



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

Discovering Data on the Web

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.



Slides by David Tarrant

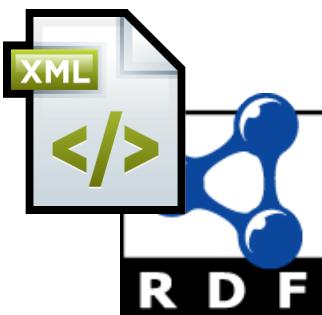
Data Formats



Tabular Data



Hierarchical / Tree Data



Data Exchange Formats

Data Formats



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





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?

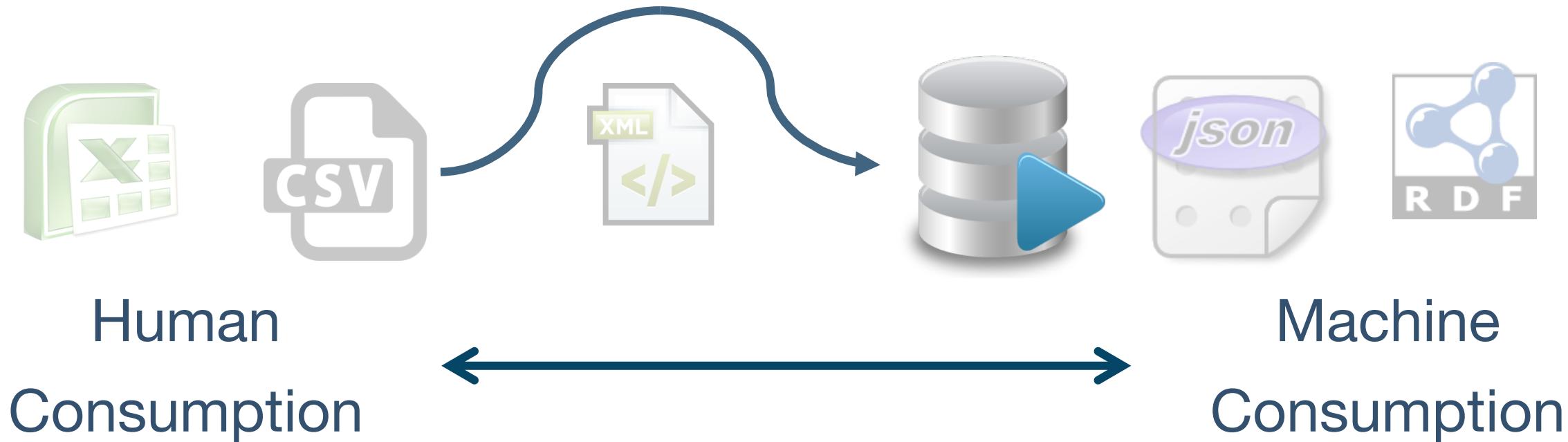


Slides by David Tarrant

What did we find?

