The Semantic Web:

Using Future Internet Technologies for Corpora

Dominic Smith,
Sir Henry Thomas Scholar,
Department of Hispanic Studies,
University of Birmingham

In the beginning was the Web...

The original vision of the Web:

- Read/Write environment
- Users connect to a server and can view a document marked-up semantically in Hypertext Markup Language (HTML)
- The user's preferences define how the markup should display

Burners-Lee (1989), HTML 1.0 (1993)

```
<HTML>
 <TITLE>Page Title</TITLE>
 <H1>Main heading</H1>
 <H2>Sub-heading</H2>
 <P>Text in a paragraph, line-wrapping will sort itelf out.
 <P><A HREF="http://www.server.tld">Linked text</A> contained
    within a paragraph
 <UL>
  <LI>An item in a bulleted list
 </UL>
 <P><IMG SRC="icon.gif">
</HTML>
```

In the beginning was the Web...

THE UNIVERSITY
OF BIRMINGHAM

Welcome to the University of Birmingham

From this page you can find information about all aspects of the University.

You can also use eXcite to search for documents on **www.bham.ac.uk** containing one or more keywords

There is also a text only version of this page

Birmingham University website, January 1997

But no mechanism for embedding tables of data. <TABLE> is introduced in HTML 3.2 in 1996.

Also adds <B> (bold) <I> (italic) <U> (underline)

For the first time, tags representing format (not semantics) is introduced.

```
<TABLE>
 <TR>
  <TD>Data Cell
  <TD>123
 <TR>
  <TD><B>Another Cell</B>
  <TD>456
</TABLE>
```

- Designers then realise that they can use <TABLE> to position text on the page which isn't really a data table at all.
- Further departure from the original semantic model when, after public pressure, HTML 4 added <FONT> in December 1997.
- The rise of the 'navigation bar' means that linked text is no longer part necessarily part of the main content

The situation in late 1998

- Reach of the WWW increased dramatically after Microsoft Internet Explorer 3 shipped with Windows 95 and shortly after AOL and CompuServe combined their networks with the WWW
- Major browsers (IE4, Netscape 6) are not read/write
- Tendency for the WWW to be seen as a publication rather than dissemination medium
- Number of pages on the WWW increasing rapidly. Manual indexing libraries (eg. The Virtual Library) can't keep up. Major search engines (Yahoo!, Altavista) using automated indexing.
- Realisation that automated indexing would be more efficient had the semantic model been adhered to.
- Becoming evident that HTML is not suitable for every type of document.

Meanwhile...

