The World Wide Web
Virtual Library
The WWWVL : Experiences in Building the Semantic Web

Michael Chapman
WWWVL Secretary

Paper presented to TMRA 2006
Wednesday, October 11th 2006 — Leipzig
abstract: The WWW Virtual Library is the oldest catalogue on the Web, having been created by Tim Berners-Lee when he invented the Web. The VL central catalogue now has a six-figure number of users each year seeking information contained in its 200+ individual VLs. In early 2005 the central catalogue was moved over to a topic map (with a marked increase in its effectiveness). Since then various experiments have taken place to allow deeper access to the data contained in the highly diverse and individual member VLs (a classic meta-indexing challenge). These experiments have initially centred on ‘easy’ PSI topics but are now moving to a broader indexation. Description of these is set against statistics of the resulting usages. With such a large, well-established, user-base the WWWVL is in a (possibly) unique position to experiment with applying topic maps to a real life situation on the Web and and to make its own contribution to building the semantic web of the future.
1. Introduction

Overview.html

Information by Subject

See also arrangements by organization or by service type. Mail www-request@info.com.ch if you know of online information not in these lists.

Aeronautics
- Mailing list archive index
- Abstract Indexes (down?)
Bio Sciences
- See separate list
Computing
- See Networking, Jargon, newsgroups, Software Technology, Languages, Algorithms
Geography
- CIA World Fact Book, India, Miscellaneous information, Thai-Yunnan, Davis collection
Law
- US Copyright law
Libraries
- Few libraries currently have servers - you have to log on to them. But you can find out how with the library index
- Project Gutenberg, two classic books a month. See their explanations, the index and newsletter, books published in 1991, 1992, and reserved for the USA
 Humanities
- BNCR classical reviews, Poetry, SciFi reviews, See also electronic journals
Mathematics
- CIRM library (French)
Meteorology
- US weather, state by state. Also WARS weather (around MIT)
Music
- MIDI interfacing, Song lyrics (apparently disabled for copyright reasons)
Physics
- High Energy Physics, Astrophysics abstracts
Politics & Economics
- US politics, includes campaign 1992
Reference
- Roget's Thesaurus, Experimental English dictionary
Religion
- The Bible (King James version), The Book of Mormon, The Holy Qur'an
Social Sciences
- Cooptec papers archive
November 3rd 1992:

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

Done.
“... there would be the occasional ‘Hey, I’ve just set up a server, and it’s dead cool. Here’s the address.’

With each new message I would enter in info.cern.ch a hyper-text link to the new web site . . .”

“In July and August 1991 there were from ten to one hundred ‘hits’ (pages viewed) a day.”

Information by Subject

See also arrangements by organization or by service type. Mail www-request@info.cern.ch if you know of online information not in these lists...

Aeronautics
  Mailing list archive index.

Astronomy and Astrophysics
  Abstract Indexes (down?)

Bio Sciences
  See separate list.

Computing
  See Networking, large, newsgroups, Software Technology, Languages, Algorithms.

Geography
  CIA World Fact Book, India; Miscellaneous Information, Thai-Yunnan; Davis collection.

Law
  US Copyright law.

Libraries
  Few libraries currently have servers; you have to log on to them. But you can find out how with Art
  St.George's list of library systems, about "Library in the internet resource guide", and the hyrnet index.

Literature
  Project Gutenberg: two classic books a month. See their explanations, the index and newsletter, books
  published in 1991, 1992, and reserved for the USA.

Humanities
  BMCR classical reviews, Poetry, Sci-fi reviews. See also electronic journals.

Mathematics
  CIRM library (French)

Meteorology
  US weather, state by state. Also WAIS weather (around MIT:).

Music
  MIDI interfacing, Song lyrics (apparently disabled for copyright reasons)

Physics
  High Energy Physics, Astrophysics abstracts.

Politics & Economics

Reference
  Roget's Thesaurus, Experimental English dictionary.

