



Search

Search 124,426,951,803 web pages

[About Cuil](#) | [Preferences](#) | [Add Cuil to Firefox](#)

[Modified Privacy Policy](#) | © 2009 Cuil, Inc.

# Adventures In SearchLand

Valeria de Paiva  
July 2009  
PARC



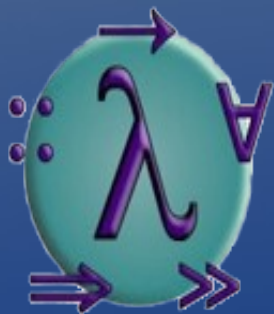
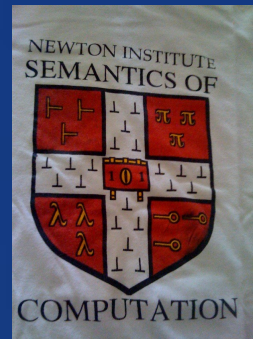
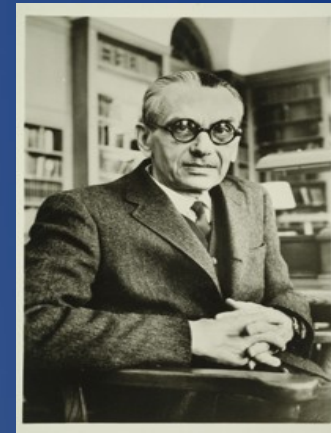
# Outline

- Personal background
- What is a search engine?
- How do they work?
- SearchLand?
- Cuil!
- Adventures...
- and Opportunities



# Yours truly...

- Pure mathematics in Cambridge
- Work on Category Theory
- Programming languages
- Natural language & KR in PARC
- Search...



# Search engines...

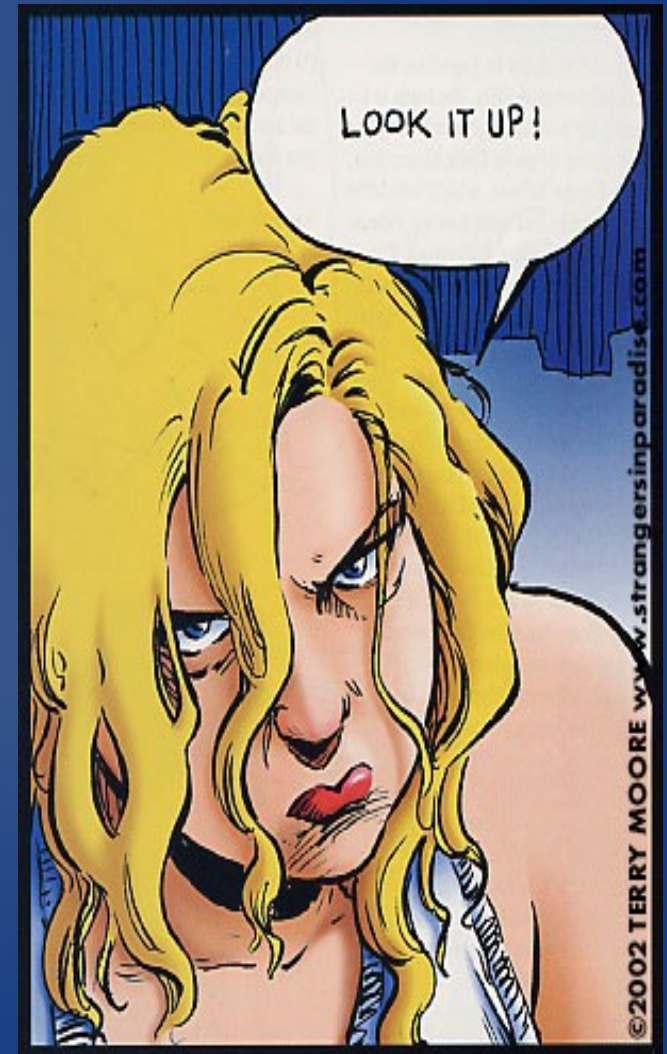
- Until last year my understanding of search engines was like my understanding of telephones or cars...
- I know **when** they're working and how to use them.
- I have no idea why or how they work...
- Assuming you're like this too, some tidbits...