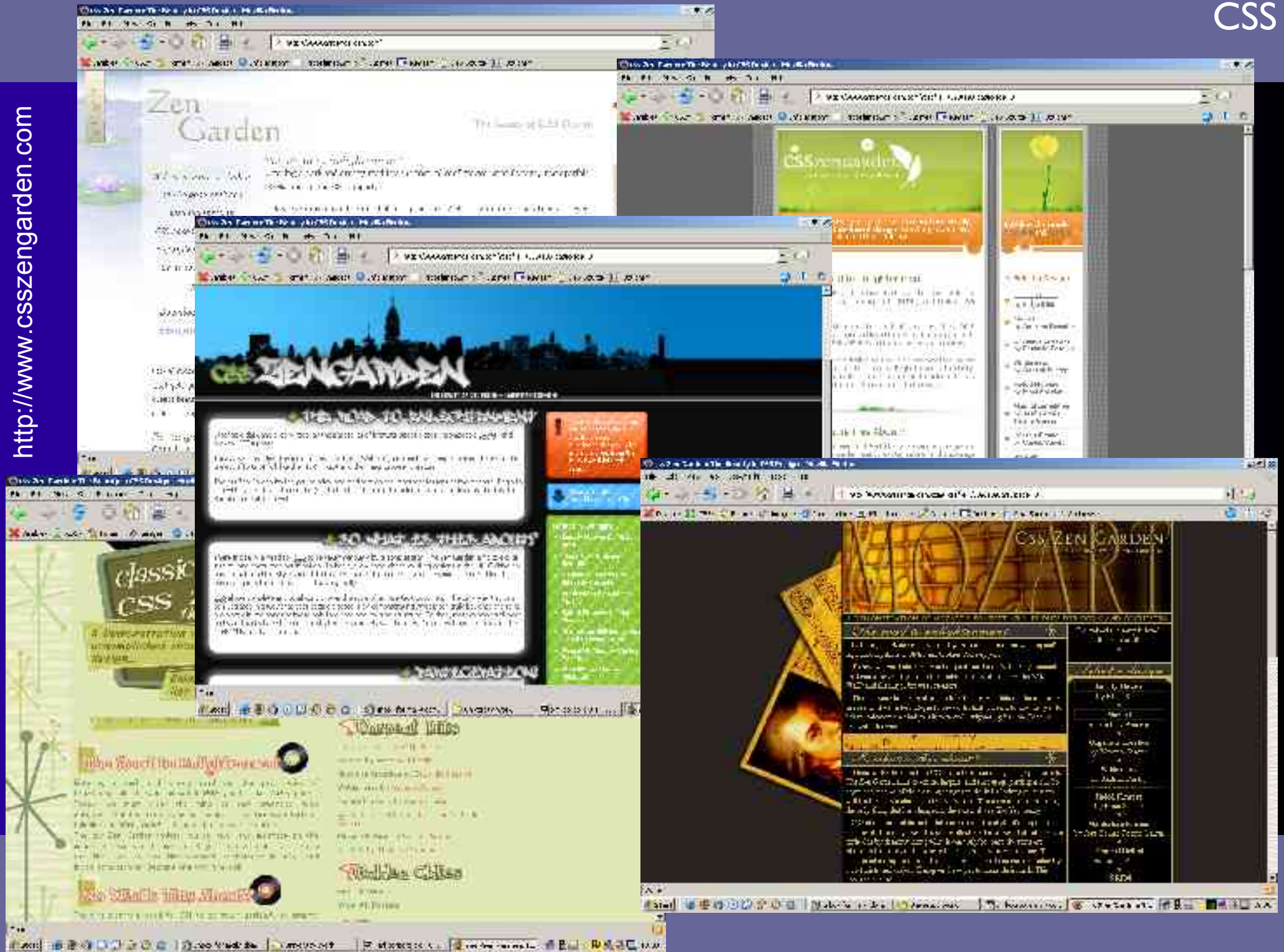
- Work progressing on CSS (Cascading Stylesheets)
- CSS allows a separate file to control formatting. Makes it easier to ensure that corporate pages have a consistent look, seems to be a way to get rid of font etc. markup in HTML

```
<HTML>
 <BODY>
 <H1>Heading</H1>
...
```

```
h1 {font-family:Arial,sans-serif;
    font-size:1.5em; color:black;}
```

- Accessibility concerns have been pushing this: CSS can be redefined in the browser, or sites can offer many CSS options
- Allows page divisions to be positioned exactly
- Also low-bandwidth

http://www.csszengarden.com



Meanwhile...

- Work progressing on a means to represent databases on the web using a language called XML (eXtensible Markup Language)
- Extensible represents the fact that any user can define tags for their own purposes

```
<?xml version=”1.0”?>
<person>
 <firstname>Joe</firstname>
 <surname>Bloggs</surname>
</person>
<person>
 <firstname>John</firstname>
 <surname>Doe</surname>
</person>
```

The birth of XHTML

- XML can be used for any document type
- Combined with CSS promises a return to the vision of a *Semantic Web.*
- HTML was slightly incompatible with XML so, in January 1999, HTML 4 was officially superseded by XHTML 1.0.
- XHTML 1.0 deprecated formatting markup in favour of CSS

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en">
<head>
  <title>Page Title</title>
  <link rel="stylesheet" type="text/css" href="stylesheet.css" />
</head>
<body>
 <h1>Heading</h1>
 <p><img src=”icon.gif” alt=”Icon” /> Text in a paragraph. </p>
</body>
</html>
```

- In 1999, Netscape proposed RSS (Rich Site Summary) – a file that could be put on every website to help search engines by summarising all pages on the site. The idea never took off.
- But news websites used this XML language to distribute newsfeeds to other websites and end-users; RSS became rebaptised 'Really Simple Syndication'

```
<item>
  <title>Asbo powers target 'enviro-crime'</title>
  <description>Anti-social behaviour orders are to be used to
tackle environmental crime such as fly-tipping and
graffiti.</description>
  <link>http://news.bbc.co.uk/go/rss/-1/hi/uk/4545542.stm</link>
  <guid isPermaLink="false">
http://news.bbc.co.uk/1/hi/uk/4545542.stm</guid>
  <pubDate>Tue, 20 Dec 2005 15:15:23 GMT</pubDate>
  <category>UK</category>
</item>
```

- The 'extensible' part of XML means that individuals can define their own tags and use them as they please.
- You can embed other peoples' user-defined tags in a standard document by referring to 'namespaces'

