

Why words matter

the vagaries of language, and the nature of book indexes versus web search

Glenda Browne
www.webindexing.biz

**How would you index
this?**



For a book index, you'd use:

red-back spiders (or redback spiders)

If you had scientific readers, you might add:

Latrodectus hasselti, see red-back spiders

And for people who didn't know exactly what they wanted, perhaps you'd also use:

spiders, 15, see also red-back spiders

Then you could get creative ...

arachnids, *see* mites; scorpions; spiders
dangerous animals 93, *see also* red-back
spiders

'Red-back on my toilet seat' 65

red-back spiders

habitat 57

songs about 65

webs 34

For all their simplicity...

...book indexes provide a lot of control

- One word form (lexical variant) gathers all content (**redbacks** or **red-backs**)
- Access from alternative terms (from scientific to common names)
- References from broader terms (**spiders**, **dangerous animals**) – chosen for the specific work and expected users

For all their simplicity (2)...

- Access to related topics that may also be of interest (eg, the song title in alpha proximity or as a reference or subheading)
- Subheadings to clearly show what aspect of the topic is being presented (eg, **habitat**)

You can get book-style index benefits on the web

[abstracting \(1996 conference paper\)](#)

[ACT Region Branch](#)

[contacts](#)

[established](#)

[advertising in *ANZSI Newsletter*](#)

[annual report indexing](#)

[courses](#)

ANZSI

[about](#)

[aims and services](#)

[background](#)

[branches and groups](#)

[conferences](#)

[constitution](#)

- www.anzsi.org
- Browne & Jermey, *Website indexing*, 2nd ed, 2004, www.webindexing.biz

Unfortunately...

...most of this is much harder to do with search on the web or on intranets

- because sites rarely use **browsable** indexes, and therefore rely more on automatic expansion of searches (if any) or search engine algorithms

And people get so much stuff from online searches, they don't realise that they're missing anything

Language issues to deal with

- Word variants
- Synonyms
- Homonyms
- Broader, narrower and related terms
- Aspects of topics

Word variants – typo help

- On Google I typed, in haste, **redbacs** and got the hint:
Did you mean: redbacks
- A search for exema brings 62,900 hits and the hint:
Did you mean: *excema* (199,000 hits)
- But if you ignore their hint, and search for **eczema**, you get 8,680,000 hits!

Word variants – typo hindrance

- Similarly, a search for **accomdation** in an intranet brought the hint:

Did you mean: *accomodation*

- A search for **hypoglyceamia** suggested:

Did you mean: *hypoglycemia*

- There were no hits for this. There were for hypoglycaemia, but it wasn't suggested!

Word variants – alternative spellings

- On the web, searches for **jewellery** and **jewelry** retrieve different hits, including different sponsored hits and related links, with no suggestions offered
- A search for **automatic categorisation**, OTOH, brings the hint:
Did you mean: **automatic categorization**

Google clearly has a point of view...

...and it is American, as a search for **automatic categorization** (the US norm) provides no such suggestion

- because Google's suggestions are based on algorithms not logic

How much difference can one letter make?

- Searches for **automatic categorisation** and **automatic categorization** on 6/7/08 retrieved NO common hits in the top 10
- You also get extra useful stuff in a search for **automated categorisation** (including the Wikipedia topic **document categorisation**)

Does it matter to you?

- As a searcher, it means you have to think of alternative search terms if you want maximum recall
- As a webpage provider, it means you have to provide alternative word forms if you want people from around the world to find your site

Easy, obvious solutions

- For an intranet, a synonym list allows you to give searchers an option to expand their search
- For the web, wouldn't it be nice if Google had a synonym list that it could use to supplement its user behaviour (primarily US) based algorithms?
- Or if you could add alternatives as subject keyword metadata that would be searched?

Homonyms are harder...

- Try searching Google for information on 'travelling children'
- Defined by the Prime Minister's Office (UK) as *Persons of nomadic habit of life whatever their race or origin*. Includes gypsies, New Age Travellers, and bargees and boat communities
- Searching for 'travelling children' finds nothing relevant in the top hits. Searching for 'travelling children gypsies' narrows the search, but at the risk of missing stuff
- www.six.somerset.gov.uk/equalities/do_download.asp?did=23642

Homonyms in indexes

- In a browsable index, homonyms are easily distinguished using parenthetical qualifiers (glosses)
 - Howard, John (actor)
 - Howard, John (ex-Prime Minister)
- In a searchable, controlled bibliographic index, they are managed in the same way, and searchers consult the thesaurus to find the appropriate term
 - races (structures)
 - (referring to narrow passages for livestock)

Homonyms on the web

- On the web, how can you distinguish:
Pandora (NLA archiving service)
from
Pandora (Internet music radio
station)?

Wikipedia does it well

Pandora (disambiguation) (selected items only)

Pandora may refer to: [Pandora](#), the Greek mythological woman, as well as her granddaughter [Pandora II](#).

People

[Pandora \(band\)](#), a Mexican music group

[Pandora Peaks](#), a former [exotic dancer](#).