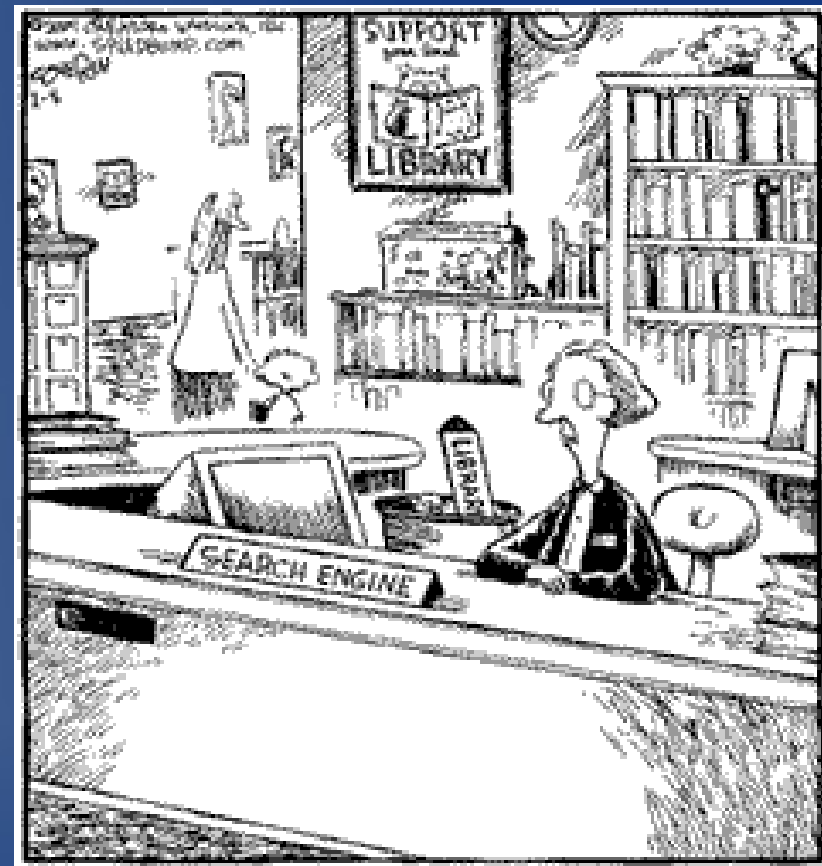
# Search Engines are like Librarians

- Have to have loads of documents a pesky user might want to see.
- Need to know the contents of the documents, to give the appropriate document.
- Need to aggregate the records of the contents of the documents in the *index*.
- When the user asks for a document, the librarian has to consult its index, decide on the most appropriate answers (the hits), find and deliver them in a *timely* and *pleasant* manner



# Metaphor continued...

- There is a building up step:  
collecting and indexing documents
- There is a serving up process:  
reading the query in, massaging it,  
finding the results, ranking results  
and serving results.
- These correspond to the modules of the search engine: crawler, indexer, query analyzer, finding and ranking algorithms, webserver magic



# Metaphor gone too far...

- Books don't arrive at a library in tens of thousands every day  
Search engines crawl the web all the time  
(and freshness is a real problem)
- Libraries get rid of books once a year  
Search engines would re-index every five minutes  
if they could
- Libraries simply hand off their goods,  
search engines differentiate themselves by how they deliver their goods