Religion
  The Bible (King James version), The Book of Mormon, The Holy Quran

Social Sciences
  Coombs papers archive.
August 1992: There are about 20 Web servers in existence.

Ciolek: Global Networking: a Timeline.

VL (with logo) first known screenshot 1993 ... next page ...

(by October 1993 — 200+ servers ...)

Ciolek: Global Networking: a Timeline.
By now the Web consisted of a small number of servers, with info.cern.ch the most interconnected with the rest. It carried a list of servers, which to a degree could coordinate people who were putting information on the Web. When the list became larger, it needed to be organized, so I arranged it in two lists, by geography and by subject matter. As more servers arrived, it was exciting to see how the subjects filled out. Arthur Secret, another student, joined me for a time and set up the lists into what we called the Virtual Library, with a tree structure that allowed people to find things.
The same concept from inception until now.

Like the Web itself, the VL is simple and logical.

Like the Web itself, it also has at its foundation the free distribution of free information.

“It is futile to discuss whether the Index or the table of Contents is the more useful in a book.”
Interfaces in various languages.
### The WWW Virtual Library
### Asian Studies

**Asian Studies**
Over 10 years of continuous service to the online community; Provision of expert information services, by the experts, for the experts. This research tool is produced by the Internet Publications Bureau, Research School of Pacific and Asian Studies at the Australian National University (ANU), Canberra.

*This resource in English is indexed under: Asian Studies.*

**Asian Continent**
This document keeps track of leading information facilities pertaining to the whole of the Asian continent.

*This resource in English is indexed under: Asian Studies.*

**Asian Studies Electronic Mailing Lists**
Electronic mailing lists associated with the extensive Asian Studies Virtual Library network.

*This resource in English is indexed under: Asian Studies.*
Online Burma/Myanmar Library

Classified and annotated links to more than 11,000 full text documents on Burma/Myanmar

Main Library (by subject): Activism; Archaeology; Aung San Suu Kyi; Children; Civil Society; Civil war; Dialogue/Transitional Justice; Drugs; Economy; Education; Environment; Ethnic and minority groups in Burma; Foreign Relations; Funding; Geography; Health; History; Human Rights; International Assistance to Burma; Internal Displacement; International Labour Organization; Labour Issues; Languages of Burma; Law; Migrants; Military (Tatmadaw); Palaeontology; Politics and Government; Refugee Social policies; Society and Culture; SPDC-related sites; Sustainability; The United Nations System; Tourism; Trafficking; Women.

Reading Room (by source): Bibliographies, universities, institutes, libraries, other research tools; Burma news (archive); Burma news (breaking); Burma news (current); Burmese Community sites; Burmese-language news sources; Conference proceedings; Dissuasion Group; Electronic Newspapers; Films, videos, photographs and other images of Burma.

Introduction | Structural Overview | What's New | Advanced Search | Feedback

Search

Done.
1.0.1. VLsearch

VLsearch --- VL

Introduction
Materials and Methods
The three phases
Conclusions
World Health Organisation:

<table>
<thead>
<tr>
<th>Search Type</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple search</td>
<td>179</td>
</tr>
<tr>
<td>“string”</td>
<td>10</td>
</tr>
<tr>
<td>without synonyms</td>
<td>113</td>
</tr>
<tr>
<td>“string” without synonyms</td>
<td>4</td>
</tr>
</tbody>
</table>
WWWVL
2. Materials and Methods

- Webserver – as most – logs visits
- Pages visited are static HTML
2.1. Webserver log