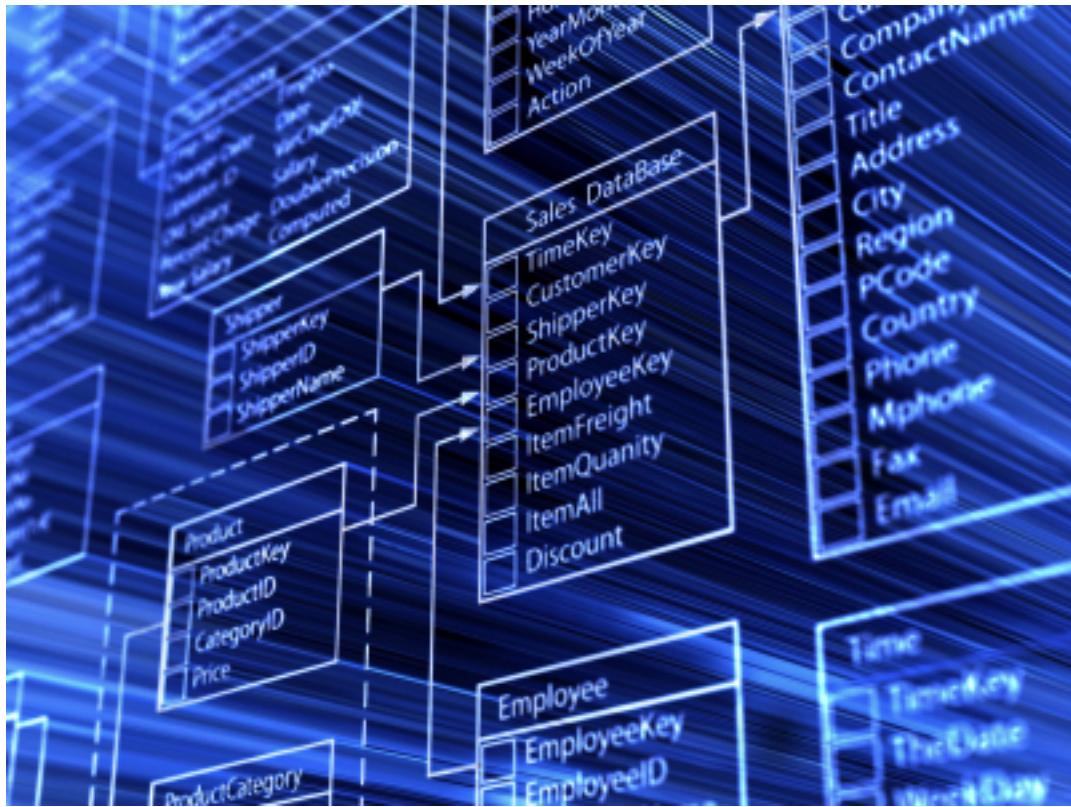


Next Step for Data





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

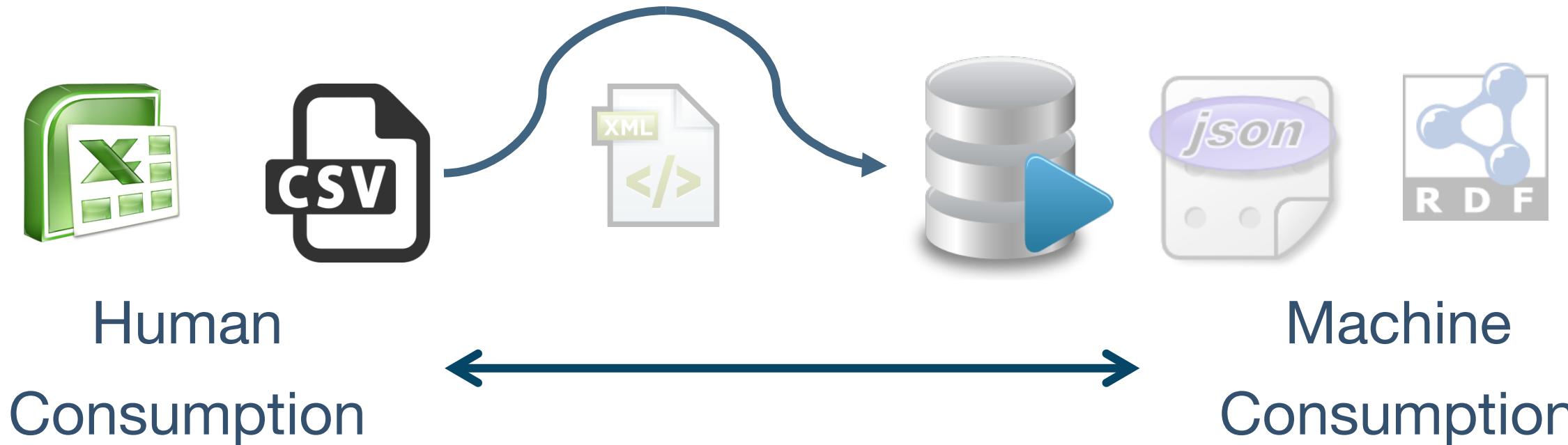


Normalisation

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		Postcode	Place Name
DOC	William Arthur Philip Louis Windsor	Duke of Cambridge		SW1A 1AA	Buckingham Palace
DSSOC	Catherine Elizabeth Mountbatten-Windsor	Duchess of Cambridge			

A diagram illustrating data normalization. It shows two tables: a main table and a detailed table. In the main table, the 'Residence' field contains 'BUCKINGHAM PALACE'. Arrows point from this cell to the 'Name' and 'Position' fields in the detailed table, which correspond to the Duke and Duchess of Cambridge. In the main table, the 'PostCode' field contains 'SW1A 1AA'. An arrow points from this cell to the 'Place Name' field in the detailed table, which is 'Buckingham Palace'.

