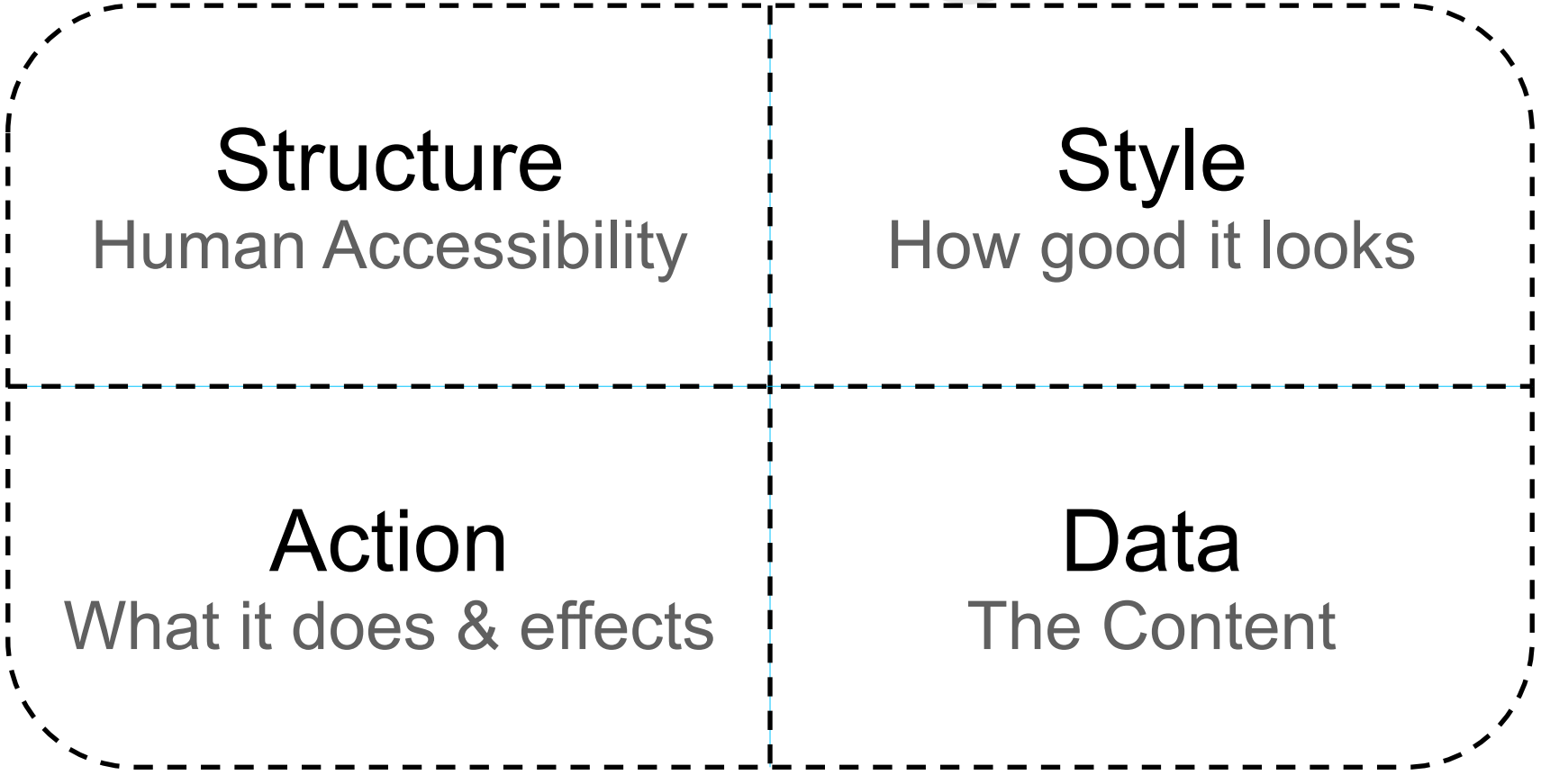
Enforces a structured relationship between elements

