The World Wide Web Virtual Library

Introduction
Materials and Methods
The three phases
Conclusions

Introduction
Materials and Methods
The three phases
Conclusions

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.

Introduction
Materials and Methods
The three phases
Conclusions

1. Introduction
2. Materials and Methods
3. The three phases
4. Conclusions

Introduction

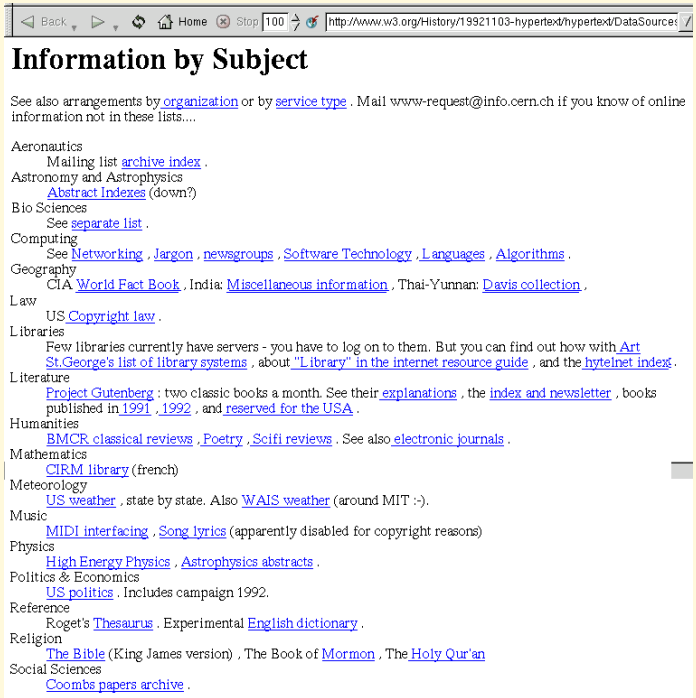
Materials and Methods

The three phases

Conclusions

1. Introduction

Overview.html



Back Home Stop 100 http://www.w3.org/History/19921103-hypertext/hypertext/DataSources

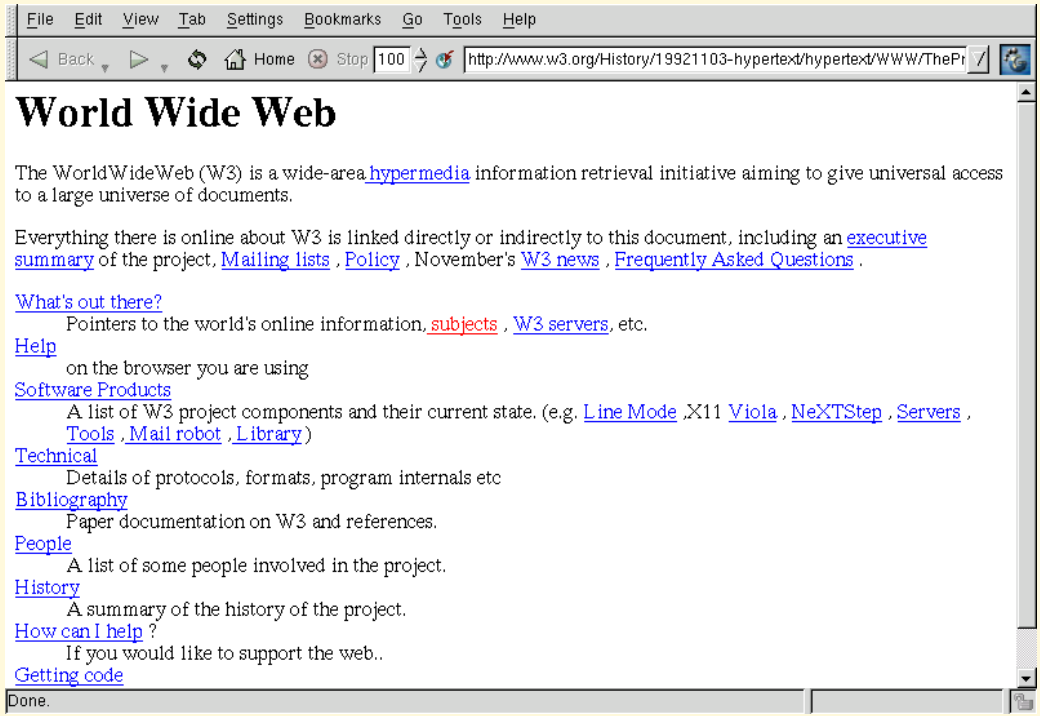
Information by Subject

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 , 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 Art St.George's list of library systems , about "Library" in the internet resource guide , and the hytelnet 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 , Scifi 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
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
Coombs papers archive .

Introduction
Materials and Methods
The three phases
Conclusions

November 3rd 1992:



Introduction
Materials and Methods
The three phases
Conclusions

"… 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."

