



UNITED STATES

National Library of Medicine

Integrating Semantic Search,  
Federated Search and  
Biomedical Knowledge Base  
Technologies for  
Improved Services by Medical Libraries

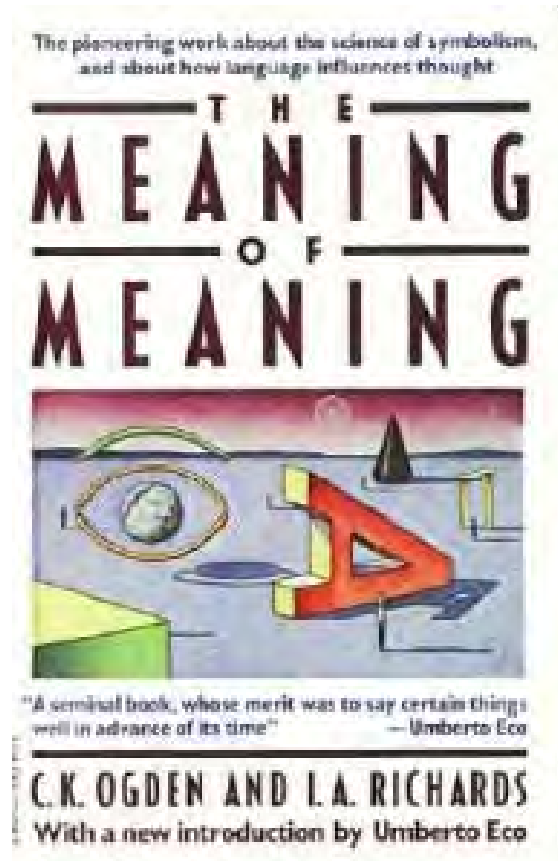
EAHIL 2010

Tamas Doszkocs, Ph.D.  
Computer Scientist  
[doszkocs@nlm.nih.gov](mailto:doszkocs@nlm.nih.gov)

“What is semantic search?”

or

## The Meaning of Meaning



“He who knows does not speak,  
he who speaks does not know”

Lao Tse

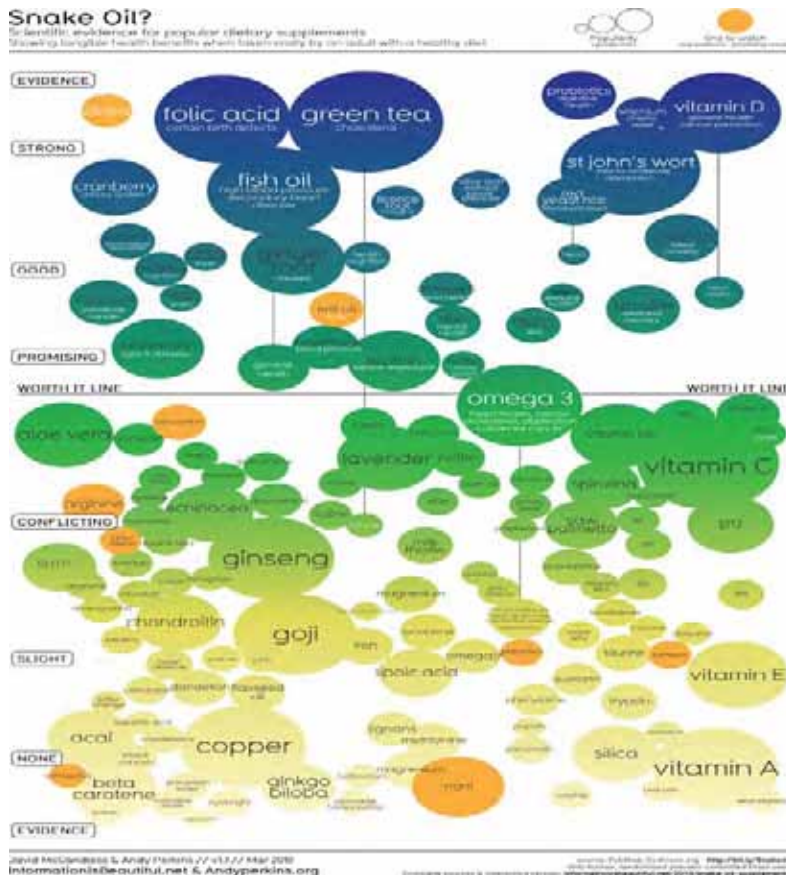
"It depends on what the meaning of  
the word 'is' is"

Bill Clinton

“the answer, my friend, is blowing  
in the wind”

Bob Dylan

# What is semantic search?



“**semantic search is a search or a question or an action that produces **meaningful results**, even when the retrieved items contain **none of the query terms**, or the search involves **no query text at all**”**