Logs: page visited, date/time of request, IP/domain of requester

Analyze using `awstats`

---

1http://awstats.sourceforge.net/
## Monthly history

<table>
<thead>
<tr>
<th>Month</th>
<th>Unique visitors</th>
<th>Number of visits</th>
<th>Pages</th>
<th>Hits</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 2006</td>
<td>2370</td>
<td>3443</td>
<td>19516</td>
<td>19516</td>
<td>50.42 MB</td>
</tr>
<tr>
<td>May 2006</td>
<td>2432</td>
<td>3703</td>
<td>13207</td>
<td>13207</td>
<td>36.50 MB</td>
</tr>
<tr>
<td>Jun 2006</td>
<td>2068</td>
<td>3145</td>
<td>11628</td>
<td>11628</td>
<td>32.15 MB</td>
</tr>
<tr>
<td>Jul 2006</td>
<td>1693</td>
<td>2495</td>
<td>7871</td>
<td>7871</td>
<td>21.99 MB</td>
</tr>
<tr>
<td>Aug 2006</td>
<td>1559</td>
<td>2013</td>
<td>16509</td>
<td>16509</td>
<td>48.06 MB</td>
</tr>
<tr>
<td>Sep 2006</td>
<td>1997</td>
<td>3160</td>
<td>14708</td>
<td>14708</td>
<td>31.02 MB</td>
</tr>
<tr>
<td>Domains/Countries</td>
<td>Pages</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>30662</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Network</td>
<td>20136</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>11014</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>7134</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>5308</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>1670</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>1634</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canada</td>
<td>870</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>733</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td>528</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>526</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>389</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td>370</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Profit Organizations</td>
<td>262</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>2965</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robots/Spiders visitors (Top 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 different robots*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Googlebot</td>
<td>312592</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inktomi Slurp</td>
<td>247505</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSNBot</td>
<td>182548</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown robot (identified by 'robot')</td>
<td>110439</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown robot (identified by 'spider')</td>
<td>17117</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unknown robot (identified by 'crawl')</td>
<td>8685</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BaiDuSpider</td>
<td>7469</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AskJeeves</td>
<td>6955</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GigaBot</td>
<td>2786</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASpider (Associative Spider)</td>
<td>2299</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1689</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Robots shown here gave hits or traffic “not viewed” by visitors, so they are not included in other charts.
<table>
<thead>
<tr>
<th>Search Keyphrases (Top 20)</th>
<th>Search</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>3060 different keyphrases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abreviation pays</td>
<td>202</td>
<td>2.6 %</td>
</tr>
<tr>
<td>abréviation pays</td>
<td>158</td>
<td>2 %</td>
</tr>
<tr>
<td>country name abbreviation</td>
<td>134</td>
<td>1.7 %</td>
</tr>
<tr>
<td>abréviation</td>
<td>133</td>
<td>1.7 %</td>
</tr>
<tr>
<td>abreviation</td>
<td>122</td>
<td>1.5 %</td>
</tr>
<tr>
<td>regions of the world</td>
<td>112</td>
<td>1.4 %</td>
</tr>
<tr>
<td>国家名</td>
<td>65</td>
<td>0.8 %</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>0.8 %</td>
</tr>
<tr>
<td>midp ?_?? filetype pdf</td>
<td>62</td>
<td>0.8 %</td>
</tr>
<tr>
<td>abbreviations etc</td>
<td>53</td>
<td>0.6 %</td>
</tr>
<tr>
<td>eec</td>
<td>52</td>
<td>0.6 %</td>
</tr>
<tr>
<td>短縮</td>
<td>50</td>
<td>0.6 %</td>
</tr>
<tr>
<td>acronymy</td>
<td>46</td>
<td>0.5 %</td>
</tr>
<tr>
<td>abréviations pays</td>
<td>44</td>
<td>0.5 %</td>
</tr>
<tr>
<td>cee</td>
<td>40</td>
<td>0.5 %</td>
</tr>
<tr>
<td>european economic community</td>
<td>39</td>
<td>0.5 %</td>
</tr>
<tr>
<td>abréviation des pays</td>
<td>38</td>
<td>0.4 %</td>
</tr>
<tr>
<td>etc abbreviation</td>
<td>37</td>
<td>0.4 %</td>
</tr>
<tr>
<td>eu</td>
<td>32</td>
<td>0.4 %</td>
</tr>
<tr>
<td>abbreviation of words</td>
<td>31</td>
<td>0.4 %</td>
</tr>
</tbody>
</table>
2.2. Pages