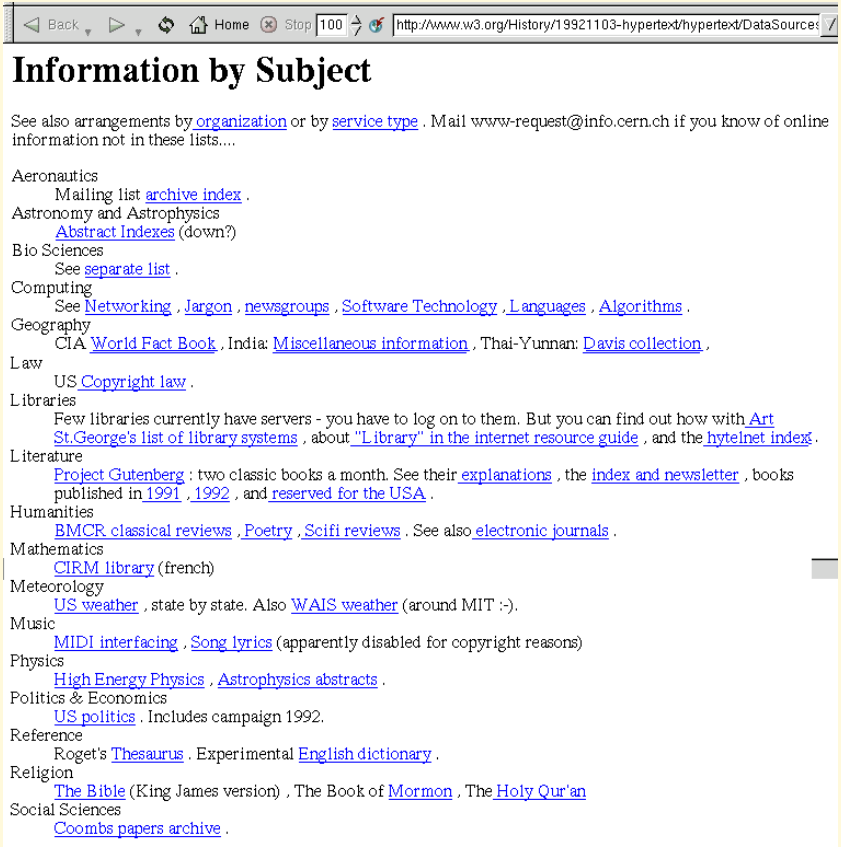
Tim Berners-Lee *Weaving the Web* (Harper Business ISBN 0-06-251587-X), pp. 47 & 49.

Introduction

Materials and Methods

The three phases

Conclusions



Introduction
Materials and Methods
The three phases
Conclusions

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 . . .)

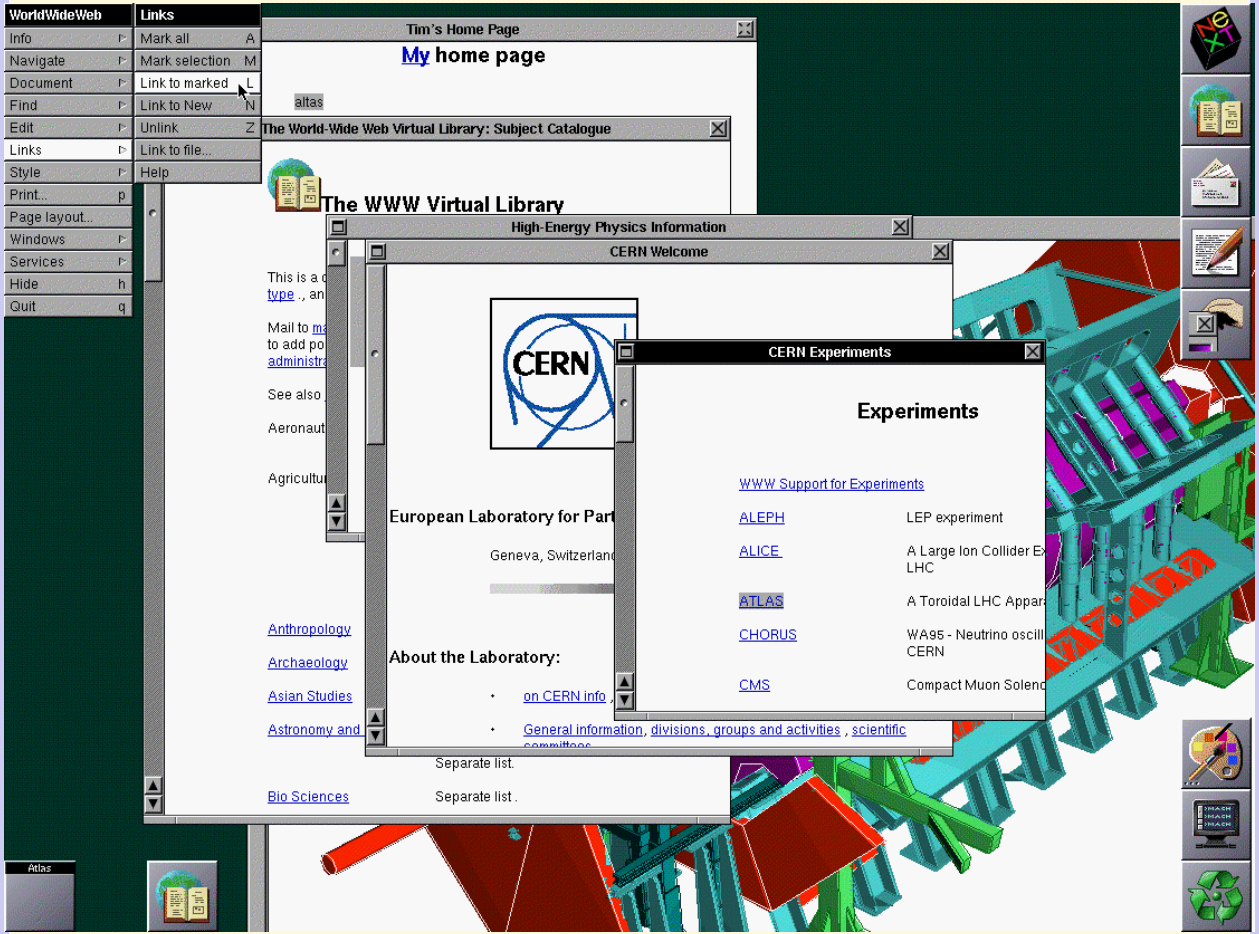
Ciolek: Global Networking: a Timeline.