```
<rss xmlns="http://purl.org/rss/1.0/"
  xmlns:ev="http://purl.org/rss/1.0/modules/event/"
  xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
[...]
<item>
 <title>Going to Cyprus</title>
 <description>I booked my holiday to Paphos in April today</description>
 <link>http://www.myserver.com/blog/2005/12/25/Going_to_Cyprus.html</link>
 <ev:type>Holiday</ev:type>
 <ev:location>Paphos, Cyprus
   <geo:Point>
    <geo:lat>34.8</geo:lat>  <geo:long>32.4</geo:long>
   </geo:Point>
 </ev:location>
 <ev:startdate>2006-04-03T09:00:00Z</ev:startdate>
 <ev:enddate>2006-04-09T22:00:00Z</ev:enddate>
</item>
```

- XML also allows URIs (Uniform Resource Identifiers) to reference potentially any count noun, as well as URLs (Uniform Resource Locators, aka internet addresses).

```
<html xmlns:bk='urn:loc.gov:books'>

[...]

<p>As discussed by <bk:author>Prof. Baggins</bk:author> in her seminal work
<a href="uri:isbn:0123456789X"><bk:title>Why academia is
important<bk:title></a> [...] </p>

</html>
```

- Or perhaps...????

```
<p>As discussed by <a href="uri:people:uk:idcard:0123456789">
<bk:author>Prof. Baggins</bk:author></a> in her seminal work
<a href="uri:isbn:0123456789X"><bk:title>Why academia is
important<bk:title></a> [...] </p>
```

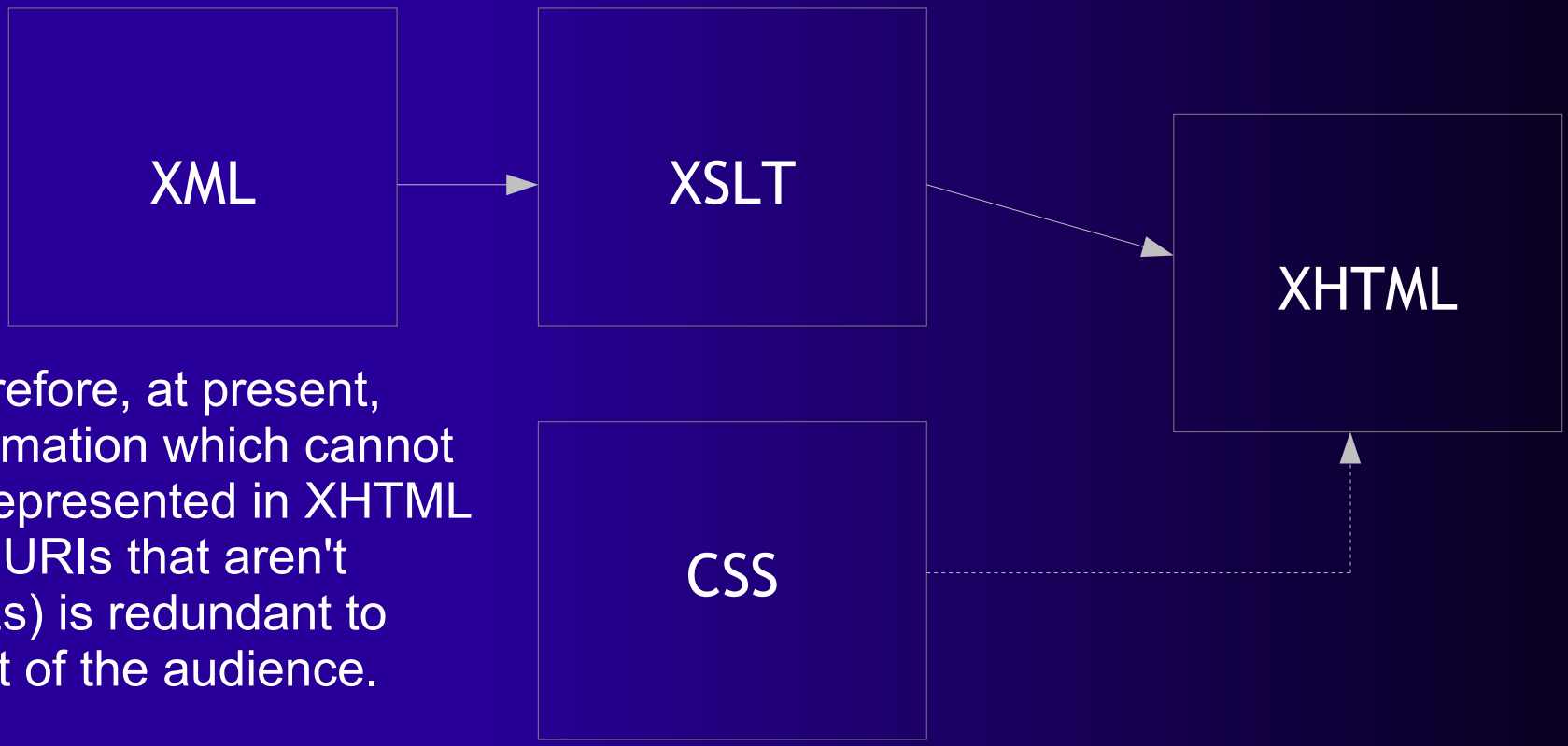
- Other popular XML-based formats:
- MathML
- MusicXML
- RDF
- SMIL
- TPEG
- TV-Anytime
- [...]