- Pages visited are static HTML
- using a CSS stylesheet
- They are created automatically from a Topic Map
- written in LTM\(^2\)
- using simple bespoke P\(ERL\) scripts

\(^2\)http://www.ontopia.net/download/ltm.html
WWW Virtual Library --- Topics Database

Benin

Capital city: Porto-Novo
Towns: Porto-Novo
broader categories
Africa
The Francophonie

This record is of type: Countries, States, Jurisdictions, etc.
WWW Virtual Library --- Topics Database

The Francophonie

Quick search:
Organisation Internationale de la Francophonie external link provides a definitive list of members, and other details of the Francophonie.

Relevant VL member library:

Francophonie

WWW Virtual Library Arbitration Database
This subject is also covered in the WWW Virtual Library Arbitration Database.

WWW Virtual Library Law
This subject is also covered in the WWW Virtual Library Law.

Note
All “members” are listed here. That is the forty-nine full members, the four “états associés” (Albania, Andorra, Greece and Macedonia) and the ten “états observateurs” (Armenia, Austria, Croatia, the Czech Republic, Georgia, Hungary, Lithuania, Poland, Slovakia and Slovenia).

This record is of type: Regions of the World.
francophonie et Francophonie

La francophonie désigne la communauté composée des personnes qui ont en commun l’usage de la langue française. Avec un F majuscule, elle désigne la Francophonie institutionnelle, dont le principal acteur est l’Organisation internationale de la Francophonie.

À la base d’un travail de Doctorat sur le thème de « La géopolitique des aires linguistiques : le cas de la Francophonie institutionnelle », ce site vise à répertorier un certain nombre de sites internet touchant au thème de la francophonie.

La liste non exhaustive sera complétée au fil des recherches.
3. The three phases

3.1. Phase I

- A hierarchical catalogue of the two hundred or so member libraries was moved from hand-written webpages to a topic map.
- Hierarchical webpages are then built and published automatically each night, from that topic map.
At the same time:

- Each VL was indexed under as many relevant category names as logical. (Thus *Admiralty and Maritime Law* is indexed under *Law*, *Transportation* and *Water*.)
- A description was added to each entry. Typically of some fifty words.
- Some, at first sight, perhaps bizarre cross-categories were created. These included topics such as *Sound*, *Light* and *Water*. These allow users to ‘think outside the box’: arts-based Music and the scientific vibration engineering now occur on the same page, as do *Admiralty and Maritime Law* and *Oceanography*. 
Simultaneously, the catalogue was offered in various other languages. Initially French, then Spanish and basic Chinese. The language choice is automatic relying on a procedure known as ‘content negotiation’: the user of the web browser states language preferences when setting up the browser and these are relayed with each webpage request.
Results:

The effect on visitors to the catalogue was neutral. (An increase of 1.9% in visits.)
The anecdotal effect on member libraries was highly positive. Many maintainers emailed to not only compliment the better æsthetic quality of the new catalogue, but also to report they were receiving more users accessing their individual VLs.

Three example libraries’ number of visits were analysed. VLs ‘A’ and ‘C’ were complex multi-page VLs, whilst ‘B’ was a classic one–page links catalogue. All were mature sites (four years and more on the Web) though ‘B’ had only been entered in the wwwwvl catalogue at the turn of 2004/5.
**A and B:** Increases of 29% and 78% on average for the first nine months of 2006.
Only one VL reported a decline in visitors. This was a dramatic fall, and the possible reasons for it are discussed below (*Unforeseen results*).