Introduction

Materials and Methods

The three phases

Conclusions

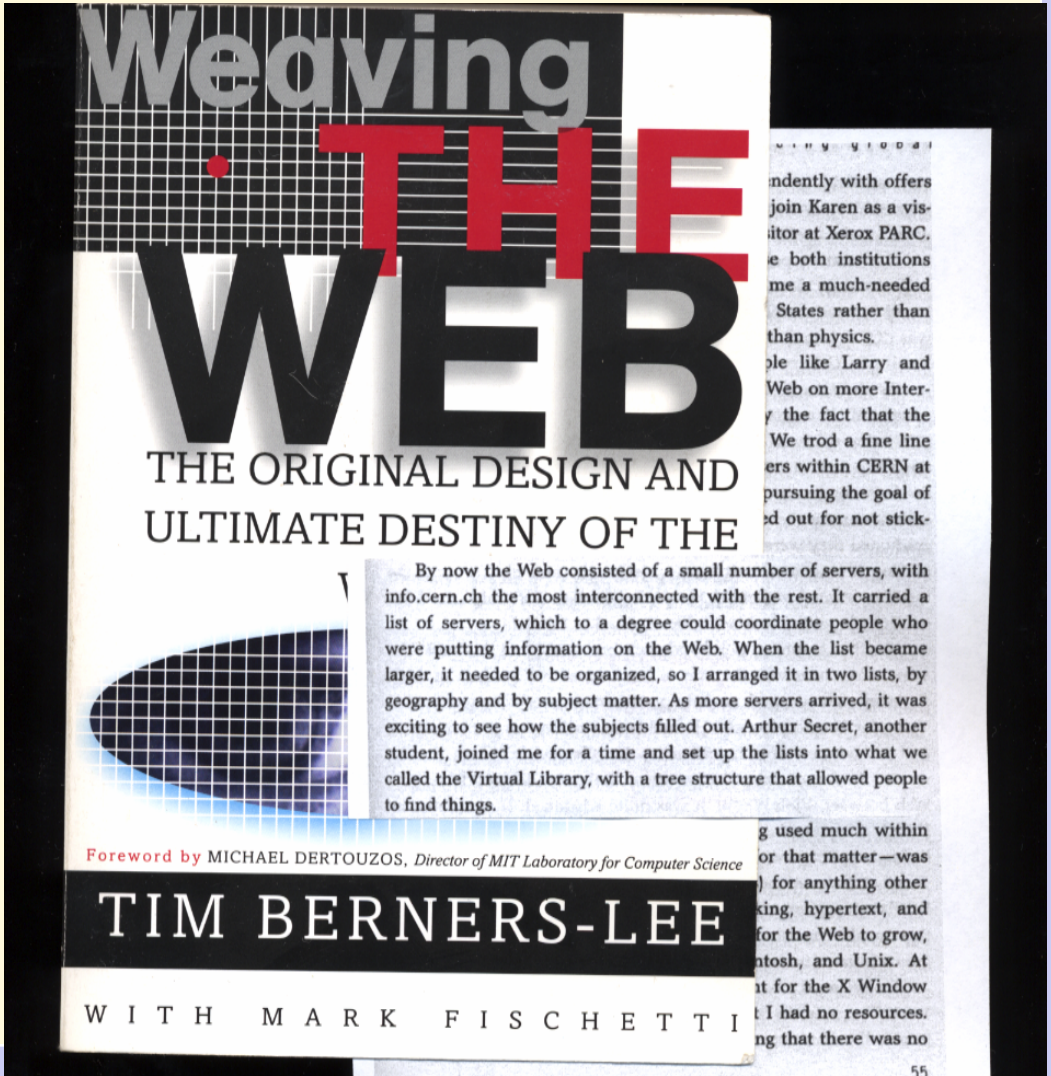


Introduction

Materials and Methods

The three phases

Conclusions



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.

Introduction
Materials and Methods
The three phases
Conclusions

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.

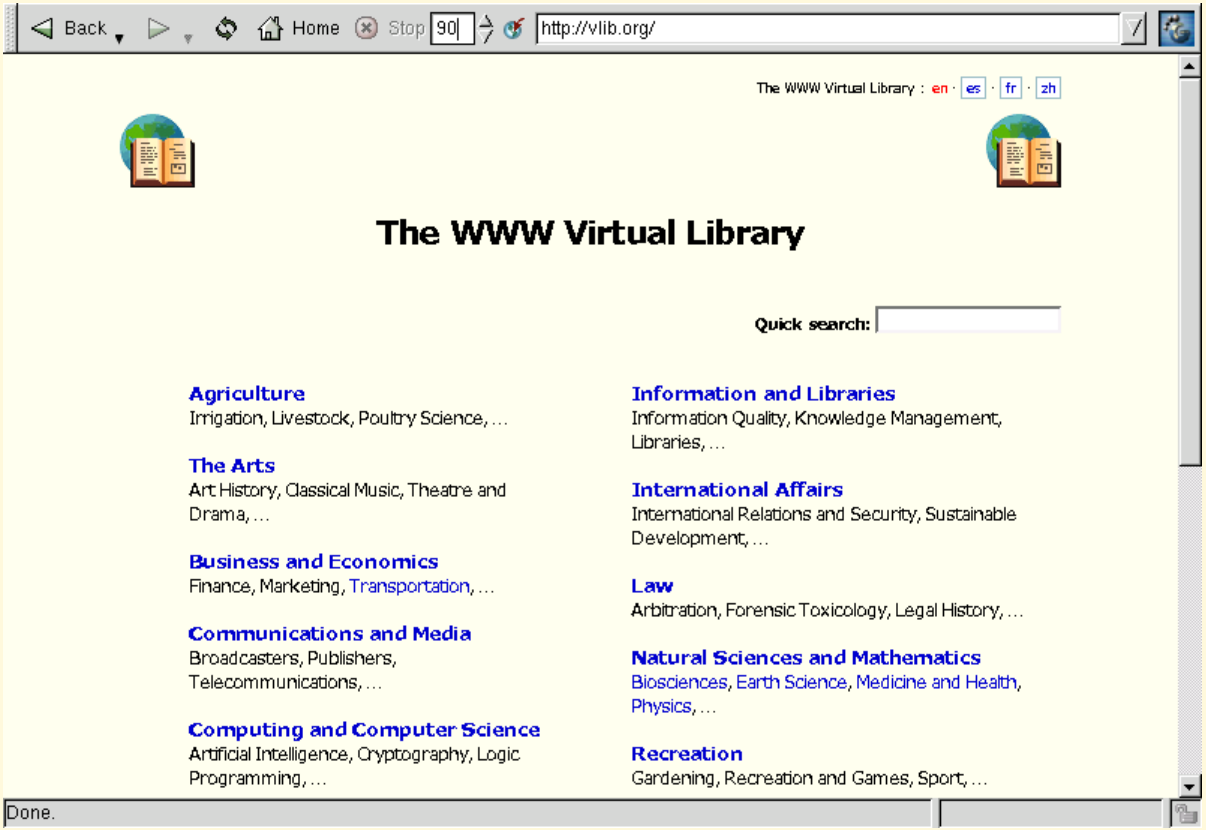
"It is futile to discuss whether the Index *or the table of* Contents *is the more useful in a book."*