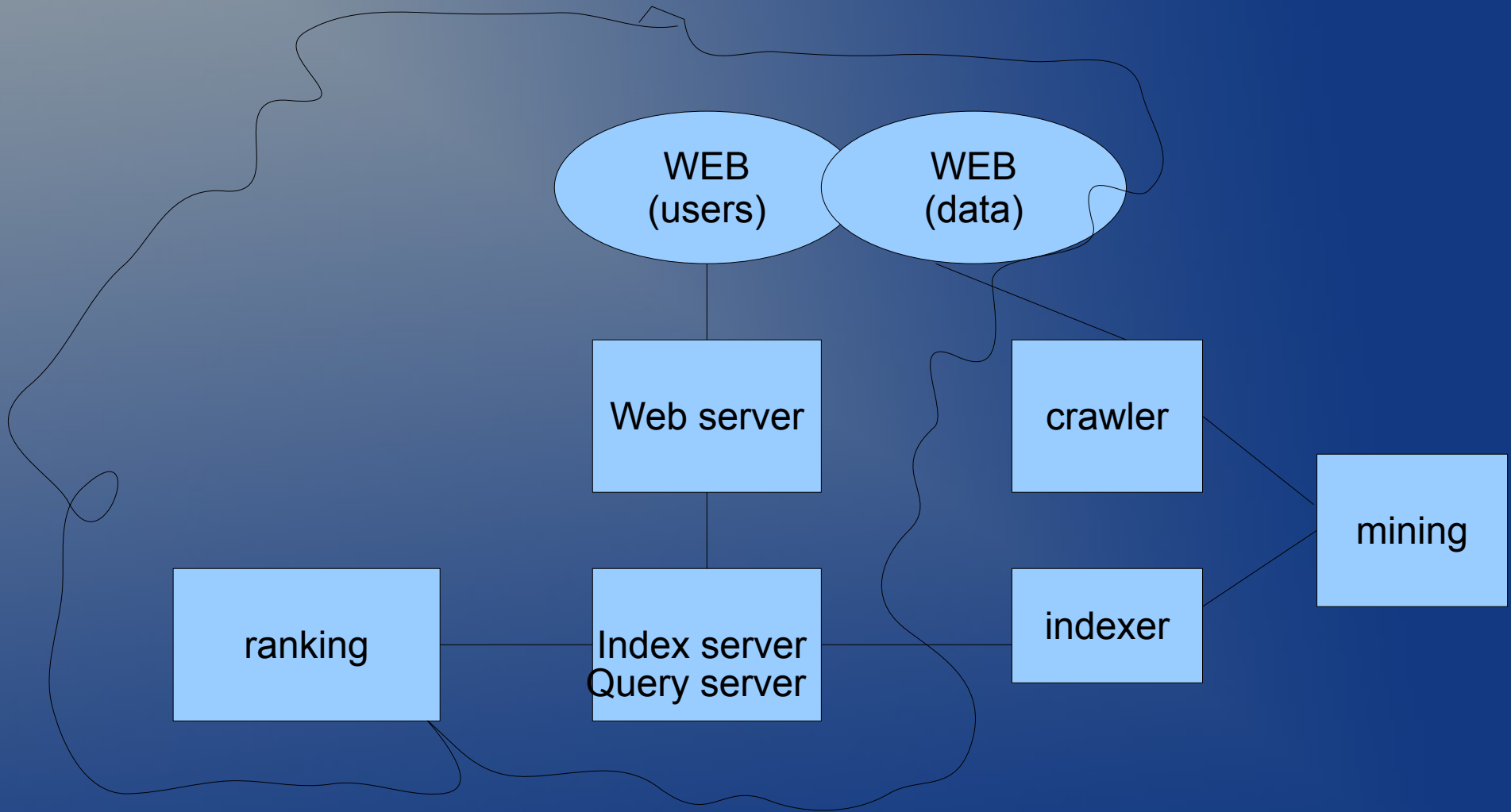
# Search Engine Basics

A search engine has modules

- Crawler
- Indexer
- Query analyzer
- Searcher
- Ranking
- Webserver

*Why writing your own search engine is hard*  
Patterson, ACM Q, 2004  
*Building Nutch: Open Source,*  
Cafarella and Cutting, 2004  
*Search technologies for the internet*  
Henzinger, Science. 2007

# Search Engine Scheme



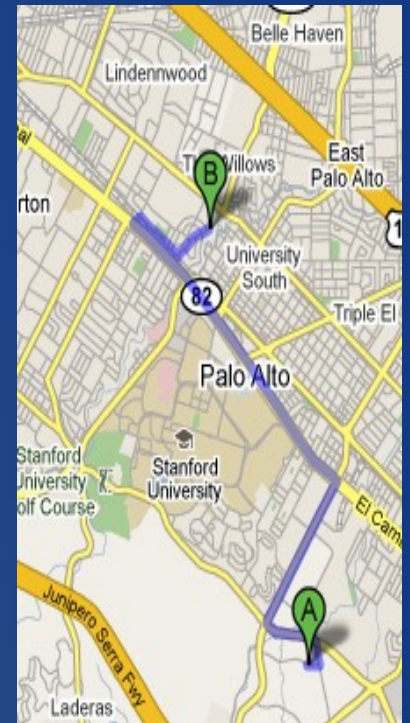
# SearchLand...