Places

The village of Pandora, an unincorporated community in [Wilson County, Texas](#)

Astronomy

[Pandora \(moon\)](#), a moon of [Saturn](#)

Life

[Pandora \(bivalve\)](#), a genus of mollusc

Culture

[Pandora \(music service\)](#), an Internet music site and radio station with various third-party tools associated with it

[Pandora \(novel\)](#), a novel by Anne Rice about a vampire of the same name who has appeared in several of her other books

The [Pandora Archive](#), an online archive run by the National Library of Australia

Clusty.com grouping

All Results (247)

Jewelry (34)

Mythology, Greek (25)

Photos (26)

Box (26)

Music (18)

Software (10)

Artist (7)

Response (6)

Australia (7)

No easy, obvious solution

- But perhaps Google could provide a list of homonyms, such as Wikipedia provides, and let people allocate their site to one of the options

Hierarchies don't automate well

- It's a lot harder to automate provision of hierarchical information
 - Thesauruses allow you to provide information on BTs, NTs, and RTs
 - Faceted schemes allow users to move up and down a number of hierarchies

Thesaurus expansion and guidance

Tips for searching

Search on these terms?

[occupational health and safety](#)

Broader Terms

[health](#)
[safety](#)

Related Terms

[illness](#)
[industrial accidents](#)
[injuries](#)
[work environment](#)
[workplace harassment](#)

Search Results

Go to [Advanced Search](#)





Your search for:

Title = "OH & S"; "OH&S"; "OHS"; "occupational health and safety" or

Description = "OH & S"; "OH&S"; "OHS"; "occupational health and safety" or

Subject = "OH & S"; "OH&S"; "OHS"; "occupational health and safety"

Returned: 66 results.

-  [Occupational health and safety policy statement](#) [policy statement]
Occupational health and safety policy statement and associated policy circular from Director General, D.B. O'Connor.
Date: 1998
application/msword [73Kb]
-  [Occupational health and safety](#) [resource guide]
Index page to occupational health and safety and injury management and return to work policy, procedures and forms.
Date: 2003
text/html [13Kb]
-  [Workplace issues](#) [resource guide]
Index page to resources on workplace issues such as access and equity, conduct, grievances, occupational health and safety, etc.
Date: 2003
text/html [14Kb]
-  [Injury management and return to work : policies, procedures, forms](#) [policy statement]
Policy circular from Commissioner, D B OConnor.
Date: 2003
application/msword

Thesaurus behind the scenes – www.austrade.gov.au

- A 'hidden' thesaurus at Austrade:
 - Offers a broader term if no hits are retrieved:
0 results for *mandalay*.
Would you like to broaden your search? You could try: Myanmar
 - Automatically expands searches to include synonyms:
16 results for *burma* and similar terms including *Myanmar*.

Faceted search

facetmap.com/demo/browse.jsp



Browse Varietal

[Red Wines \(171\)](#), [White Wines \(149\)](#), [Bubbly \(40\)](#), [Pink Wines \(30\)](#), [Dessert/Fortified Wines \(41\)](#)

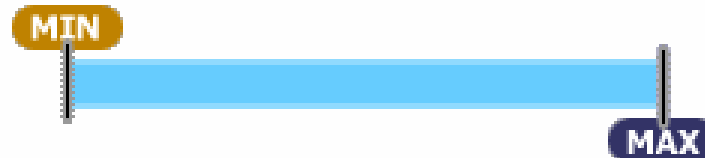
Browse Region

[French \(55\)](#), [German \(6\)](#),
[Italian \(67\)](#), [New Zealand \(2\)](#),
[Other European \(8\)](#),
[Portuguese \(19\)](#), [South American \(4\)](#), [Spanish \(15\)](#),
[USA \(255\)](#)

Browse Price

[Bargains under \\$20 \(237\)](#), [Top shelf \(over \\$100\) \(11\)](#)

Set your own Price:



(Drag pointers to select)

You've narrowed your search down to 431 results. Here are the top 10.
Turning Leaf 2000 White Zinfandel

Faceted search

- See also the NCSU libraries catalogue at www.lib.ncsu.edu/catalog/advanced.html for browsing via subject, genre, format, call number location, library, language, region, time period, author
- System is complex, but at least a student user group can develop familiarity with a system
- We need more experiments!

Aspects of topics

- Good book indexes use subheadings to show aspects of topics that have been discussed

Aspects of topics

- Search engines do this occasionally:

Refine results for **eczema**:

Treatment Tests/diagnosis For patients From medical authorities

Symptoms Causes/risk factors For health professionals Alternative medicine

- ...and even revise them as you go:

Refine results for **eczema**:

Treatment Tests/diagnosis For patients From medical authorities

Symptoms Causes/risk factors ***For health professionals*** Alternative medicine

Patient handouts Clinical trials Continuing education Practice guidelines

Language issues to deal with

- For all of these language and concept issues, search engines try, but don't do everything they could to improve findability

Something they both do badly – bias

- Manually-crafted indexes suffer from intentional or unintentional bias
- Search engines can be distorted by mass activity (link bombing and spamdexing)
- Search engine algorithms show specific viewpoints

Google shows the 'sunny side'...