Introduction

Materials and Methods

The three phases

Conclusions



Introduction

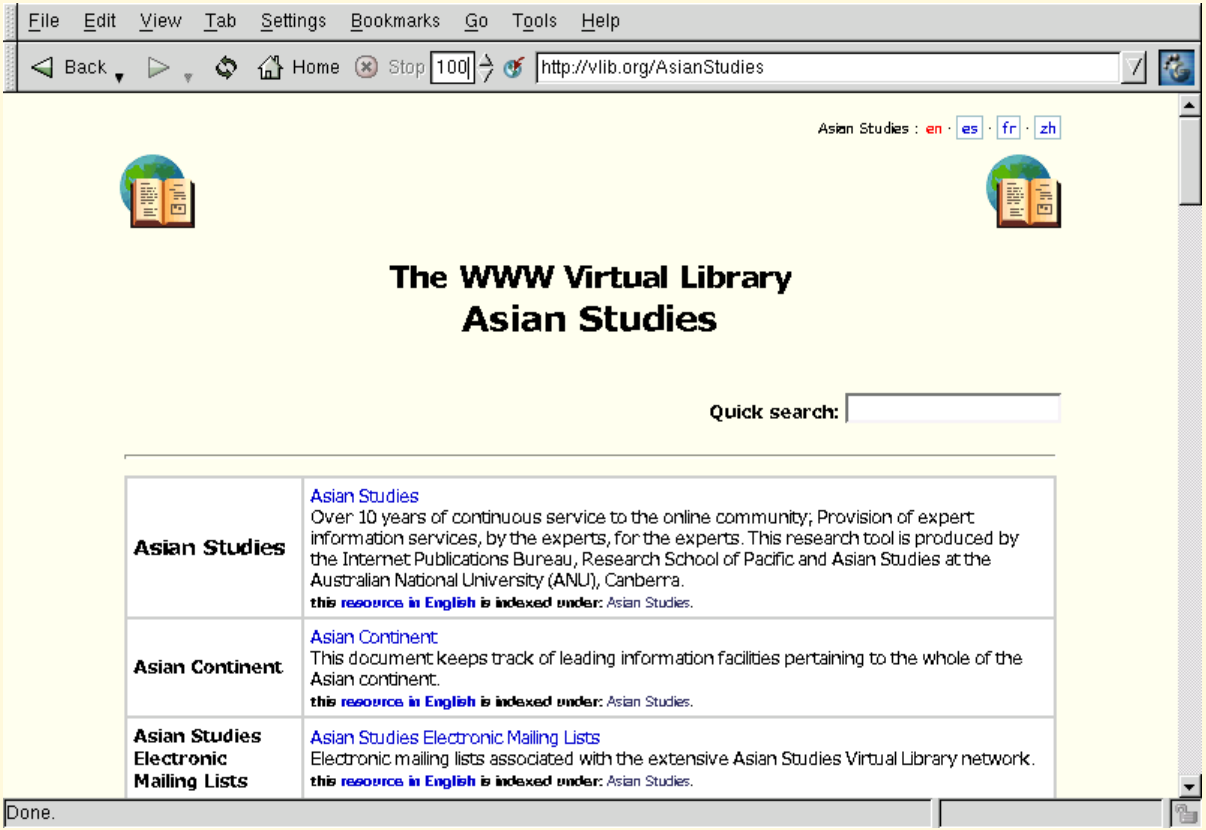
Materials and Methods

The three phases

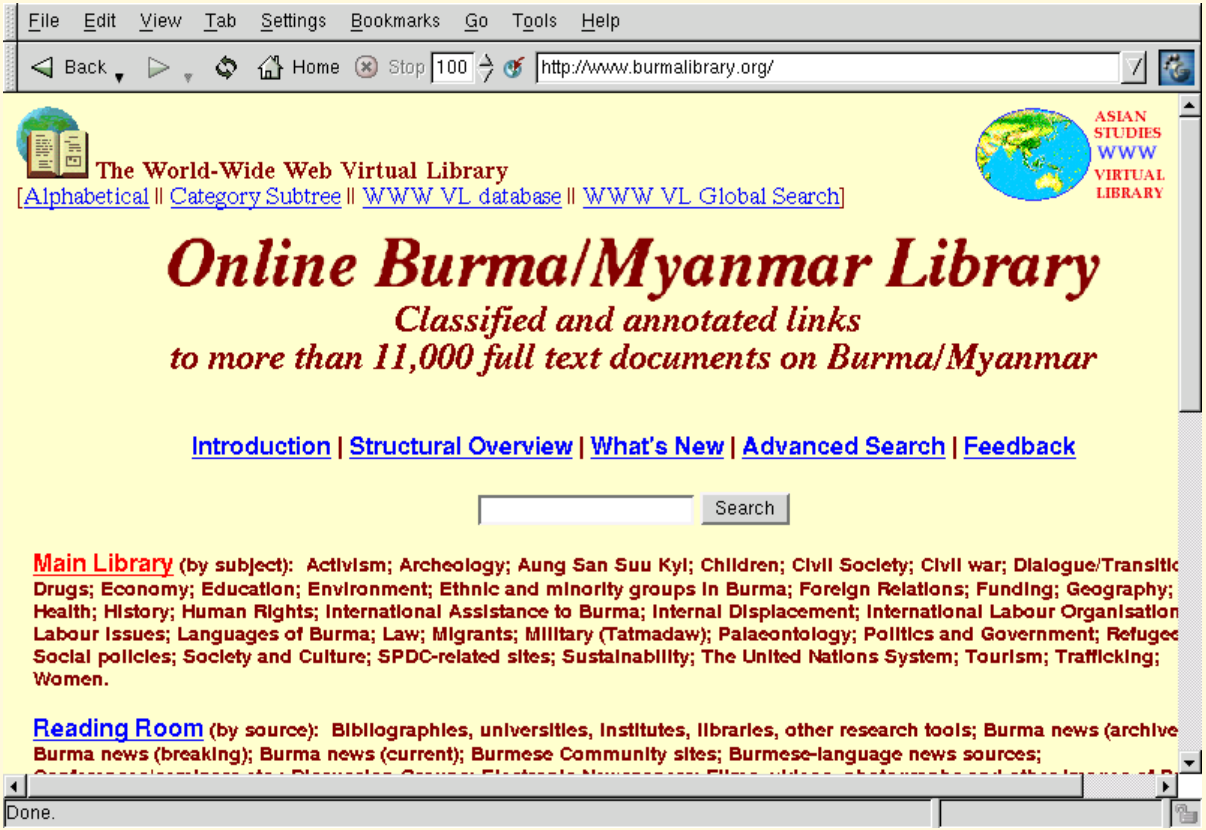
Conclusions

Interfaces in various languages.

Introduction
Materials and Methods
The three phases
Conclusions



Introduction
Materials and Methods
The three phases
Conclusions



Introduction
Materials and Methods
The three phases
Conclusions

1.0.1. VLsearch

Introduction
Materials and Methods
The three phases
Conclusions

World Health Organisation —simple search

Introduction
Materials and Methods
The three phases
Conclusions

World Health Organisation:

simple search	179
“string”	10
without synonyms	113
“string” without synonyms	4

Introduction

Materials and Methods

The three phases

Conclusions

WWWVL

Introduction
Materials and Methods
The three phases
Conclusions

2. Materials and Methods

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

Introduction
Materials and Methods
The three phases
Conclusions

2.1. Webserver log

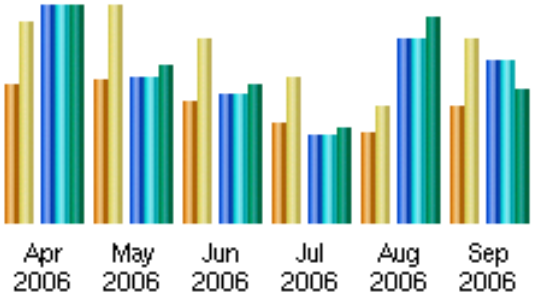
Logs: page visited, date/time of request, IP/domain of requester

Analyse using *awstats*[1]

[1] http://awstats.sourceforge.net/

Introduction
Materials and Methods
The three phases