```
<tpeg_message>
<originator country="UK" originator_name="BBC Travel News"/>
<summary xml:lang="en">M60 Greater Manchester - Contraflow both w
<road_traffic_message message_id="80972" message_generation_time=
<network_conditions><position position="&rtm1C_0;"/><roadworks ro
<location_coordinates location_type="&loc1_3;">
<WGS84 latitude="53.413811" longitude="-2.265107" />
<location_descriptor descriptor_type="&loc3_7;" descriptor="M60;"
<location_descriptor descriptor_type="&loc3_24;" descriptor="Grea
<location_descriptor descriptor_type="&loc3_1C;" descriptor="Prin
<location_descriptor descriptor_type="&loc3_32;" descriptor="M60;
<location_descriptor descriptor_type="&loc3_8;" descriptor="A5103
<WGS84 latitude="53.440896" longitude="-2.335596" />
<location_descriptor descriptor_type="&loc3_7;" descriptor="M60;"
<location_descriptor descriptor_type="&loc3_24;" descriptor="GREA
<location_descriptor descriptor_type="&loc3_1C;" descriptor="Carr
<location_descriptor descriptor_type="&loc3_32;" descriptor="M60;
<location_descriptor descriptor_type="&loc3_8;" descriptor="A6144
<direction direction_type="&loc2_2;"/>
</location_coordinates>
</location_container>
</road_traffic_message>
</tpeg_message>
```


Ice cream made in festive flavour
A Lancashire company says its Christmas pudding ice cream is a success with its customers.
published at 09:30 on the 22nd December
in Feed: BBC News | England | Lancashire | UK Edition
North Atlantic Ocean
North Sea
English Channel
St George's Channel
Irish Sea
Donegal Bay
Belfast
Dublin
Cardiff
Bristol
Plymouth
Exeter
Southampton
Brighton
Norwich
Liverpool
Blackpool
Sheffield
Manchester
Google
Map data ©2005 Tele Atlas


Google Earth
File Edit View Add Tools Help
Fly To
Local Search
Directions
Search
Places
Layers
Wuhan
Shanghai
Chengdu
Xian
AS-006
AS-131
AS-075
AS-132
AS-139
AS-110
AS-094
AS-185
AS-143
AS-116
AS-162
South China Sea
Image © 2006 MDA EarthSat
Google
Start
Google Earth

- Problem 1: Nobody maintains a register of namespaces, so duplicates occur
- Problem 2: If anyone can define tags, how does a browser know how to display them?



Therefore, at present, information which cannot be represented in XHTML (eg. URIs that aren't URLs) is redundant to most of the audience.

- XML returns to the notion of *The Semantic Web*
- Allows powerful APIs (ISBN searches...)
- When combined with the rise in blogging could be seen as a return to the notion of the read/write web

but...

- XML cannot be efficiently written by editing programs
- Most editing programs even don't even bother with CSS
- The vast majority of webpages written today are still in HTML 4
- How can we train secretaries to learn about semantic markup if it requires hours looking up namespaces which will mark information that the reader can't see on the page?

Why should corpus linguists care?

- *Semantic Web* makes mark-up useful
- Parsers are readily available for most programming langauges (eg XML::Parser for Perl from CPAN)
- Namespaces mean that similar types of information can be compared wherever they are on the web
- Possibility of a huge, useful corpus with building automated by web-crawlers (eg. GoogleBot)

The Chávez Bot

- My PhD involves a corpus analysis of newspaper articles in Spanish about President Hugo Chávez of Venezuela
- Looking for intertextual reference, shifts in viewpoints
- My MPhil suggested that as opinions change over time, so does strength of collocation

- Needed a DIACHRONIC, CONTINUOUSLY COLLECTED corpus, in which the articles come from MANY SOURCES and are GROUPED BY THEME

The Chávez Bot

- ChávezBot is actually two Perl scripts set on a Chrontab:
- ABN-RSS.pl (hourly) – collects RSS feed from Agencía Bolívarania de Noticias and adds new stories to database
- chavwatch.pl (every six hours):

- 1) Loads Google News Spanish with search term 'Chávez'
- 2) Goes to each external URL
- 3) Collects text, removes tags and linked text
- 4) Saves plain-text copy
- 5) Compares wordlist to all previously-collected texts and extracts keywords
- 6) Compares keywords to ABN feeds and links the article if appropriate
- 7) For each keyword: concordance and patterns from all previous
- 8) Uploads report to www.domsmith.co.uk/chavwatch/

The Chávez Bot

Concordances and Picture for keyword titulares