(my definition)

# Thinking of the Web and Semantic Search

“Wholly new forms of **encyclopedias** will appear, ready made with a **mesh** of **associative trails** running through them, ready to be dropped into the **MEMEX** and there **amplified**”

Vannevar Bush “As We May Think” Atlantic Magazine

JULY 1945

# Trends in Searching 2009-2010

- *The Web* is the *Memex*
- *Vertical Search*
- *Universal Search*
- *Discovery Search*
- *Social Search*
- *Real Time search*
- ***Semantic Search***
- *Semantic Web*
- *Connected Mobility*
- *Focus on Consumers*
- *Information Democracy*
  - *Social content*
  - *Social search*
    - *Social interaction*
- *Information Monopolies*
- *manipulating text,*  
*context, meaning,*  
*people*

# Semantic Search Engines Mean Well

- *A long history ...*
- *At a loss for words*
- *Ranking Results by Relevance*
- *Searching for Meaning*
- *Google and the Rest*
- *“Power to the People”*
- *“Pure” Semantic Search Engines*
- ***Specialized Semantic Search Engines***
- *Demos*

# Semantic Searching: a long history

- *Libraries as Knowledge Bases*
- ***Librarians as Search Engines***
- *The Web as the Knowledge Base*
- ***Search Engines as Librarians***

# Toward Meaning-Based Search

- ***Understanding content***
- ***Understanding context***
- ***Understanding people***
- ***Semantic Search vs. The Semantic Web***
- ***Meaningful results***
- ***Web 1.0 linking pages***
- ***Web 2.0 linking content and people***
- ***Web 3.0 linking data and people and applications***
- ***Web 4.0 ; inferences, conjectures, the pursuit of happiness ☺ ?***



# Search Engines: at a loss for words

“In the beginning was the Word”

- *WebCrawler (1994)*
- *AltaVista (1996)*
- *InfoSeek (1996)*
- *“bag of words”*
- *Stopwords*
- *Boolean logic*
- *No linguistics*
- *Meaningless function words?*
  - *“to be or not to be”*
  - *What is it all about?*
  - *Why me?*
- *Ignoring syntax*
  - *“state police” vs.*
  - *“police state”*

# Ranking Results by Relevance

“all animals are equal but some animals are more equal than others”



*Science Citation Index*

**Eugene Garfield** (1955, 1961)

***Google Page Rank***

**Larry Page** (1996)

*Semantic Ranking*

**Colucci et al.** (2006)

# Searching for Meaning

the difference between “sciatica” **the word** and “sciatica” **the pain**

- ***understanding language***

- *polysemy*
- *synonymy*
- *intensions*
- *context*
- *disambiguation*
- *meaning*
- *personalization*

- ***Natural Language Processing and contextual understanding in searching***

- *Semantic Resources*
- *Metacrawler* (1994)
- *WebLine* (1994)
- *NorthernLight* (1996)
- *TotalAccess* (1999)
- *AllPlus* (2004)

# Semantics in the Search Engines

- evolution by *trial and error*
- *incremental gains*
- *internal research*
- *personalization*
- *mobility*
- *Interactivity*
- *acquisitions*
- *competition*
- **Bingoogle**
- *Related Searches*
- *Categories*
- *Context*
- *Mashups*
- **Answers**
- *Timelines*
- *“Something different”*
- *Visualizations*

Of the People  
By the People  
For the People

- *web 2.0*
- *social content*
- *tagging*
- *folksonomies*
- *blogging*
- *tweeting*
- *chatroulette*
- ***Community***
- ***Creating***
- ***Indexing***
- ***Cataloging***
- ***Searching***
- ***Sharing***

# “Pure” Semantic Search Engines

- semantics from the ground up
- *understanding the Web*
  - *people and their intent*
  - *content and quality*
  - *data and attributes*
  - *entities and relationships*
  - *Context and meaning*
- *creating*
  - *actionable information*
  - *dynamic applications*
- meaningful search results
- better quality results
- better relevance
- better presented results
- better currency of results
- *better tailored results*
- *better info streams*
- *better explore/discover*
- *better learning*
- *meaningful interactions*

# Specialized Semantic Search Engines

- *domain knowledge*
- *matching people and needs*
- *improving diverse applications*
  - *real time news*
  - *finding jobs*
  - *recommendations for movies, music, goods, etc.*
  - *trend trackers*
  - *health*
  - *computable knowledge*
  - *semantic Facebook*
- ***Semantic resources and tools***
- ***Related searches/topics***
- ***Related items***
- ***Semantic mapping***
- ***Semantic synthesis***
- ***Semantic/linguistic annotations***
- ***Clustered search***
- ***Faceted search***
- ***Answers***