Conclusions

Monthly history



Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Apr 2006	2370	3443	19516	19516	50.42 MB
May 2006	2432	3703	13207	13207	36.50 MB
Jun 2006	2068	3145	11628	11628	32.15 MB
Jul 2006	1693	2495	7871	7871	21.99 MB
Aug 2006	1559	2013	16509	16509	48.06 MB
Sep 2006	1997	3160	14708	14708	31.02 MB

Introduction
Materials and Methods
The three phases
Conclusions

Visitors domains/countries

	Domains/Countries		Pages
?	Unknown	ip	30662
	Network	net	20136
	China	cn	11014
	Australia	au	7134
	Commercial	com	5308
	Switzerland	ch	1670
	France	fr	1634
	Canada	ca	870
	Netherlands	nl	733
	Belgium	be	528
	Japan	jp	526
	Taiwan	tw	389
	United Kingdom	uk	370
	Germany	de	365
	Non-Profit Organizations	org	262
	Others		2965

Introduction

Materials and Methods

The three phases

Conclusions

Robots/Spiders visitors (Top 10) - Full list - Last visit

25 different robots*	Hits
Googlebot	312592
Inktomi Slurp	247505
MSNBot	182548
Unknown robot (identified by 'robot')	110439
Unknown robot (identified by 'spider')	17117
Unknown robot (identified by 'crawl')	8685
BaiDuSpider	7469
AskJeeves	6955
GigaBot	2786
ASpider (Associative Spider)	2299
Others	1689

* Robots shown here gave hits or traffic "not viewed" by visitors, so they are not included in other charts.