Without the Web



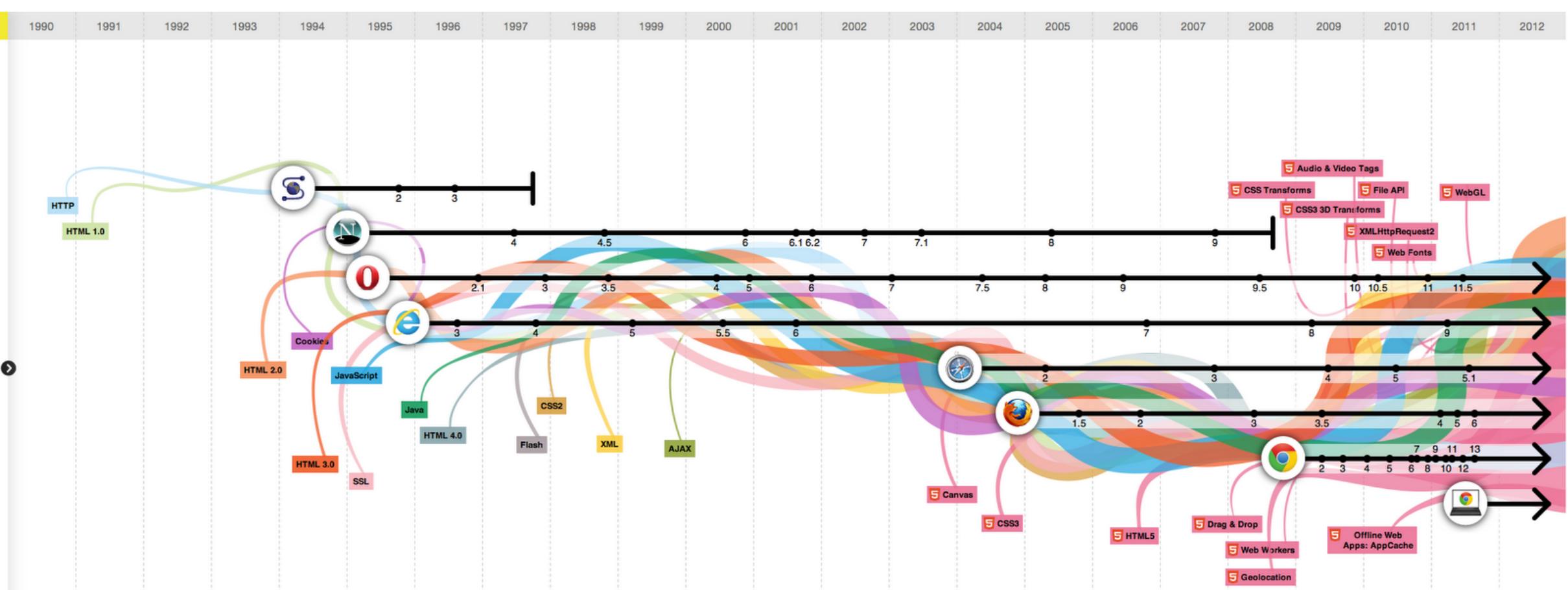
The Web



Human
Consumption

Machine
Consumption

Evolution of the Web



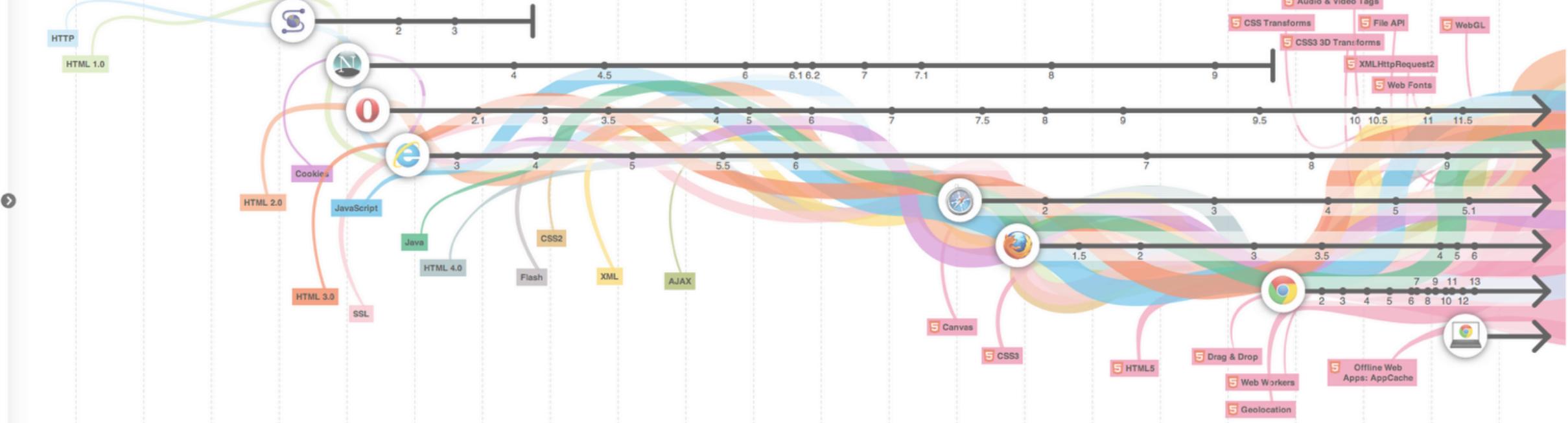
Slides by David Tarrant



<http://www.evolutionoftheweb.com/>

Evolution of the Web

Files → Static Pages → Dynamic Content → Applications

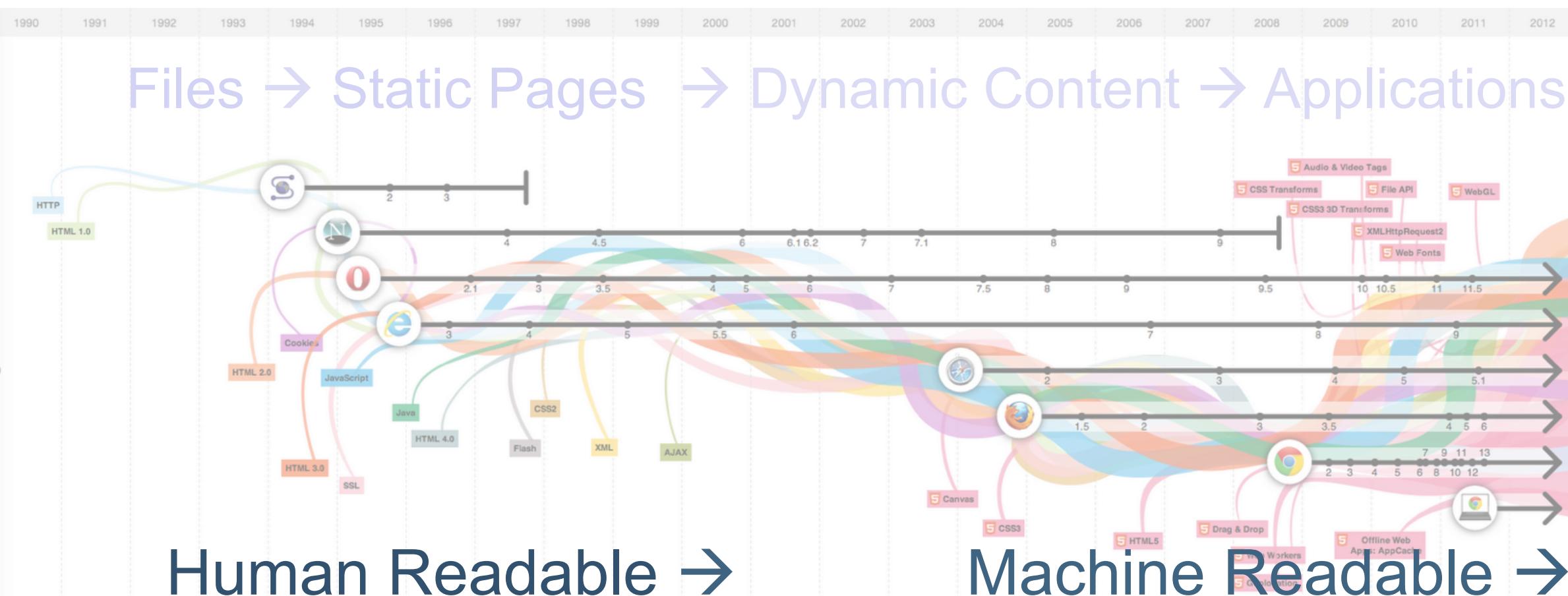


Slides by David Tarrant



<http://www.evolutionoftheweb.com/>

Evolution of the Web



Slides by David Tarrant



<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?



Slides by David Tarrant

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



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

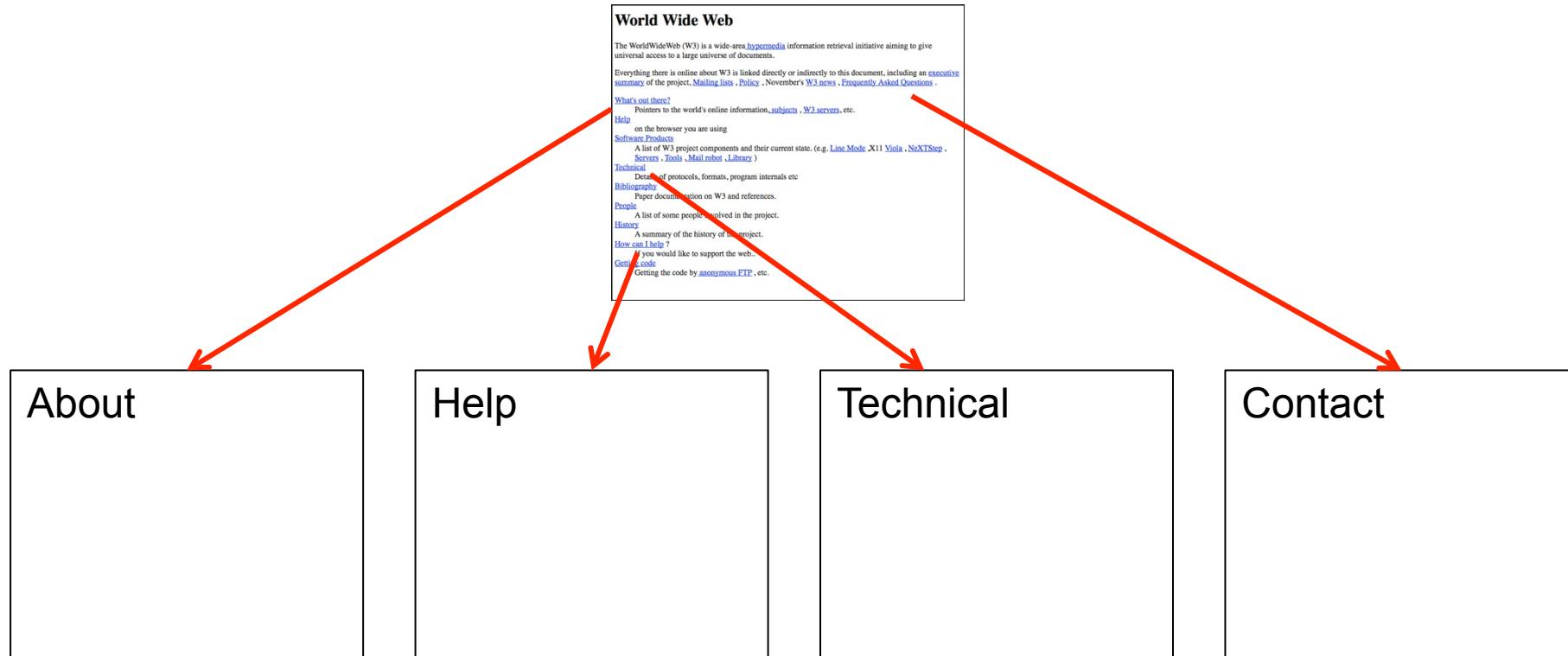
Getting the code by [anonymous FTP](#) , etc.