- Google tends to show the 'sunny side' of controversial issues, because its algorithms are based on authors' links and searchers' choices

Susan L Gerhart (2004) 'Do Web search engines suppress controversy?' *First Monday* v.9 n.1,
firstmonday.org/issues/issue9_1/gerhart/index.html

Some things the web does better

- Place
- Specific, popular searches
- Search refinement
- Multiple viewpoints via tagging

Place

- Online maps and route finders
- Transport Infoline for timetables, maps and route diagrams – www.131500.info/realtimed/newjourney.asp
- Instant mapping of search results

*Search for **Blaxland, NSW***



Specific, popular, recent

- YouTube - Sort Of Dunno Nothin' - Peter Denahy
- A video of a dad asking an average teenager questions and the teenager gives average teenager 1 word responses. It's a funny video and you should watch...
- Yep, yep, nothin', nothin', sort of, dunno, nowhere, good, yep, nup, dunno, no one, sort of,

Refinement – limiting content using metadata

- Let users shrink the information space based on metadata queries, eg, if you know the year something was published, put that in.

Peter Morville, via Mike Moran

http://www.mikemorán.com/biznology/archives/2006/06/peter_morville.html

Refinement – Advanced Search

- Language: Esperanto
- File type: Adobe Acrobat PDF (.pdf)
- Content type: Media release
- Search within a site or domain: .edu

Refinement - Audience

- Refine results for **eczema**:

Treatment Tests/diagnosis For patients From medical authorities

Symptoms Causes/risk factors ***For health professionals*** Alternative medicine

Patient handouts Clinical trials Continuing education Practice guidelines

- Metadata-driven websites do this well

www.nps.org.au

Multiple viewpoints via tagging

- Social bookmarking brings the power of the people to indexing the web
- Selfish desires for re-findability lead people to tag sites of interest to them
- This tagging then offers useful links and pathways for other people

When does tagging help?

- Tagging helps:
 - When many people do it
 - When there is no other text, eg, for finding photos on Flickr
 - When the tagger's interests match the searcher's interests, and when it is carefully done (eg, del.icio.us.com/GlendaIndexer/website_indexes; del.icio.us/ACS_library)
 - As a complement to structured metadata/as a first step towards structured metadata
 - pace layering: the slow layers provide stability; the fast layers drive innovation

Pace layering

- Stewart Brand's concept of Pace Layering
 - Originally applied to buildings
 - Layers are able to evolve and change at different rates and paces
 - Folksonomies adapt quickly and allow for innovation and responsiveness
 - Thesauruses evolve slowly and provide stability and quality
 - Folksonomies can inform thesauruses

Trends in tagging - 2

- Specific suggestions for improving tagging:
 - More structure and control
 - The ability to subdivide tags
 - Providing tag definitions
 - Offering tag suggestions (such as the way delicious prompts you with tags)
 - Allowing true phrases (such as the ability to write **creative nonfiction** rather than **creativenonfiction**)

Gene Smith, www.iasummit.org/proceedings/2008/tagging_five_emerging_trends

K.G. Schneider,

freerangelibrarian.com/2008/05/21/ebony-and-ivory-tagging-and-taxonomies

How good is tagging?

- Roger Hudson reported at an earlier Open Publish conference on an experiment in which he showed an image of a redback spider to 40 people with various backgrounds, and asked them how they'd tag it.

Hudson, Roger. 'Information Classification and Retrieval - The folks and their tags' www.usability.com.au/resources/tags

Variety is the spice of life...

67 different tags used. Most common tags:

Redback / Redback spider (33) (various forms)

Spider (27)

Spider web (8)

Web (8)

Prey (6)

Australian spider (5)

Insect (5)

Poisonous spider (5)

Predator (4)

Arachnid (4)

Arachnology (2)

Entomology (2)

...and there's more

47 unique tags including:

Australian wildlife

Australian bush

biology, black widow

dangerous, death

eating

exotic animal

feeding

green

locust

sheds

Slim Dusty

spider killing

venom

Does accuracy matter?

- More people tagged it as an **insect** (which it isn't) than as an **arachnid** (which it is)
- Someone tagged it with the name **Slim Dusty**. But it was **Slim Newton** who sang **Red-back on my toilet seat!**

Can (and should) tagging be improved?

- Nothing can be imposed
- Guidelines should be presented. Simple suggestions such as joining words with underscore means we collocate subjects, instead of spreading them as:
 - informationarchitecture
 - information architecture [searched as 2 independent words]
 - information_architecture
 - IA

Best of the old and new

- Search works when the words on your website match the words that your users search with.
- Simple solutions such as the use of synonym lists, and more complex approaches such as the use of thesauruses and browsable lists, can enhance findability.
- The more that people experiment, the more options we'll have to consider.
- When information is important, let's aim to have the best of both worlds.