Exercise

Building a Web Page

that reads the BBC News Feed

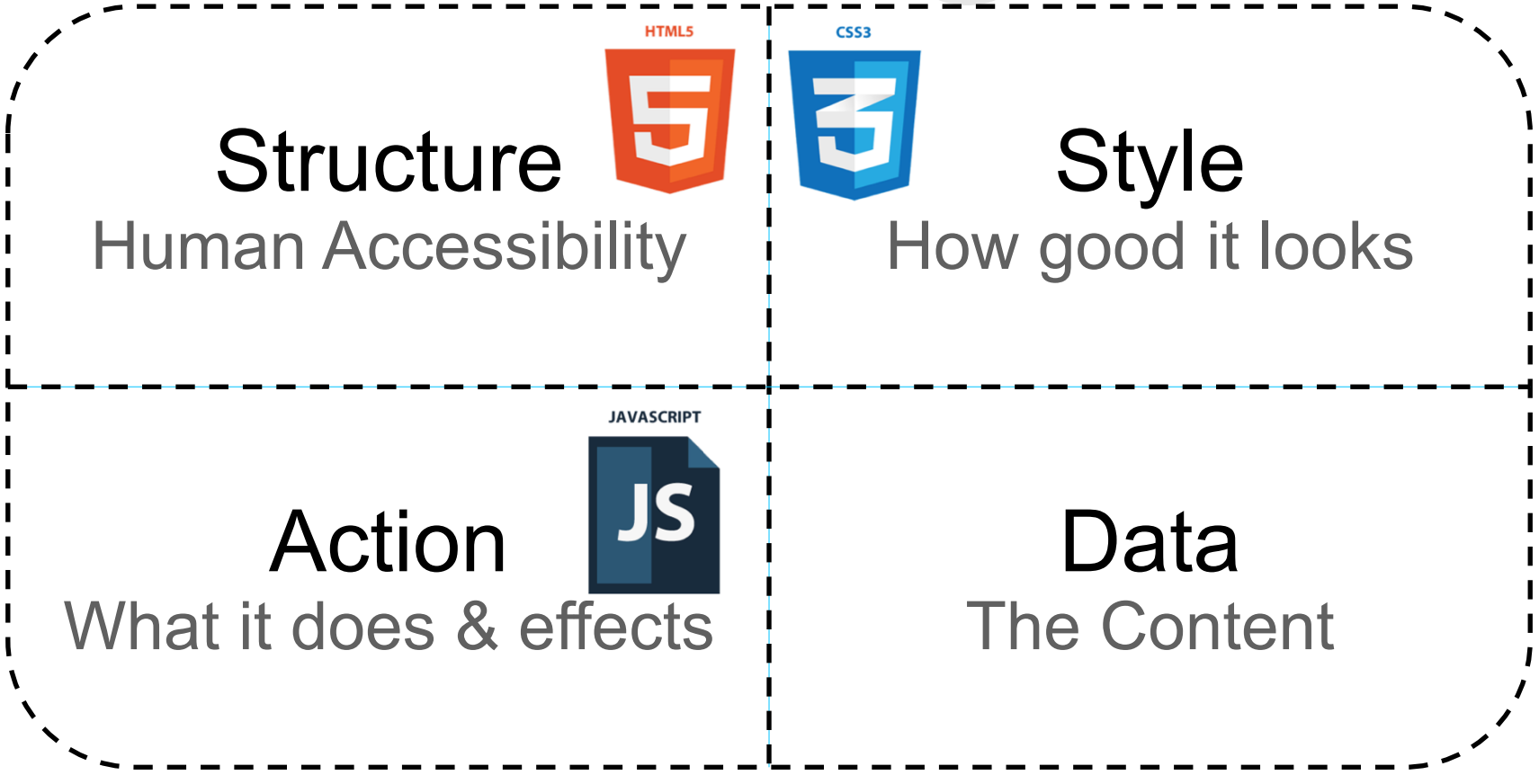
Building Blocks



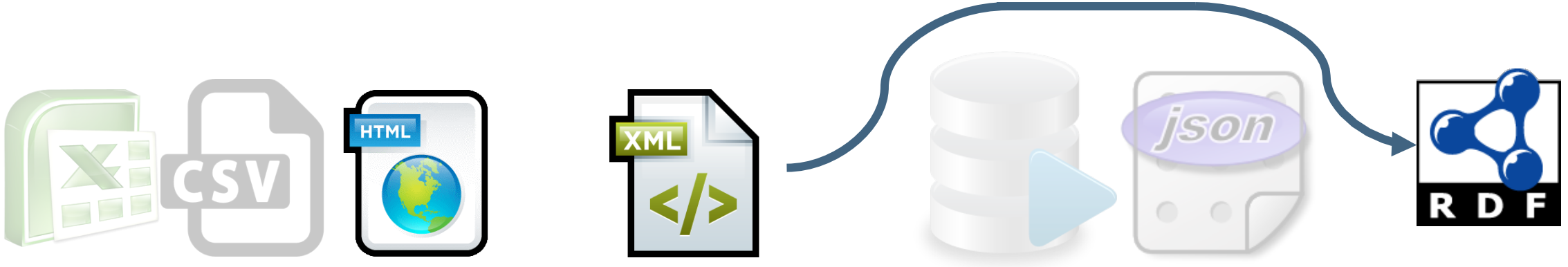
The Language of the Web

UNIVERSITY OF Southampton

Building Blocks



The Web



Human Consumption

Machine Consumption

Problems with XML

Unbelievably, XML is a challenge for a machine to parse.

All the parser[illegible] hard to use,

XML has [illegible] semantics

```
<v>
  <x>
   <y> a="ppppp"</y>
   <z>
      <w>qqqqq</w>
   </z>
  </x>
</v>
```

Introducing RDF

The **R**esource **D**escription **F**ramework is a data model.

(It is not a file format)

The RDF data model is similar to classic conceptual modeling approaches such as entity–relationship or class diagrams (from databases).