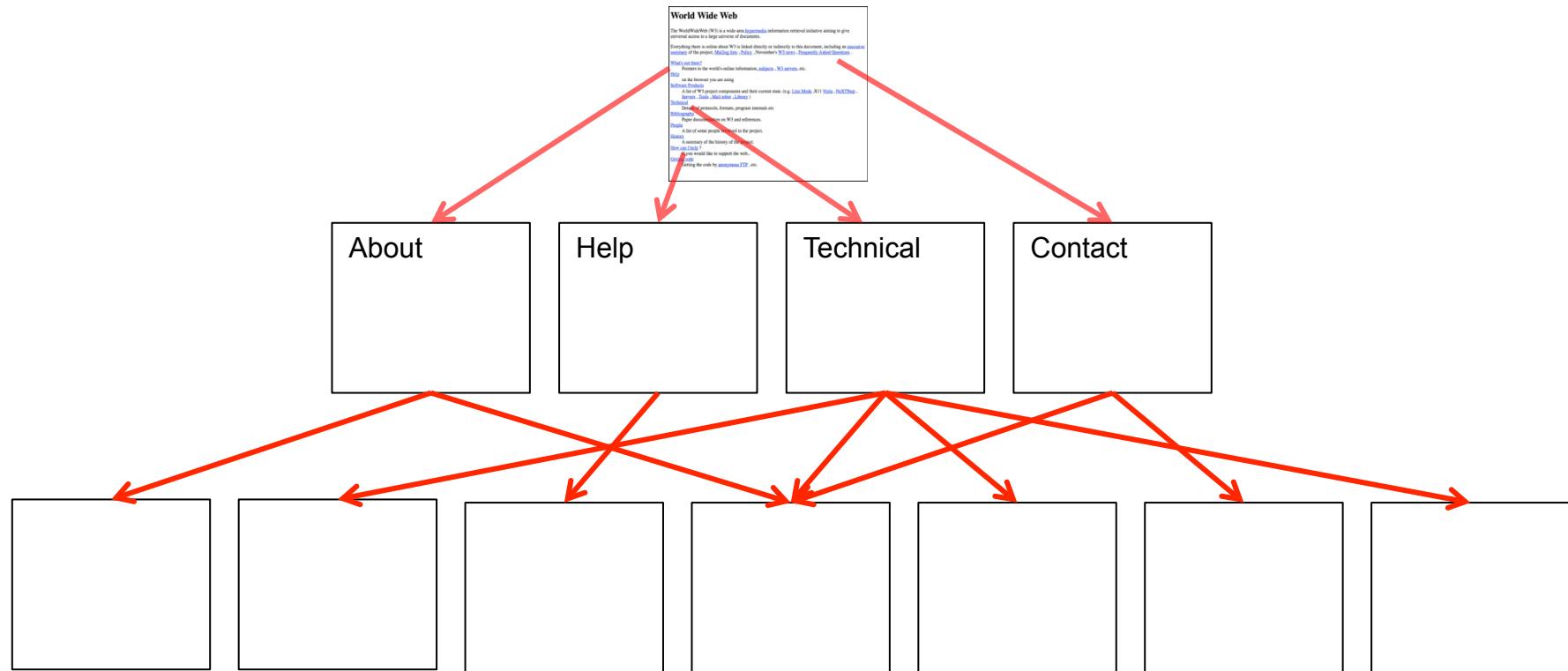
</list>

Markup Links & Tags

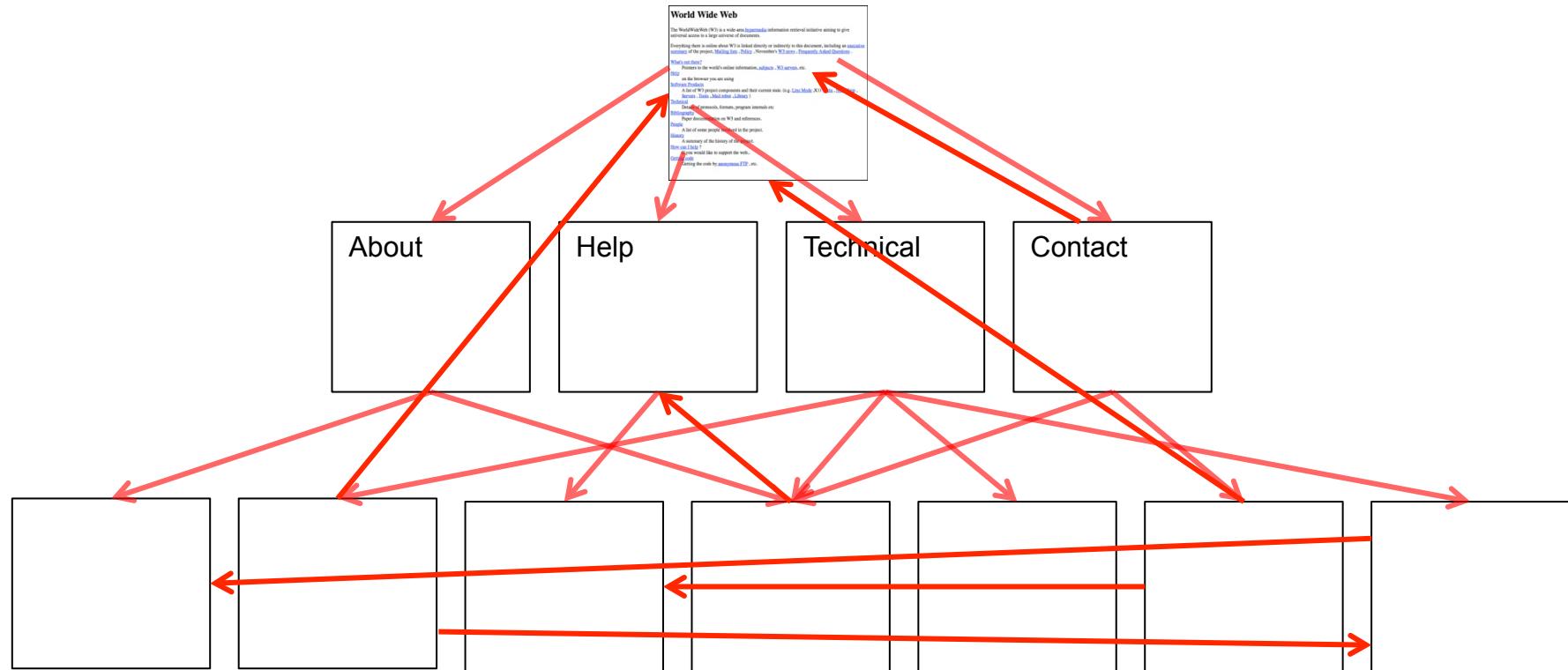


Nodes and Links



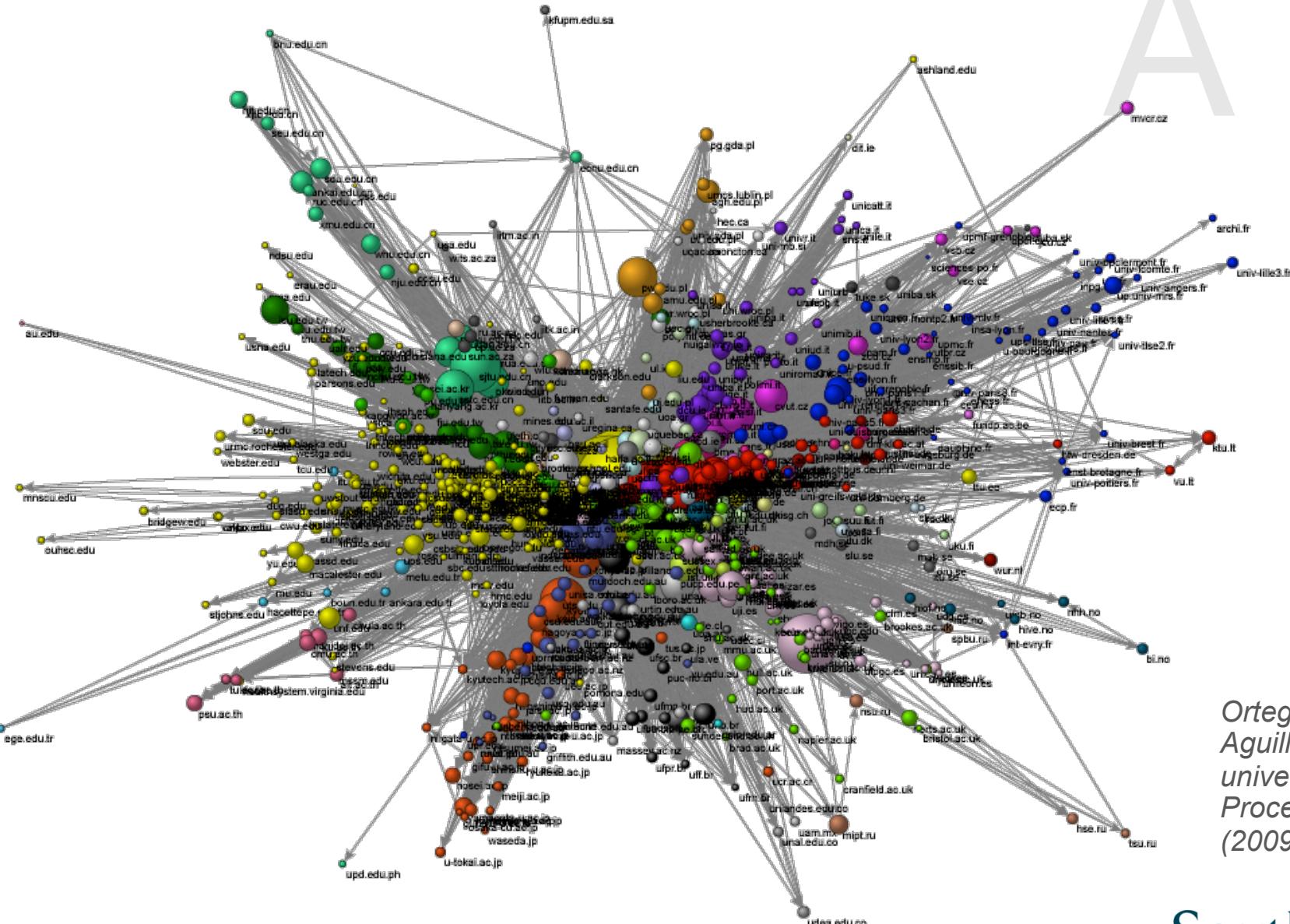


Nodes and Links





A Web



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

The Web



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>

</marquee>



Slides by David Tarrant

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>



Slides by David Tarrant

A Link

`Link to BBC News`



HTML 5

<menu>

<summary>

<figure>

<details>

<nav>

<legend>

<input>

<label>

<header>

<title>

<section>

<option>

<footer>

<a>

<blink>

<marquee>





HTML 5

html																			col	table		