- So far, so good.
- Like Alice in the Wonderland in the Oxford meadows with her sister
- Then she follows the rabbit into the hole and things began to change..



# Getting there

- PARC: a big change from academia. There are things that you cannot tell your friends about your industrial research
- Timing is an art: you cannot publish too early, as IP has to be protected. Wait too much and there's nothing to publish.
- But PARC is still much closer to academia than I realized. It's research! It must become a product. Pretty soon. But it isn't one to begin with.



# Are we there yet?

- Start-up landscape is different: no offices, an open plan with individual desks and machines
- No book shelves, no work phones
- No four All Hands per year, one every week.
- Release of new code once a week usually more
- Life moves fast...





# SearchLand: Cool Cuil!

- How did I get there?

Anna Patterson and Tom Costello  
are friends of many years.

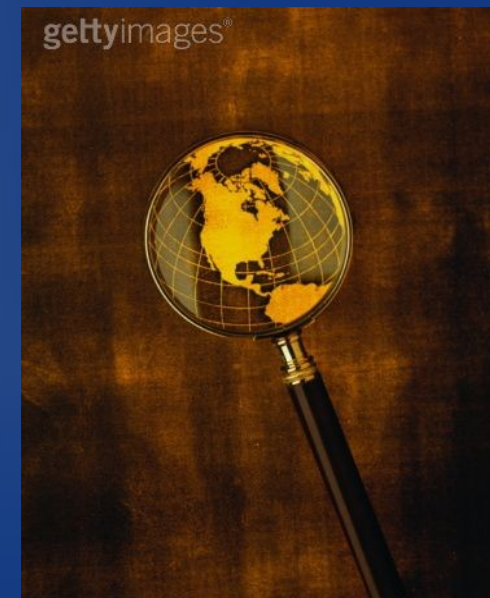
How did they get there?



- They did a search start-up called Xift in 1999. Then Anna designed, wrote and sold Recall—the largest search engine in 2004 to Google. Also architect of Google's TeraGoogle in early 2006.
- Tom worked in IBM on the prototype of WebFountain and on Storage Systems Strategy worldwide
- Then they decided to work together in Cuil

# The reasons for Cuil

- There are many search engines. But their results tend to be very similar. Are we seeing everything?
- Reports estimate we can see only 15% of the existing web. This is decreasing
- Probing the web is mostly **popularity based**. You're likely to see what others have seen before.  
But your seeing increases the popularity of what you saw, thereby reducing the pool of available stuff.
- Deep Web too?...



# The reasons for Cuil

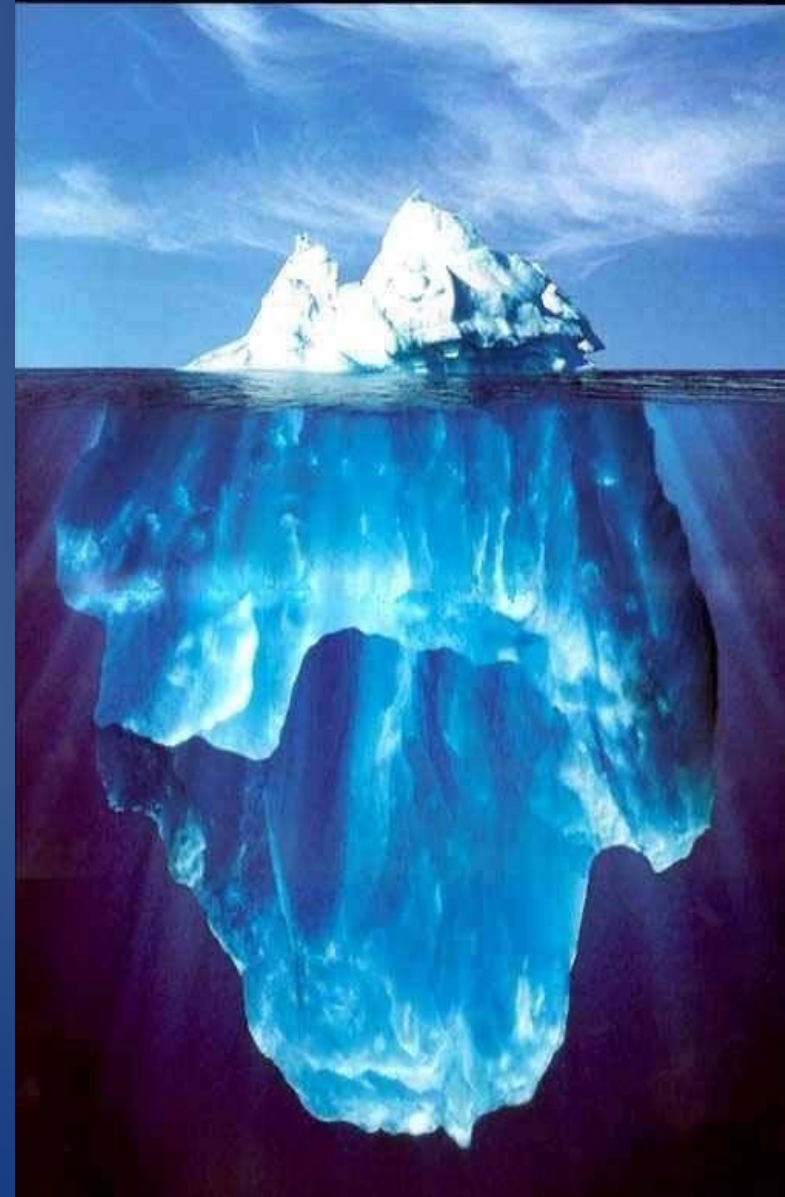
Much rubbish on the web.