Introduction
Materials and Methods
The three phases
Conclusions
Home Page
Title Page
Page 25 of 51
Go Back
Print
Close
Quit

Search Keyphrases (Top 20)

Full list

3060 different keyphrases	Search	Percent
abreviation pays	202	2.6 %
abréviation pays	158	2 %
country name abbreviation	134	1.7 %
abréviation	133	1.7 %
abreviation	122	1.5 %
regions of the world	112	1.4 %
国家名	65	0.8 %
	64	0.8 %
midp ?_?? filetype pdf	62	0.8 %
abbreviations etc	53	0.6 %
eec	52	0.6 %
縮寫	50	0.6 %
acronyme	46	0.5 %
abréviations pays	44	0.5 %
cee	40	0.5 %
european economic community	39	0.5 %
abréviation des pays	38	0.4 %
etc abbreviation	37	0.4 %
eu	32	0.4 %
abbreviation of words	31	0.4 %

Introduction

Materials and Methods

The three phases

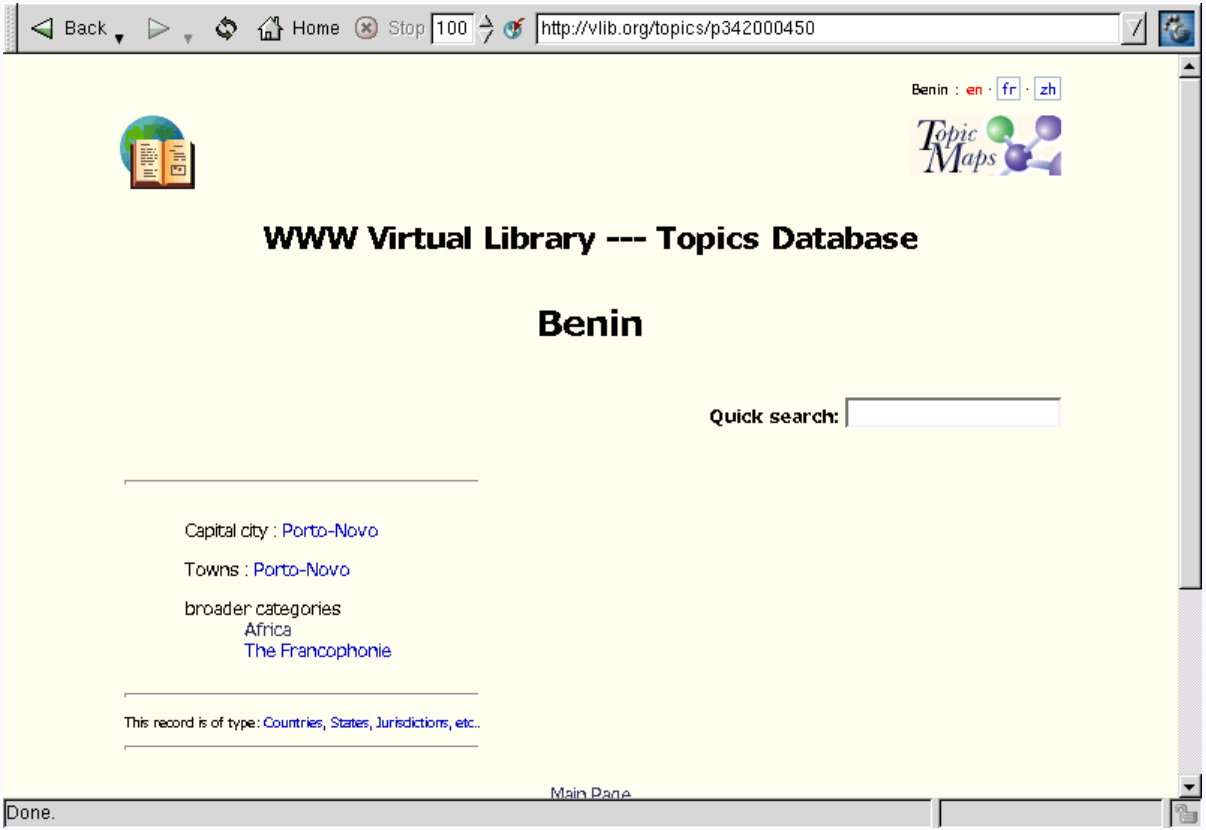
Conclusions

2.2. Pages

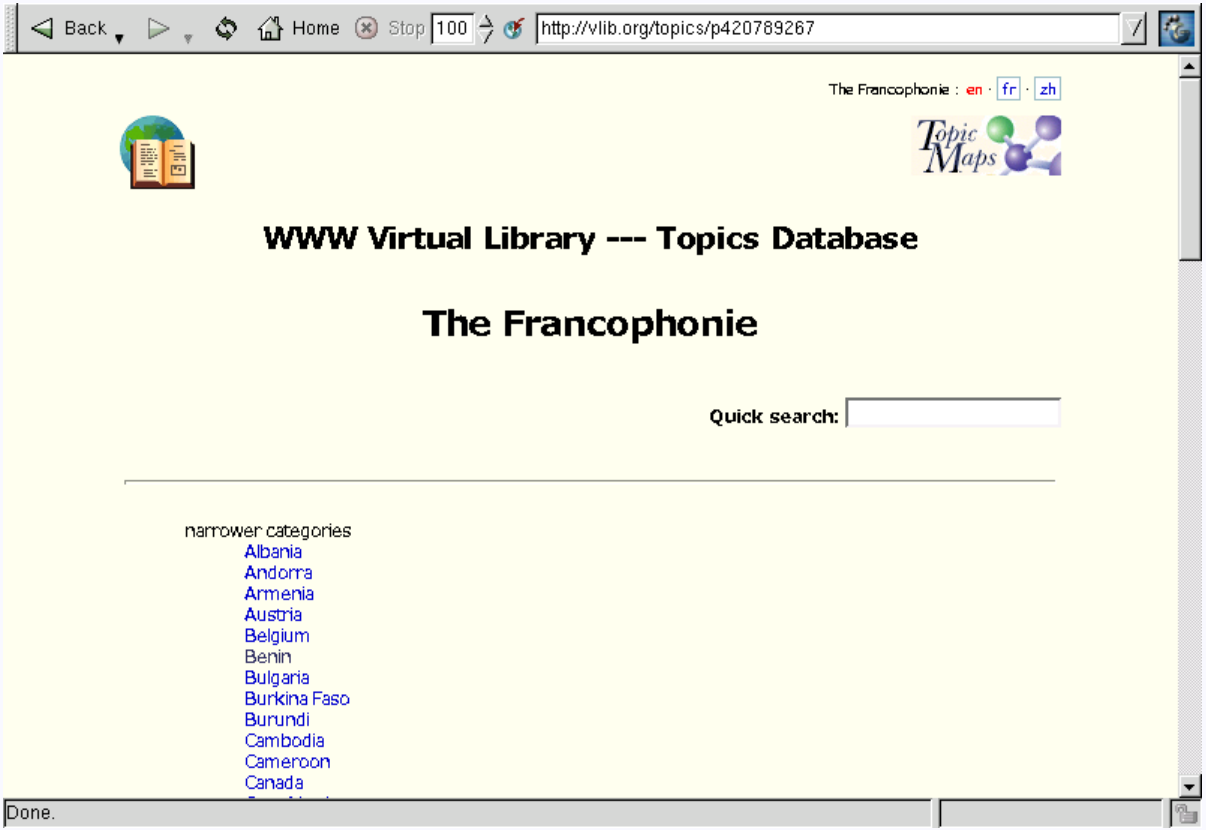
- Pages visited are static HTML
- using a CSS stylesheet
- They are created automatically from a Topic Map
- written in LTM[2]
- using simple bespoke PERL scripts