head	span																		div	fieldset		
title	a																		form	body		
meta	rt	dfn	em	i	small	ins	s	br	p	blockquote	legend	optgroup	address	h1	section	colgroup	tr	h2	header	caption		
base	rp	abbr	time	b	strong	del	kbd	hr	ol	dl	label	option	datalist	h3	nav	menu	th	h4	article	command		
link	noscript	q	var	sub	mark	bdi	wbr	figcaption	ul	dt	input	output	keygen	h5	footer	summary	thead	h6	hgroup	details		
style	script	cite	samp	sup	ruby	bdo	code	figure	li	dd	textarea	button	progress	h6	hgroup	details	tfoot					
											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



Slides by David Tarrant

The Web



Data on the Web

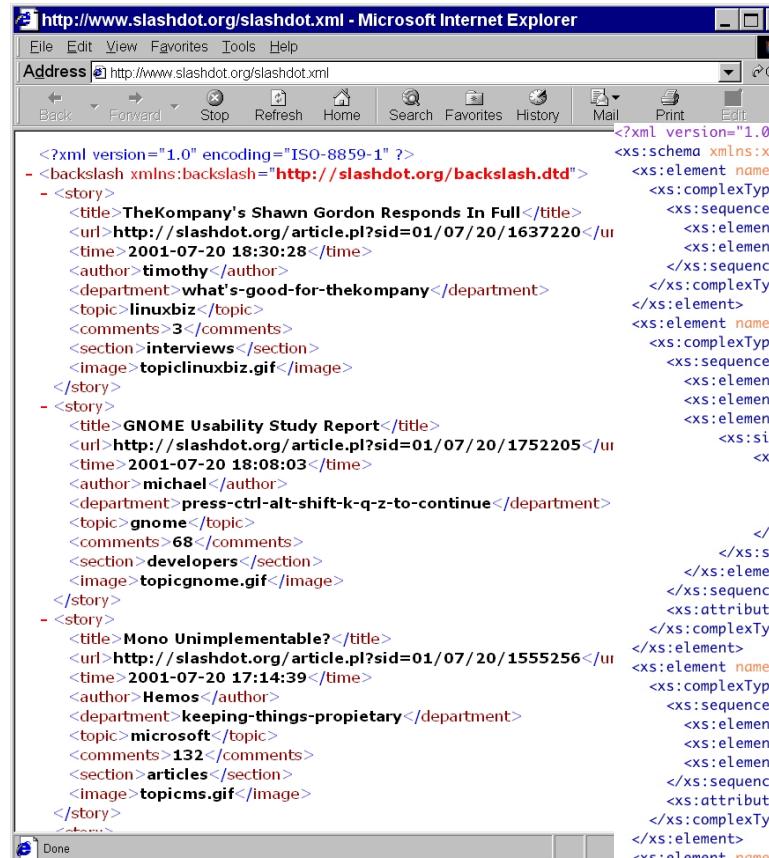
more markup...