Some say **all** we don't see is web rot: web spam, porn, mindless duplication of non-content...

Cuil says let's check it out, let's analyze contents of the pages.

People want to find information important to them, even when it's not popular.

[e.g. vanity search yields long lost brother]



# The reasons for Cuil

- Cost and natural resources
- Users don't pay directly for using search engines and their server farms
- But costs to the environment should be part of the equation
- Cuil can serve a bigger index using a small fraction of the number of machines
- Cheaper for the environment and for the company





# The reasons for Cuil



- Cuil doesn't need to know your search history and habits.
- So we don't.
- no names, no IP addresses, and no cookies
- Your search history is your business, not ours.



# The reasons for Cuil

- There is (too much) information on the web.
- Cuil 'organizes' the web so that you can find information that you didn't know you wanted..



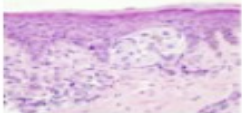
melanoma Search Preferences English

All Results Malignant Melanoma Ocular Melanoma Nodular Melanoma more... Melanoma Vaccine

**Melanoma**  
Wikipedia: Amelanotic (colorless or flesh-colored) **melanomas** do not have pigment and may not even be visible. Lentigo maligna, a superficial **melanoma** confined to the topmost layers of the skin (found primarily in older patients) is often described as a "stain" on the skin. Some patients with metastatic **melanoma**  
[en.wikipedia.org/wiki/Melanoma](http://en.wikipedia.org/wiki/Melanoma)

**Melanoma – skin cancer reviewed**  
Nodular **melanoma** (about 15% of cases) – This type of **melanoma** rises very rapidly and is the most aggressive type of skin cancer. From the beginning this type of **melanoma** grows vertically up and down. So there is a danger from the beginning that **melanoma** will spread in to inner tissues. This type of **melanoma**  
[melanoma.blogsome.com/](http://melanoma.blogsome.com/)

**MRF Home**  
To support medical RESEARCH for finding effective treatments and eventually a cure for **melanoma**. To EDUCATE patients and physicians about the prevention, diagnosis and treatment of **melanoma**. To act as an ADVOCATE for the **melanoma** community to raise the awareness of this disease and the need for a cure.  
[www.melanoma.org/](http://www.melanoma.org/)



# Organizing the web...

- Images can help.
- Longer snippets help.
- Tabs and categories show new stuff.

[Search](#) [Preferences](#) | [English](#) ▼

cuil

987,499 results for jane austen novels

### Allison Thompson

Novels, plays, and poems of the past are often fruitful sources for a dance historian to learn more about the attitudes towards or conventions of the dance. Despite the fact that **Austen** never describes a specific dance, her **novels** are particularly useful sources, as we have seen.

[www.jasna.org/persuasions/on-line/v...](http://www.jasna.org/persuasions/on-line/v...)