# Semantic Search Engines for Health

- *Rich Semantic Resources in Biomedicine*
- *Small Number of Important Facets*
  - *Diseases and conditions*
  - *Causes*
  - *Tests and diagnosis*
  - *Treatments and procedures*
  - *Drugs and side effects*
  - *Alternative medicine*
  - *Life style and prevention*
- *Participatory Medicine*
- *E-patients*
- *first, second and third opinions* on and off the Web
  - *Health Professionals*
  - *Friends*
  - *Family*
  - *The “long tail” of health concerns*
- ***Informed Personal Health Decisions***



# Trusted Health Information is vital for

- dealing with health problems,
- promoting healthy behavior,
- making healthy decisions and
- for overall well being
- However, as Mark Twain put it:  
**Be careful of reading health books.  
You may die of a misprint!**

# examples of health search engines with Semantic Search Capabilities

- [HealthLine](#)
  - uses own taxonomy of > 250,000 health terms
  - thousands of Indian doctors and pharmacists
- <http://www.everydayhealth.com/>
  - Meta-data clusters
  - Topical clusters
  - Second most popular site after WebMD
- <http://righthealth.com/>
  - federated search engine
  - taxonomy of several million nodes
  - organized into a graph by
  - using a combination of human operators and algorithms
- <http://MedStory.com>
  - high-level categorizations or popular URLs
  - Purchased by Microsoft
  - <http://health.msn.com>

# examples of health search engines with Semantic Search Capabilities

- <http://www.semanticmedline.com/>
  - Semantic NLP "understands" word and phrase meanings within context
- <http://skr3.nlm.nih.gov/SemMedDemo/>
  - [Research prototype](#)
  - [summarizes MEDLINE citations returned by a PubMed search](#)
  - [Natural language processing is used to analyze salient content](#)
- <http://healthbase.netbase.com/>
  - [Based on language understanding](#)
  - [Surfaces facts, events, behaviors and connections among them"](#)
  - combines classical keyword-based Web search with text-mining and ontologies
- <http://healthmash.com/>
  - Best-of-class semantic search engine
  - Powered by an automatically generated
    - Health Knowledge Base

# the best health search engines

offer information that is

- » **Reliable**
- » **Relevant,**
- » **Recent and**
- » **Related meaningfully to the search topic or user action**

# HealthMash

anatomy of a semantic health search and discovery engine

- <http://healthmash.com/> is an innovative next generation **semantic health search** engine, currently in public beta
- HealthMash developers have been working on NLM and NIH R & D projects for over 5 years
- HealthMash utilizes a pragmatic mix of **natural language processing tools**, semantic engineering techniques and **multiple knowledge sources**, including a proprietary **Health Knowledge Base**, to achieve both high precision and relevancy in its search results

# HealthMash Demo

The screenshot shows a Mozilla Firefox browser window displaying the HealthMash website. The browser's address bar shows the URL <http://test.polymeta.com/search/u7/searchfr.jsp?sc=900&notahonynaccess=true>. The website features a green header with navigation links: [Trusted Health Information](#), [Health News](#), [Video](#), [Image](#), [Drugs and Substances](#), [Clinical Trials](#), [Integrative Medicine](#), [Web](#), and [Articles](#). Below these are links for [Books](#), [Blogs](#), and [Twitter](#). The main content area has a blue cross logo with the text "Health Mash" and a search input field. To the right of the input field are buttons for "Search", "Clear", and "Select Sources". Below the search area, the text "Revolutionary Health Knowledge Base and Semantic Search Engine" is displayed. At the bottom of the page, there is a "WebQb" logo. The browser's taskbar at the bottom shows several open applications, including "Start", "Friday", "icnl\_update", "ICPL Pocket Program K (...)", "aand", "Seeking Health ICPL 200...", and "HealthMash - Medical...". The system tray on the right shows the time as 3:08 PM.

# The **Explore and Discover** data comes from the **Health Knowledge Base**

- The *Health Knowledge Base* is automatically generated from trusted health content sites and diverse knowledge sources, such as **MeSH** and **UMLS**. HealthMash also utilizes the **Web** itself as a **data base**.
- The *Health Knowledge Base* contains explicit knowledge about
  - Health Concerns, Causes, Signs and Symptom, Tests, Procedures, Treatments, Drugs and Substances and Alternative, Complementary and Integrative Medicine
- The *Health Knowledge Base* is available via a **web service API**
- The *Health Knowledge Base* facilitates **exploration** and **discovery**

# UMLS: Strengths, Weaknesses and Potential

- MeSH vs. UMLS Scale
- Missing concepts
  - Aichmophobia; pca3 test
- Missing semantic types
  - Alternative and complementary medicine
- Missing NLM applications ):
- Robust R&D: NLM Semantic MEDLINE