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



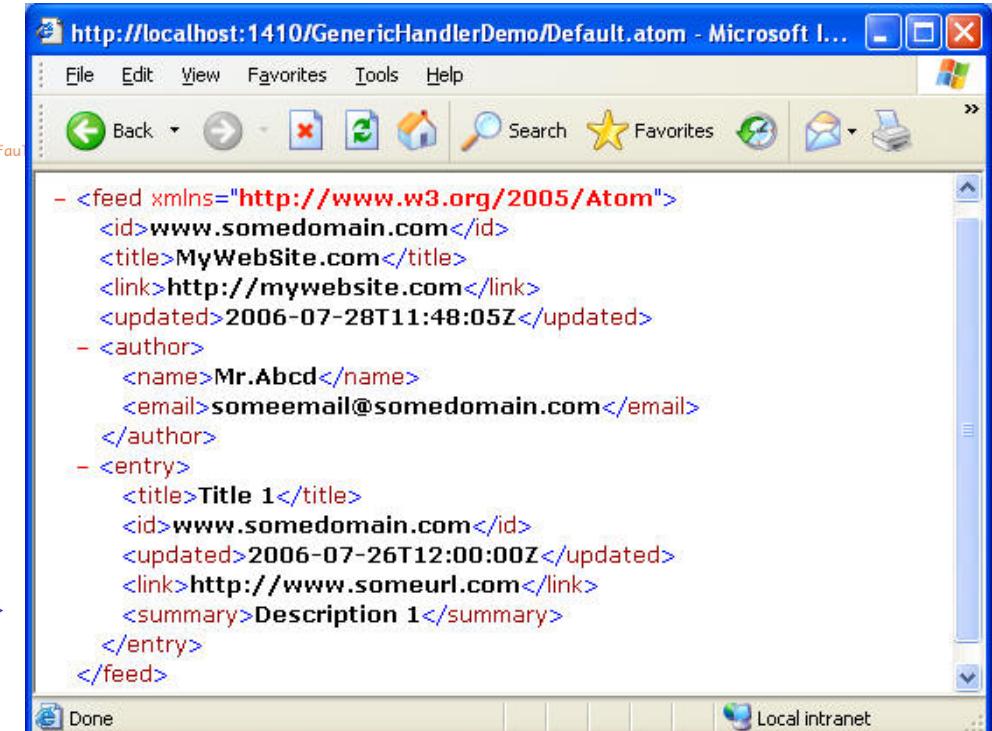
Slides by David Tarrant

Markup for anything



A screenshot of Microsoft Internet Explorer version 6.0 displaying XML content from <http://www.slashdot.org/slashdot.xml>. The XML structure includes a root element with attributes like `<?xml version="1.0" encoding="UTF-8"?>`, nested elements such as `<xs:schema>` and `<xs:element name="xypair">`, and various attributes and values throughout the document.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="xypair">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="xaxis"/>
        <xs:element ref="yaxis"/>
      </xs:sequence>
    </xs:complexType>
    <xs:element name="xaxis">
      <xs:complexType>
        <xs:sequence>
          <xs:element ref="property"/>
          <xs:element ref="value"/>
        </xs:sequence>
        <xs:attribute name="axistype" use="required" type="xs:NCName"/>
      </xs:complexType>
      <xs:element name="yaxis">
        <xs:complexType>
          <xs:sequence>
            <xs:element ref="property"/>
            <xs:element ref="value"/>
            <xs:element ref="unit"/>
          </xs:sequence>
          <xs:attribute name="axistype" use="required" type="xs:NCName"/>
        </xs:complexType>
      </xs:element>
    </xs:element>
  </xs:element>
</xs:schema>
```



A screenshot of Microsoft Internet Explorer version 6.0 displaying an Atom feed from <http://localhost:1410/GenericHandlerDemo/Default.atom>. The feed includes entries with titles, IDs, updated dates, authors, and summaries.