### Jane Austen novels on Lists of Bests

The Watsons: A Fragment (Jane Austen Library, Vol 4)




### Rowlinson

**Austen, Jane.** The Novels of Jane Austen.

**Austen, Jane.** Jane Austen's Letters to her Sister Cassandra and Others. 5, 30). **Jane Austen**, on the other hand, was less interested in doing good for the masses than in getting to know individuals. Not until her last work did **Jane Austen** reflect the changing mores of the times: Mr.

[www.jasna.org/persuasions/on-line/v...](http://www.jasna.org/persuasions/on-line/v...)



### Reception history of Jane Austen

### Explore by Category

#### Television Programs Based On Jane Austen Novels

[Pride and Prejudice](#), [Persuasion](#), [Sense and Sensibility](#), [Northanger Abbey](#), [Mansfield Park](#)

#### Novels By Jane Austen

[Pride and Prejudice](#), [Persuasion](#), [Sense and Sensibility](#), [Northanger Abbey](#), [Mansfield Park](#), [Lady Susan](#), [The Watsons](#), [Sanditon](#), [Love and Freindship](#)

#### Films Based On Jane Austen Works

[Pride & Prejudice](#), [Clueless](#), [Bridget Jones's Diary](#)

# Organization is fundamental

- Definitions —easier then going to a dictionary
- Timelines - show the evolution of your concept
- Maplines — new connections
- Videos from Hulu, maps from Mapquest.

Firefox File Edit View History Bookmarks Tools Window Help

solar eclipses - Cuil

http://www.cuil.com/search?q=solar+eclipses

Most Visited Getting Started Latest Headlines B Workqueue QPS-AT-B C Workqueue mining Gmail - FW: Talk: Ad... Cuil Trac - Trac ibourbaki / FrontPage D workqueue staging WorkqueueB

Search Google Gmail Event Picsear Alice's solar... Cuil Sea Dr Wilfr SQIG at IT Index AYSO S...


solar eclipses Search Preferences English

cuil

11,982 results for solar eclipses


### Solar eclipse

Wikipedia: The next anticipated simultaneous occurrence of a **Solar eclipse** and a transit of Mercury will be on July 5, 6757, and a **Solar eclipse** and a transit of Venus is expected on April 5, 15232. Only 5 hours after the transit of Venus on June 4, 1769, there was a total solar eclipse  
[en.wikipedia.org/wiki/Solar\\_eclipse](http://en.wikipedia.org/wiki/Solar_eclipse)




### Solar Eclipses

In this section we consider **solar eclipses** and in the next we discuss **lunar eclipses**. Geometry of **Solar Eclipses** The geometry associated with **solar eclipses** is illustrated in the following figure (which, like most figures in this and the next section, is illustrative and not to scale).  
[csep10.phys.utk.edu/astr161/lect/time/eclipses.html](http://csep10.phys.utk.edu/astr161/lect/time/eclipses.html)




### Curious About Astronomy? Lunar and Solar Eclipses

Though a total **solar eclipse** may be seen more than once a year on Earth, from a given spot on the planet these events are almost as rare as they are spectacular. The relative motions of the Earth and the Moon cause **solar eclipses**  
[curious.astro.cornell.edu/eclipses.php](http://curious.astro.cornell.edu/eclipses.php)




### IAU WWW Home Page

Eye Safety and Solar Eclipses, by Ralph Chou

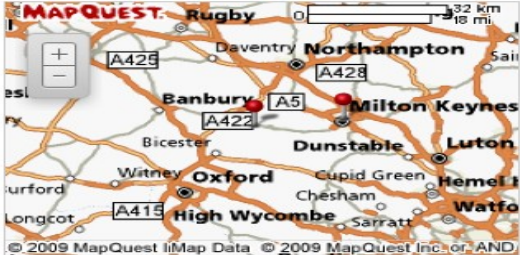


### Solar eclipses, Poland, Sosnowiec

The Sun can be viewed safely only during the few brief



### Mapline for Solar Eclipses



© 2009 MapQuest iMap Data © 2009 MapQuest Inc. or AND  
Enlarge For More

On this map: [United Kingdom](#), [Open University](#), [Assyrian eclipse](#), [Gustavia](#), [American Academy of Arts](#)  
[more](#)

There were no Total Solar Eclipses visible from the United Kingdom between 1724 and 1925.