```
       el país en los titulares de los periódicos en
      ese país en los titulares de los diarios. el
       el país en los titulares de los periódicos en
  todo lo que quiere. titulares en la prensa mundial,
  todo lo que quiere. titulares en la prensa mundial,
       el país en los titulares de los periódicos en
      ese país en los titulares de los diarios. el
       el país en los titulares de los periódicos en
```

país (5)	en (5)	los (6)	titulares	de (6)	es (6)

New text

... [illegible]
[illegible] Noticias Imprimir [illegible]

Property	Value
Source URI	http://[illegible]
Tokens	642
Types	347
TTR	[illegible]

Keyword	Freq (%)	Ref Freq (%)	Keyness
titulares	1.090	[illegible]	[illegible]

Active threads

- Heinz Dieterich: Evo Morales dignifica al indígena tras su triunfo en Bolivia 2005-12-12 11:55:02
  http://www.prensalatina.com.mx/Art [illegible]

The Chávez Bot

- This would be so much more accurate if newspapers published in a universal XML standard!
- Martha Chávez and other family members – need for URIs?
- Possibilties for mash-ups? (Plot geographically how ideas spread over time...)
- Reliant on Google (I don't have their resources!) – so may be delayed: not ideal if looking for intertextuality

- Need comparison with ABN feeds to be collocate not keyword-based
- Using Wikipedia to do this – linked to definitions
- Changes in collocational strength over time – ie. Multiple pictures
- No reference corpus, only what has come before, so it gradually gets more accurate over time
- No useful data yet... See you after the summer vacation?!

References and Credits

- HTML 2.0 http://www.w3.org/MarkUp/html-spec/html-spec_toc.html
- HTML 3.2 http://www.w3.org/TR/REC-html32
- HTML 4.01 http://www.w3.org/TR/REC-html40/
- CSS http://www.w3.org/Style/CSS/
- XHTML 1.1 http://www.w3.org/TR/xhtml11/
- XHTML 2.0 (Working Draft) http://www.w3.org/TR/xhtml2/
- XML http://www.w3.org/XML/
- XML Namespaces http://www.w3.org/TR/REC-xml-names/
- URIs, URLs, URNs http://www.w3.org/Addressing/
- XSL http://www.w3.org/Style/XSL/
- Semantic Web http://www.w3.org/2001/sw/

- BBC Backstage http://backstage.bbc.co.uk
- BBC News Maps http://www.benedictoneill.com/content/newsmap/

- ChavWatch http://www.domsmith.co.uk/chavwatch/

Thank You

Comments / Questions?

Copies of this presentation and abstract:
www.domsmith.co.uk/phd

SOME RIGHTS RESERVED

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.0 England & Wales Licence. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-sa/2.0/uk/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.