```
<feed xmlns="http://www.w3.org/2005/Atom">
  <id>www.somedomain.com</id>
  <title>MyWebSite.com</title>
  <link>http://mywebsite.com</link>
  <updated>2006-07-28T11:48:05Z</updated>
  <author>
    <name>Mr.Abcd</name>
    <email>someemail@somedomain.com</email>
  </author>
  <entry>
    <title>Title 1</title>
    <id>www.somedomain.com</id>
    <updated>2006-07-26T12:00:00Z</updated>
    <link>http://www.someurl.com</link>
    <summary>Description 1</summary>
  </entry>
</feed>
```



Data Exchange

eXtensible Markup Language

Forms the basis for hundreds of XML formats



Resource Description Framework

Enforces a structured relationship between elements



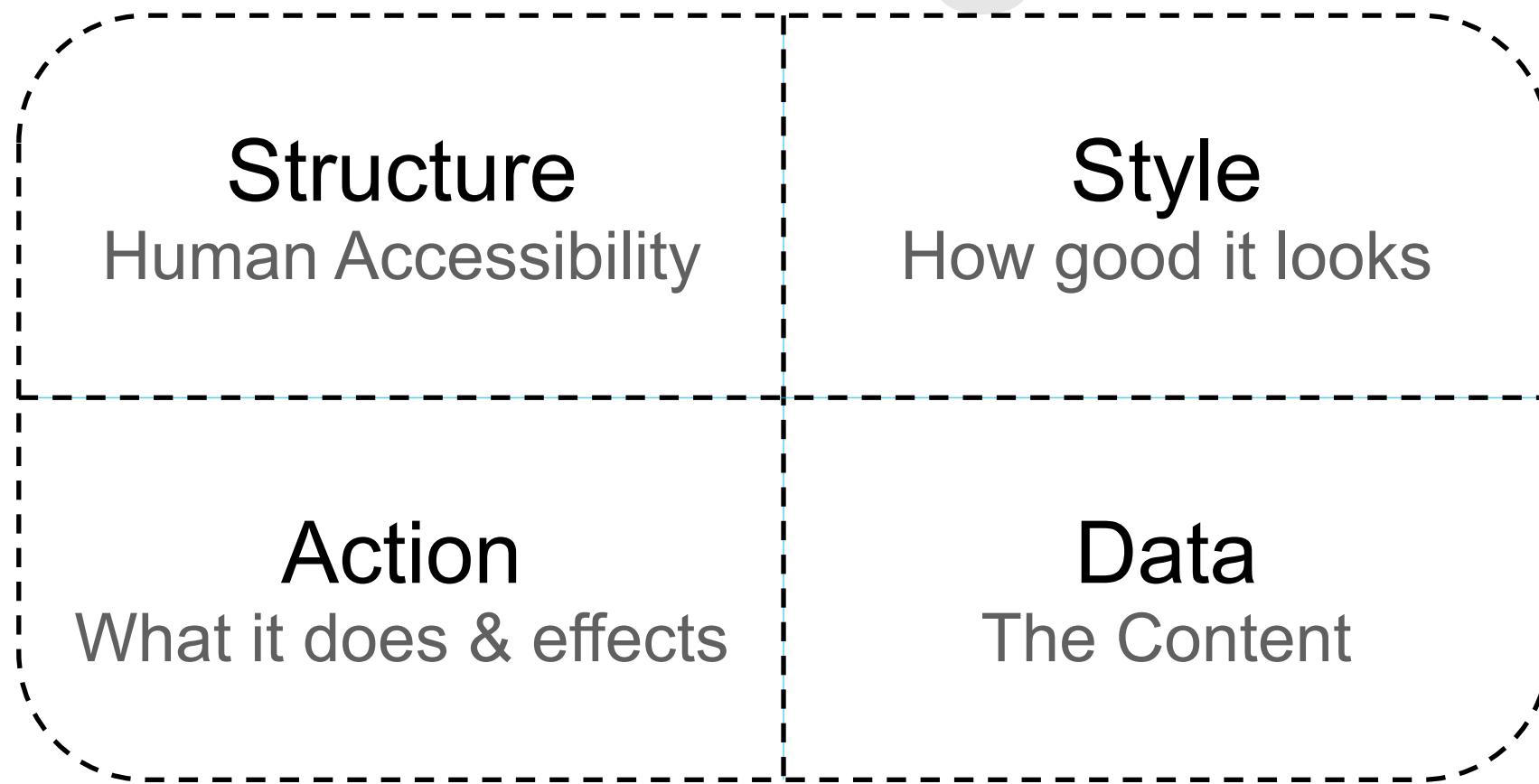
Exercise

Building a Web Page
that reads the BBC News Feed



Slides by David Tarrant

Building Blocks



The Language of the Web

HTML5



CSS3



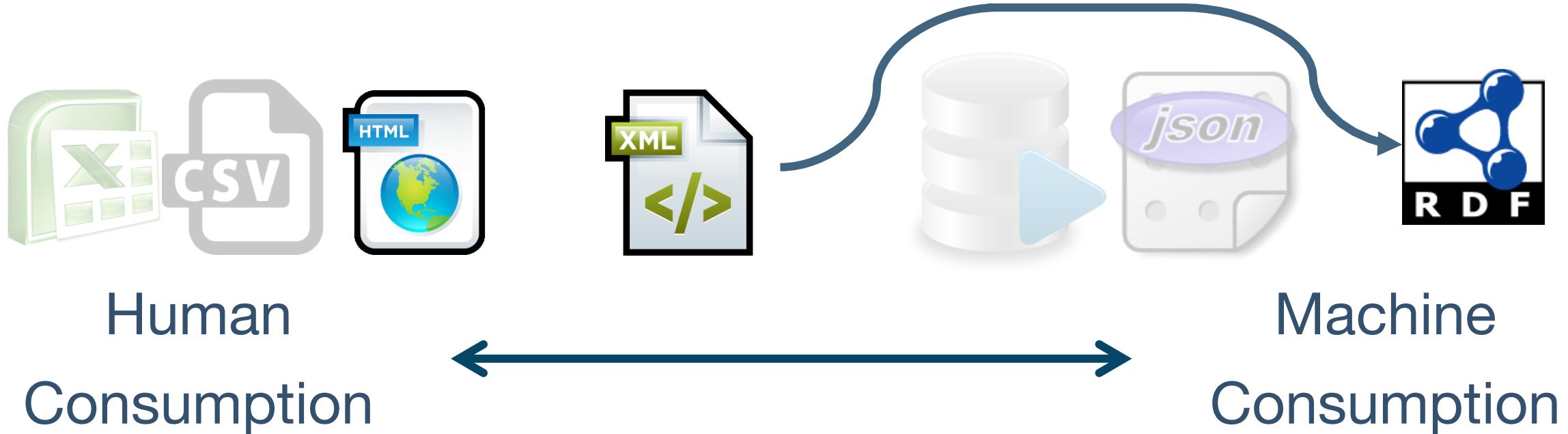
JAVASCRIPT



Building Blocks



The Web



Problems with XML

Unbelievably, XML is...
hard to learn for a machine to parse.

All the parser

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

XML has no

hard to use,
semantics



Introducing RDF

The Resource Description Framework 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		Postcode	Place Name
DOC	William Arthur Philip Louis Windsor	Duke of Cambridge		SW1A 1AA	Buckingham Palace
DSSOC	Catherine Elizabeth Mountbatten-Windsor	Duchess of Cambridge			

A diagram illustrating data normalization. It shows two tables: a main table and a detailed table. In the main table, the 'Residence' field contains the value 'BUCKINGHAM PALACE'. An arrow points from this value to the 'Name' field in the detailed table. Another arrow points from the 'PostCode' field in the main table to the 'Place Name' field in the detailed table.