It is based upon the idea of making statements about resources (in particular web resources) in the form of subject-predicate-object expressions.

Normalisation

ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA

MRF	Name	Position
DOC	William Arthur Philip Louis Windsor	Duke of Cambridge
DSSOC	Catherine Elizabeth Mountbatten-Windsor	Duchess of Chambridge

Postcode	Place Name
SW1A 1AA	Buckingham Palace

Identifiers

ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA

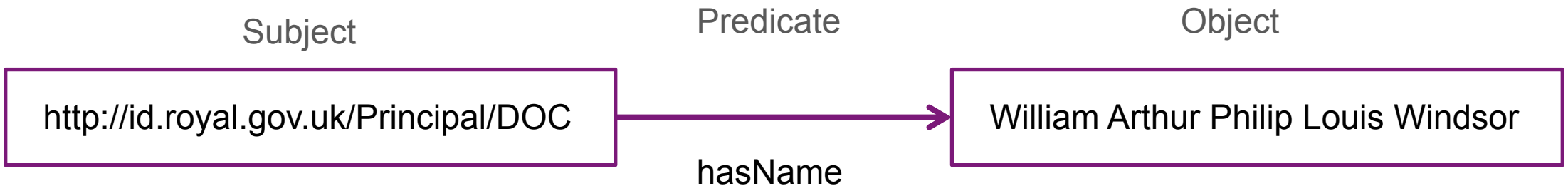
http://id.royal.gov.uk/Principal/DOC

hasName

MRF	Name	Position
DOC	William Arthur Philip Louis Windsor	Duke of Cambridge
DSSOC	Catherine Elizabeth Mountbatten-Windsor	Duchess of Chambridge