[2] http://www.ontopia.net/download/ltm.html

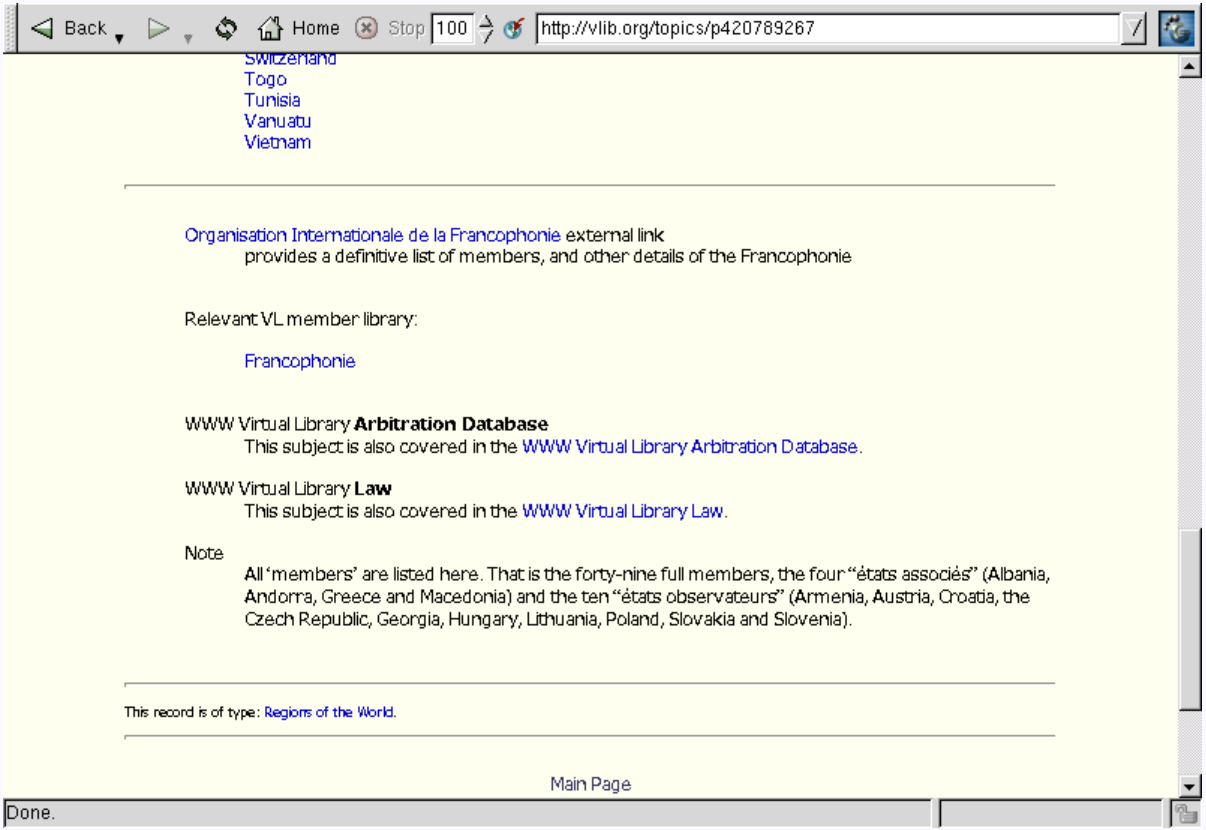
Introduction
Materials and Methods
The three phases
Conclusions



Introduction
Materials and Methods
The three phases
Conclusions



Introduction
Materials and Methods
The three phases
Conclusions



Introduction
Materials and Methods
The three phases
Conclusions

Introduction
Materials and Methods
The three phases
Conclusions

vlib.org

Introduction
Materials and Methods
The three phases
Conclusions

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.

Introduction
Materials and Methods
The three phases
Conclusions

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*.

Introduction
Materials and Methods
The three phases
Conclusions

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.