Identifiers

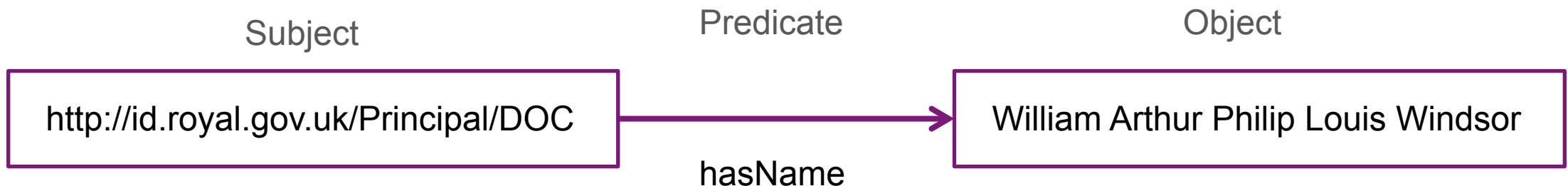
ContentID	ContentDate	Residence	Content	MRF	PostCode
22498	08/01/2012	BUCKINGHAM PALACE	The Duke and Duchess of Cambridge this evening attended the United Kingdom Film Premiere ...	DOC,DSSOC	SW1A 1AA

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

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

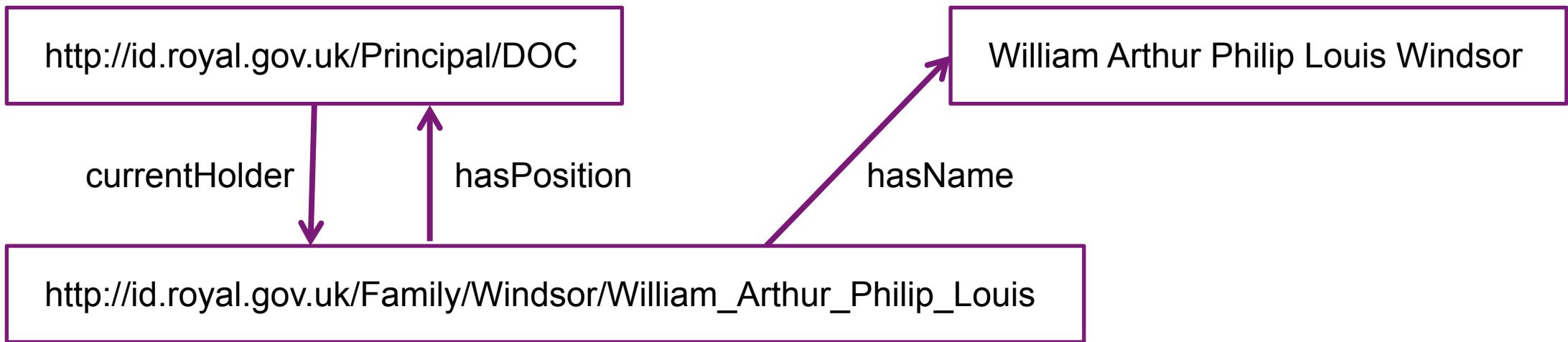
hasName

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



Slides by David Tarrant

Why RDF

It encourages use of URIs to identify things.

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



Slides by David Tarrant

Data Exchange

eXtensible Markup Language

Forms the basis for hundreds of XML formats



Resource Description Framework

Enforces a structured relationship between elements



Data Exchange

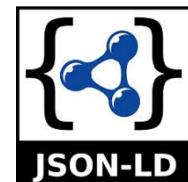
eXtensible Markup Language

Forms the basis for hundreds of XML formats



Resource Description Framework

Enforces a structured relationship between elements



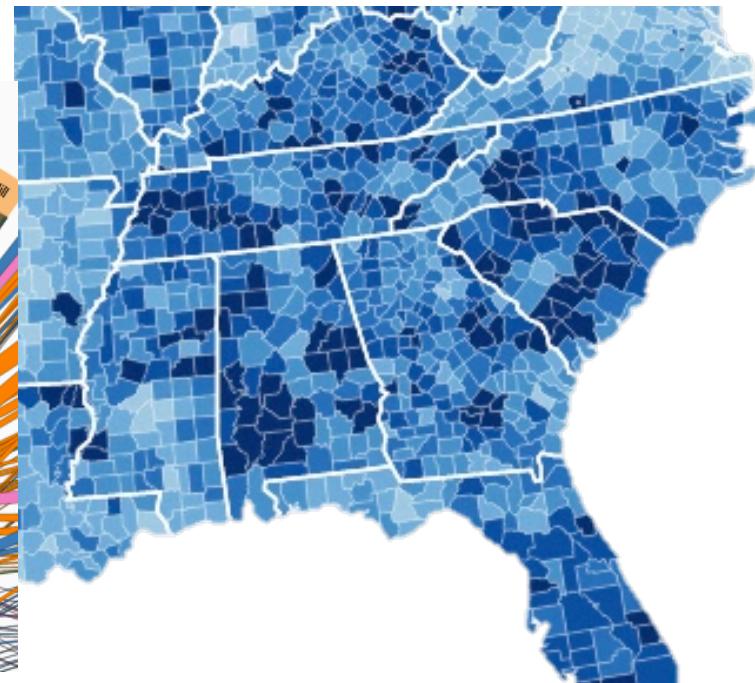
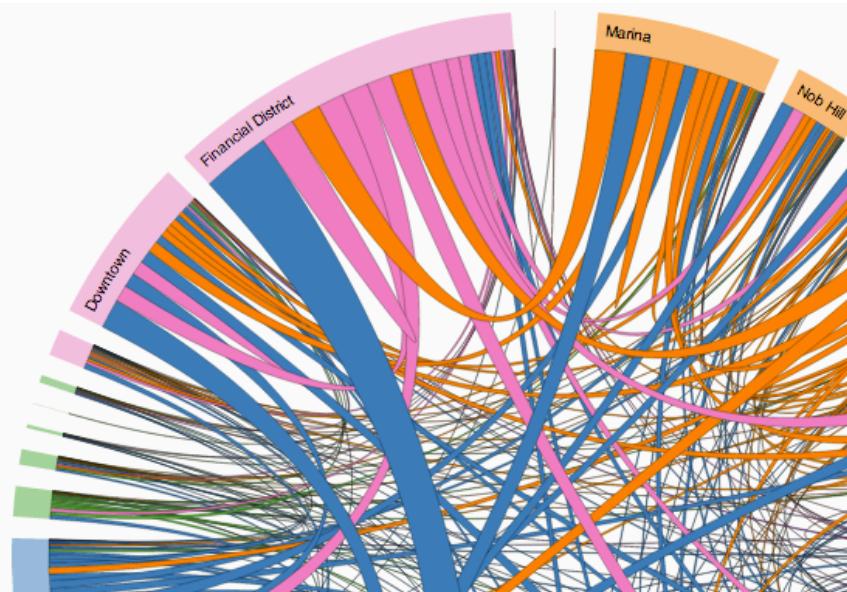
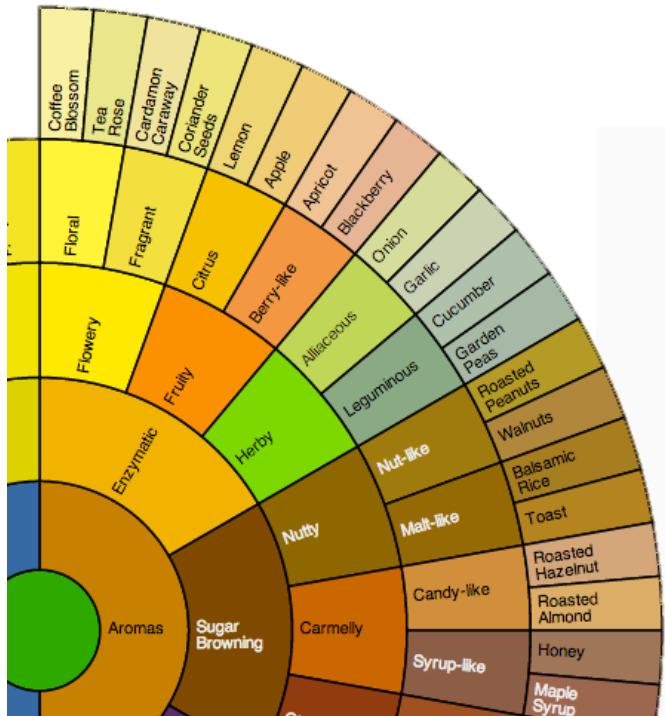
Data Formats



Data Formats



Using Data



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



Slides by David Tarrant

Recap

The Evolution of the Web

The Evolution of Data

Intro to Identifiers and Linking

Knowledge representation

Serialisations and data discovery



Slides by David Tarrant