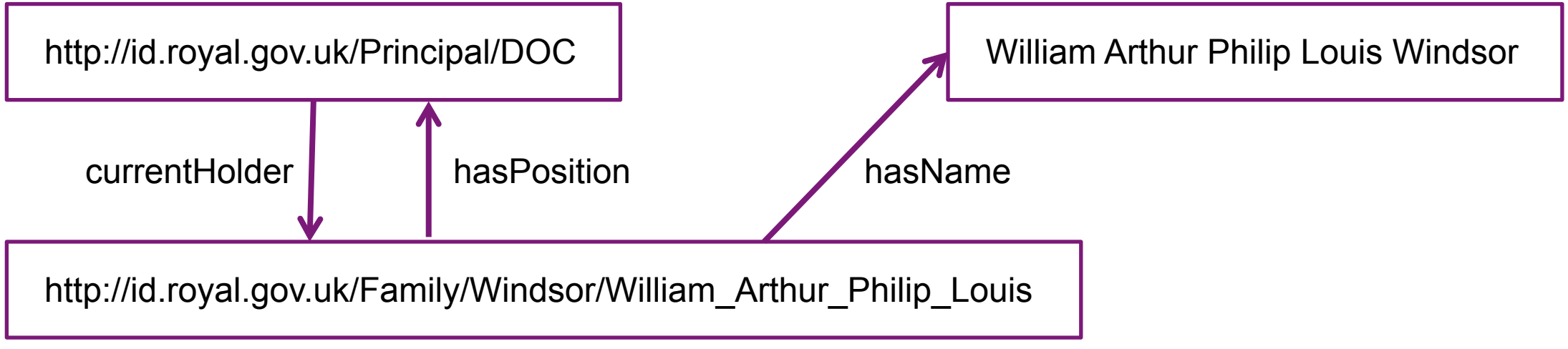
William Arthur Philip Louis Windsor

A Triple!



These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object.

Normalising



Why RDF

One syntax!

Any system that understands the syntax can use the data!

It is a near as we have come so far to a common standard for knowledge transfer.

Why RDF

It encourages use of URIs to identify things.

It is the basis for the Web of Linked Data and the Semantic Web.

Data Exchange

eXtensible Markup Language

Forms the basis for hundreds of XML formats






Resource Description Framework

Enforces a structured relationship between elements

Data Exchange

eXtensible Markup Language

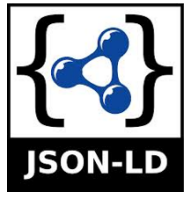
Forms the basis for hundreds of XML formats