### Explore by Category



# Adventures

- There are many.
- Talking about three:
- Launch!
  - And blogsphere...
- Timelines
- Languages



# Launching a product

- It's different from anything I had ever done before.
- Launched July 28th, less than three months from my start.
- Hoped for a “soft” launch in the middle of the summer..
- Unbelievable “flood” of interest





# After the hype, the blogs...

- Hadn't realized how much the valley runs on blogs
- Didn't know about tech celebrities or valleywag...
- Had no idea how many people make a living doing SEO
- Unbelievable that people went to the trouble of “faking” bad results.



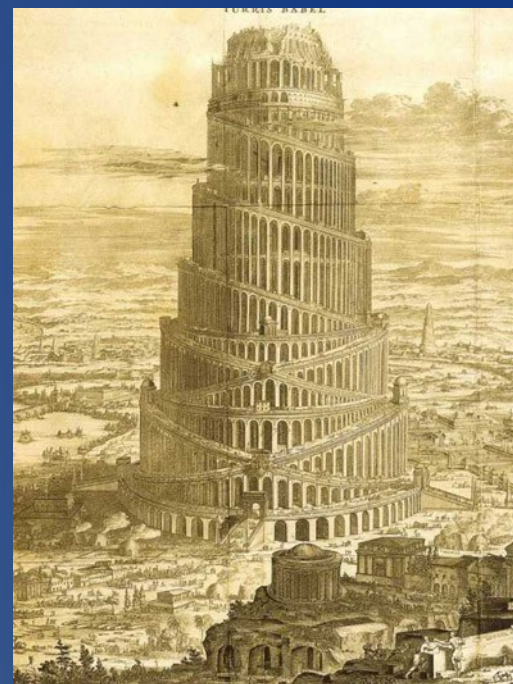
# Timelines

- Launched in March'09
- Dynamic timelines, not pre-computed for a few subjects
- Project completed in less than six weeks
- Too many? Algorithm still needs improvement
- But a personal battle won...