**Library C:** In this case ‘pages visited’ (rather than ‘visits’). The decline year on year for the first five months is 42%, though largely accounted for by January and February.
3.2. Phase II

In May 2005 [actually I await access to archive.org, to check on this!] a very simple topic map of World geography was created. This listed the countries of the world. It also had their capital cities and a few other towns (a data set carried over from another project). It grouped countries in continents, other groupings (e.g. the Francophonie) were added later.
Visitors viewed more pages per visit, and re-visited more often within the same calendar month:

<table>
<thead>
<tr>
<th></th>
<th>catalogue</th>
<th>topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>pages viewed per visit</td>
<td>2.7</td>
<td>4.6</td>
</tr>
<tr>
<td>visits per visitor/month</td>
<td>1.33</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Data for first nine months of 2006.
Unforeseen results:

Library C
3.2.1. Phase III

From Spring 2006 the basic geographical topic map was expanded.
Even after excluding robots (about half of all ‘hits’), there was not only broad use of all pages, but most pages at one time were either a point of ‘entry’ or ‘exit to the site.

<table>
<thead>
<tr>
<th></th>
<th>entry</th>
<th>exit</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>92.1</td>
<td>93.2</td>
</tr>
<tr>
<td>English en</td>
<td>95.8</td>
<td>97.0</td>
</tr>
<tr>
<td>French fr</td>
<td>85.2</td>
<td>85.9</td>
</tr>
<tr>
<td>Chinese zh</td>
<td>95.4</td>
<td>96.6</td>
</tr>
</tbody>
</table>

Percentage of total possible URLs that were reported as ‘entry’ or ‘exit’ pages (that is as the first or last page of an entire visit).
<table>
<thead>
<tr>
<th></th>
<th>‘index’</th>
<th>internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>4180</td>
<td>8587</td>
</tr>
<tr>
<td>English en</td>
<td>4133</td>
<td>2500</td>
</tr>
<tr>
<td>French fr</td>
<td>31</td>
<td>3227</td>
</tr>
<tr>
<td>Chinese zh</td>
<td>16</td>
<td>2860</td>
</tr>
</tbody>
</table>

Visits to `index.en.html` or just ‘/’, (left hand column), visits to the site’s main page(s) from internal links (right hand column).
This contrasts markedly with vlsearch.org where over 9% of visits request a non–*.en.* page.
The most popular pages were those listing regions, countries, towns, abbreviations (generally and for country names) and dealing with the United States and Canada.

Popularity was almost independent of page language.

The most popular ten pages received 3057–8656 hits each, the most popular one hundred 1469–8656, and the most popular one thousand 838–8656 hits (including robots). Whilst some of this might be accounted for by out-of-control robots, the pages receiving high scores were also those matching search engine search criteria in the referral logs.
4. Conclusions

A rich set of data has only just started to be analysed.

Even with, as yet, only very limited outlinking to member VLs the pages have drawn users, who have explored them (visiting a relatively high number of pages per visit) and have in about half the cases chosen to return within a month.

Analysis of actual paths through the topics would yield further insights into how Web users move through such a topic map.

In the coming months the range of topics mapped will move beyond the basic (largely geographical) core.
It is intended to link deeper into member VLs (rather than predominantly just to main entry pages).

1. Of great importance is the need for an easy system to query the underlying topic map(s). The web directories produced from topic maps definitely have marked advantages over their predecessors and other alternatives. However to use conventional searching of these static pages, rather than querying the actual TM loses a great opportunity.

2. Of greatest importance topics must cease to be isolated. Topics must have PSIs. Fundamental decisions will have to be made about how to use PSIs (and how far to diverge from mono– or bi–lingual PSIs in an attempt to maintain multi–lingual usage). Those decisions will in great effect be based on the discussions that take place during this week.
3. Examples of real world usages of topic maps will do much to raise their profile, and undoubtedly free-of-charge software will help in creating and enabling viewing such examples. However, my personal view, is that to engender wider *confidence* in topic maps there is a need for some basic³ *open source* tools.⁴

³comparable perhaps with the Lynx browser rather than Firefox!
⁴This final comment resulted from discussions at the Wrap Up session on the final day of the conference, rather than in this presentation, but is recorded here for completeness.
vlib.org