Resource Description Framework

Enforces a structured relationship between elements



Data Formats

Human Consumption

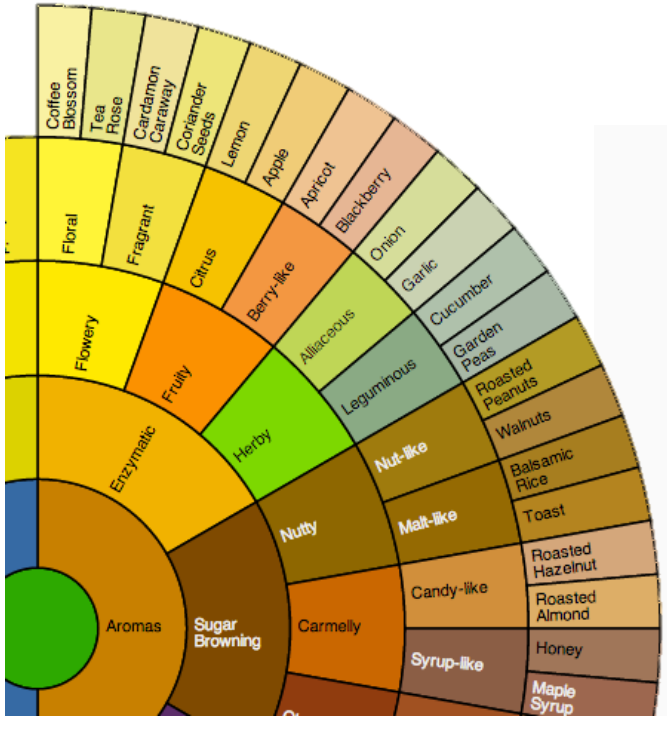
Machine Consumption

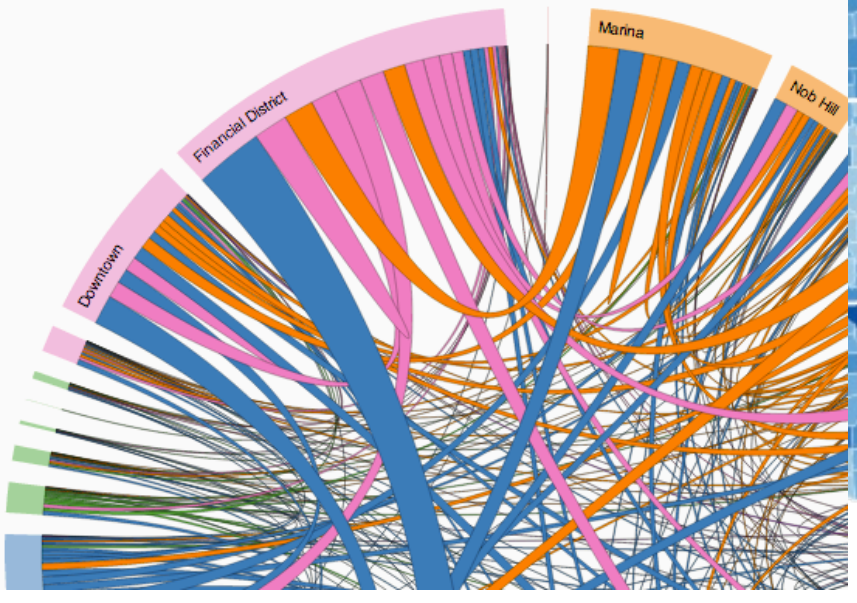
Data Formats

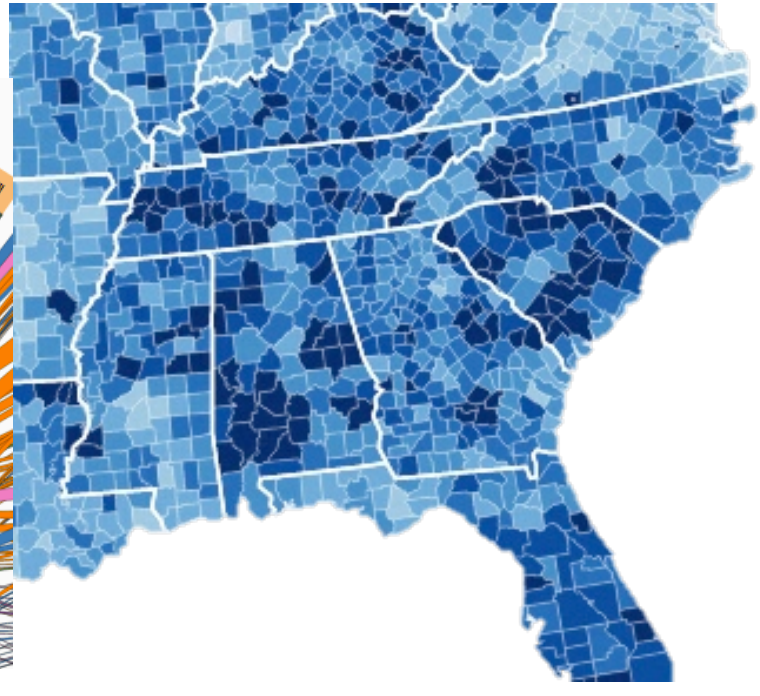
Human Consumption

Machine Consumption

Using Data







UNIVERSITY OF Southampton

Programmatic Data

JSON – JavaScript Object Notation

Can be directly used by Javascript without need to parse to different structure.

Your transfer objects and references, rather than serialised* versions of the same thing.

* JSON is a data serialisation format as well

JSON: A Natural Fit

Exercise

Building a Web Page

that reads the BBC News Feed

Recap

The Evolution of the Web

The Evolution of Data

Intro to Identifiers and Linking

Knowledge representation

Serialisations and data discovery