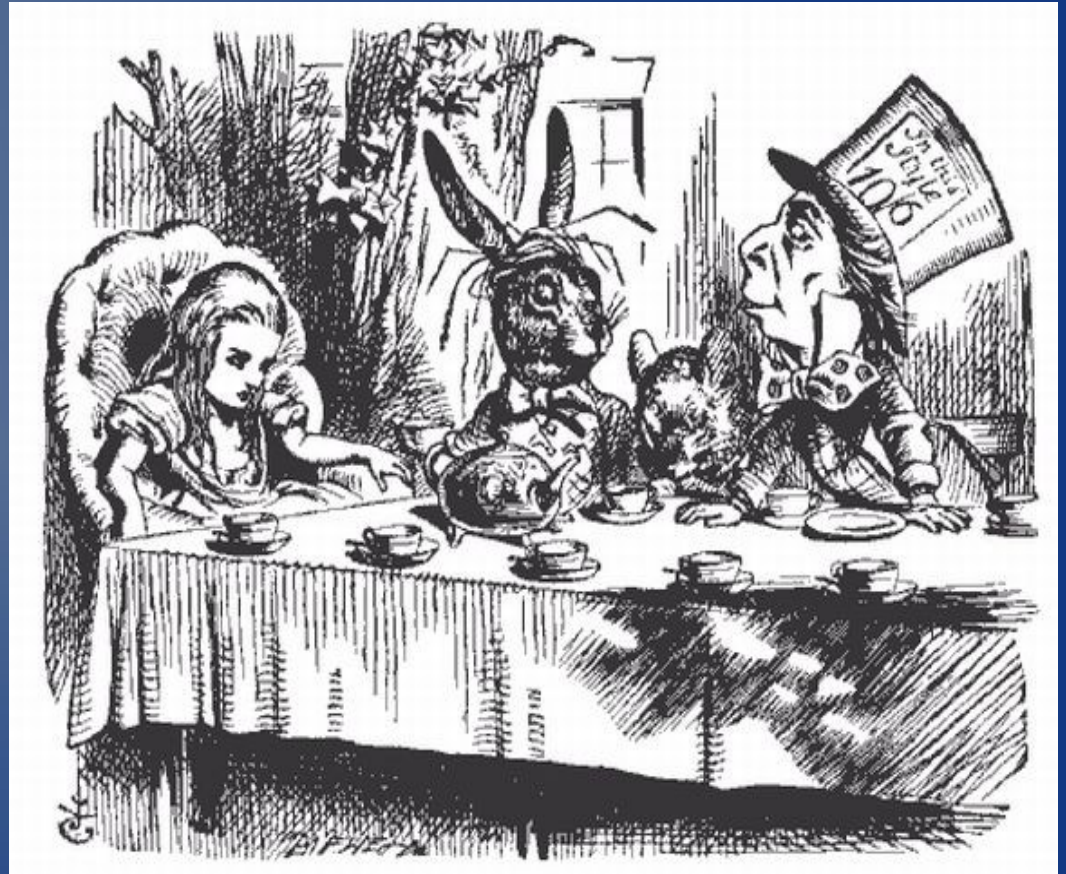
# Multiple Languages

- Launched in May'09
- Infra-structure in place, took less than a month to release
- Seven languages so far
- Evaluation hardly started
- But loads of offers to help
- All of this organization with a team of less than thirty...



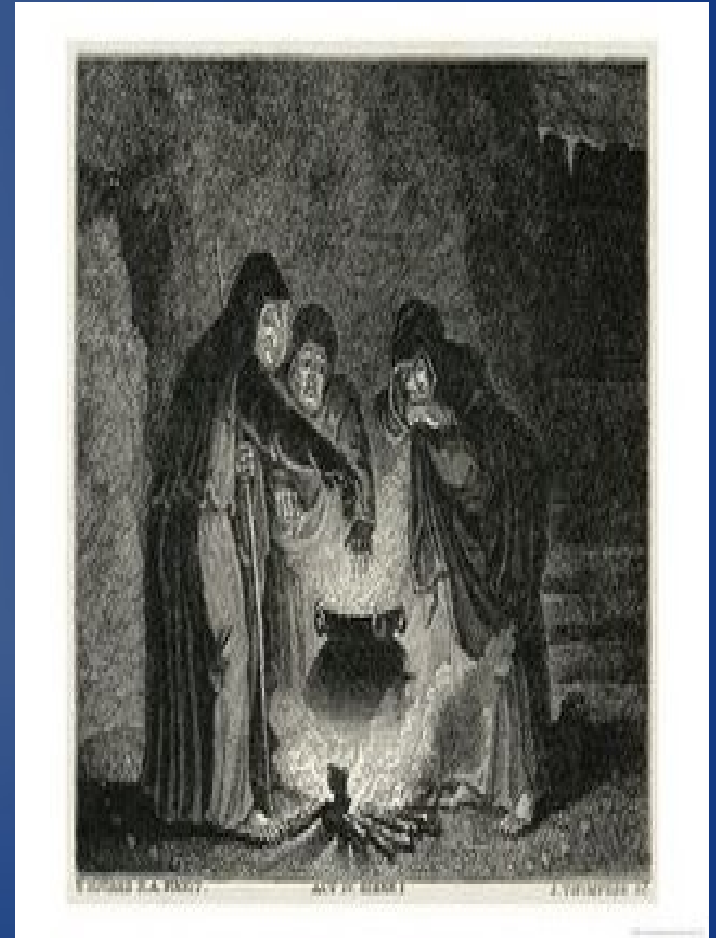
# Opportunities...

- There are many.
- Quality evaluation
- Relevance improvement
- More services...



# More Opportunities...

- Three banes of my life:
- Spam, spam, spam
- (Economics of) malware
- Attacking pornography





# Summing up

- Life in Searchland is very different
- And lots of fun!
- As Patterson says in “Why Writing your Own Search Engine is Hard”, AM Q 2004,  
“[...] once the search bug gets you, you'll be back. The problem isn't getting any easier, and it needs all the experience anyone can muster.”



And ever, as the story drained  
The wells of fancy dry,  
And faintly strove that weary one  
To put the subject by,  
“The rest next time--” “It is next time!”  
The happy voices cry.

Lewis Carroll -- Proem

Thank You!



Search

Search 124,426,951,803 web pages

[About Cuil](#) | [Preferences](#) | [Add Cuil to Firefox](#)

[Modified Privacy Policy](#) | © 2009 Cuil, Inc.