Introduction

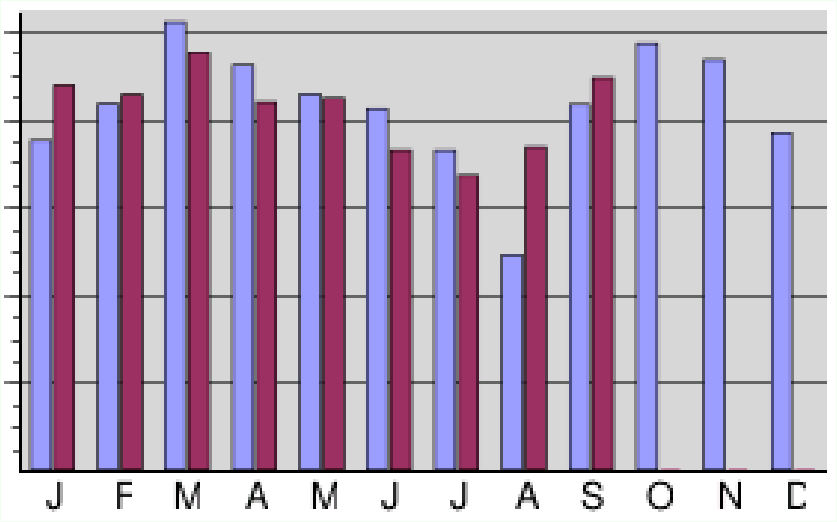
Materials and Methods

The three phases

Conclusions

Results:

The effect on visitors to the catalogue was neutral. (An increase of 1.9% in visits.)



Introduction
Materials and Methods
The three phases
Conclusions

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.

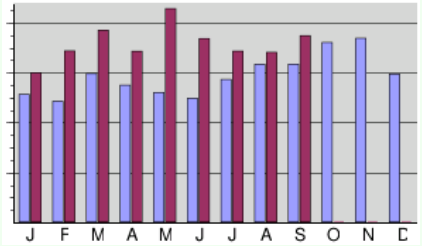
Three example libraries' number of visits were analysed. VLs '*A*' and '*C*' were complex multi-page VLs, whilst '*B*' was a classic one–page links catalogue. All were mature sites (four years and more on the Web) though '*B*' had only been entered in the WWWVL catalogue at the turn of 2004/5.

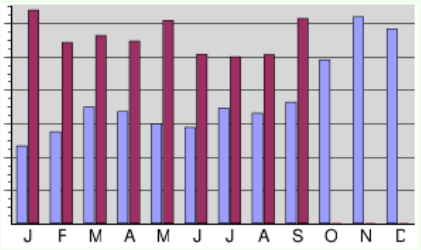
Introduction

Materials and Methods

The three phases

Conclusions



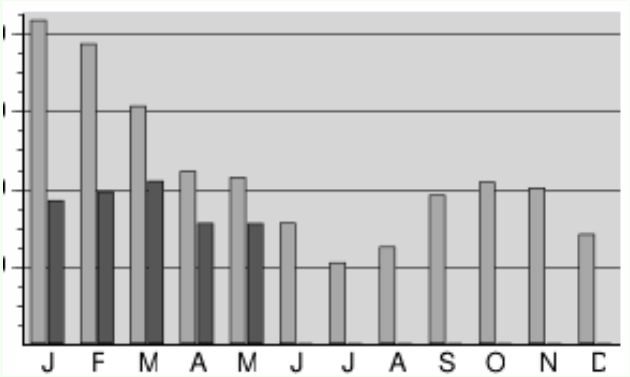


A and B: Increases of 29% and 78% on average for the first nine months of 2006.

Introduction
Materials and Methods
The three phases
Conclusions

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.



Introduction
Materials and Methods
The three phases
Conclusions

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.

Introduction
Materials and Methods
The three phases
Conclusions

Visitors viewed more pages per visit, and re-visited more often within the same calendar month:

	catalogue	topics
pages viewed per visit	2.7	4.6
visits per visitor/month	1.33	1.47

Data for first nine months of 2006.

Introduction

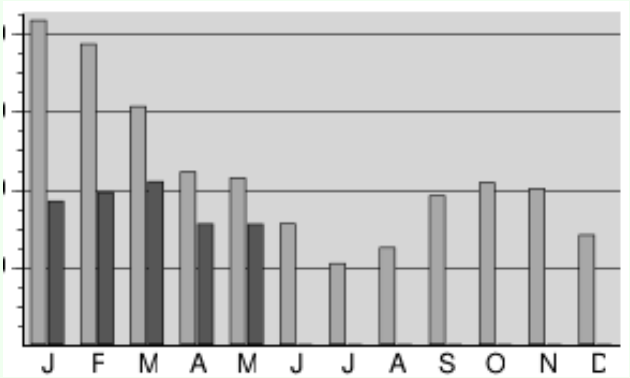
Materials and Methods

The three phases

Conclusions

Unforeseen results:

Library C



Introduction
Materials and Methods
The three phases
Conclusions

3.2.1. Phase III

From Spring 2006 the basic geographical topic map was expanded.

Introduction
Materials and Methods
The three phases
Conclusions

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.

	entry	exit
total	92.1	93.2
English en	95.8	97.0
French fr	85.2	85.9
Chinese zh	95.4	96.6

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).

Introduction

Materials and Methods

The three phases

Conclusions

	‘index’	internal
total	4180	8587
English en	4133	2500
French fr	31	3227
Chinese zh	16	2860

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.

Introduction

Materials and Methods

The three phases

Conclusions

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.

Introduction

Materials and Methods

The three phases

Conclusions

Introduction
Materials and Methods
The three phases
Conclusions

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.

Introduction

Materials and Methods

The three phases

Conclusions

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.

Introduction
Materials and Methods
The three phases
Conclusions

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 basic[3] ***open source*** tools.[4]

[3] comparable perhaps with the Lynx browser rather than Firefox!

[4] This 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.

Introduction

Materials and Methods

The three phases

Conclusions

vlib.org

Introduction
Materials and Methods
The three phases
Conclusions