# Semantic Knowledge Bases and Tools in HealthMash

- Semantic knowledge sources utilized by HealthMash
  - [MESH](#) (the [Medical Subject Headings Thesaurus](#) of the NLM) ,
  - [UMLS](#) (the [Unified Medical Language System](#) of NLM/NIH) and other semantic data repositories and
  - The [Web](#) as a knowledge base
  - **The Health Knowledge Base**
    - the **Health Knowledge Base** is the most important **semantic resource** in **HealthMash**
- Proprietary [Natural Language Processing](#) tools
  - Lexical/morphological and orthographic tools
  - Syntactic tools and
  - **Semantic tools**

behind the scenes in  
**Explore and Discover**

- **Semantic Engineering**
- **Heuristics**
- Proprietary **Semantic Knowledge Base**
- Proprietary **Semantic Search Algorithms**
- Proprietary **Semantic Ranking**

## HealthMash utilizes the following NLP resources (these tools can also be licensed)

- **PolyDictionary** (English, medical and scientific dictionary)
- **PolySpell** (English, medical and scientific spell checkers)
- **PolyTagger** (part-of-speech tagger)
- **PolyPhraser** (phrase parser)
- **PolySearch** (intelligent concept search engine)
- **PolyCluster** (search result clustering engine)
- **PolyMeta** (federated search and discovery engine)

# Putting it all together

- the **automatically generated** and automatically enhanced **Health Knowledge Base** is the key value-added semantic component of **HealthMash**
- the proprietary **semantic search**, **semantic processing** and **semantic ranking** technologies utilized in HealthMash contribute to **better search results**
- HealthMash can be integrated into any search engine via a **web service** (or **API**)

## IN SUMMARY

**HealthMash** combines **vertical** semantic search of Trusted Health Information, **mashups** (Health News, Videos etc), **Semantic Relations** in context for **Exploration and Discovery** and **Table of Contents** and **Topic Clusters** for *drill down* in search results and *dynamic query modification*

# Semantic Knowledge: HealthMash

The screenshot displays the HealthMash website interface. At the top left is the HealthMash logo. A navigation bar contains tabs for 'Trusted Health Information', 'Health News', 'Videos', 'Images', 'Drugs and Substances', 'Clinical Trials', 'Integrative Medicine', 'Web', 'Articles', 'Books', 'Blogs', and 'Twitter'. A search bar contains the text 'plantar warts', with 'Search', 'Clear', and 'Feedback' buttons to its right. Below the search bar is a green 'Explore and Discover' section with five columns of related concepts, each with a plus sign in the top right corner:

Related Concepts	Health Concerns	Tests and Treatments	Drugs and Substances	Alternative Medicine
<ul style="list-style-type: none"><li>verruca</li><li>verruca vulgaris</li><li>plantar wart</li><li>mosaic plantar wart</li></ul>	<ul style="list-style-type: none"><li>common wart</li><li>venereal wart</li><li>human papillomavirus</li><li>corns and calluses</li></ul>	<ul style="list-style-type: none"><li>cryotherapy</li><li>cantharidin</li><li>wart removal</li><li>duct tape</li></ul>	<ul style="list-style-type: none"><li>salicylic acid</li><li>imiquimod</li><li>trans-ver-sal</li><li>wart away</li></ul>	<ul style="list-style-type: none"><li>vinegar</li><li>tea tree oil</li><li>vitamin a</li><li>podophyllin</li></ul>

# Semantic Knowledge: MedStory



Search:

Results for **plantar warts**

---

 **Information that Matters™: click below to refine your search**

Conditions

**Plantar Warts**  
Warts



# Semantic Knowledge:

## Google Health

Google health   [Browse all health topics](#)

**Warts**

[Overview](#)  
[Treatment](#)  
[Causes](#)  
[Tests & diagnosis](#)  
[Prevention](#)  
[Lifestyle](#)  
[Complications](#)  
[When to contact a doctor](#)

**U.S. National Library of Medicine**  
Find this content and more from trusted sources.  
[MedlinePlus®](#)

**Overview**

Warts are small, usually painless growths on the skin caused by a virus. They are generally harmless. However, warts can be disfiguring and embarrassing, and occasionally they itch or hurt (particularly on the feet).

The different types of warts include

- Common warts usually appear on the hands, but can appear anywhere.
- Flat warts are generally found on the face and forehead. They are common in children, less so in teens, and rare in adults.
- [Genital warts](#) are usually found in the genitalia, in the pubic area, and in the area between the thighs, but they can also appear inside the vagina and anal canal.
- Plantar warts are found on the soles of the feet.
- Sublingual and perioral warts appear under and around the tongue in children.

**Symptoms**

### Search Trends

People who searched for **Warts** also searched for:

### Conditions

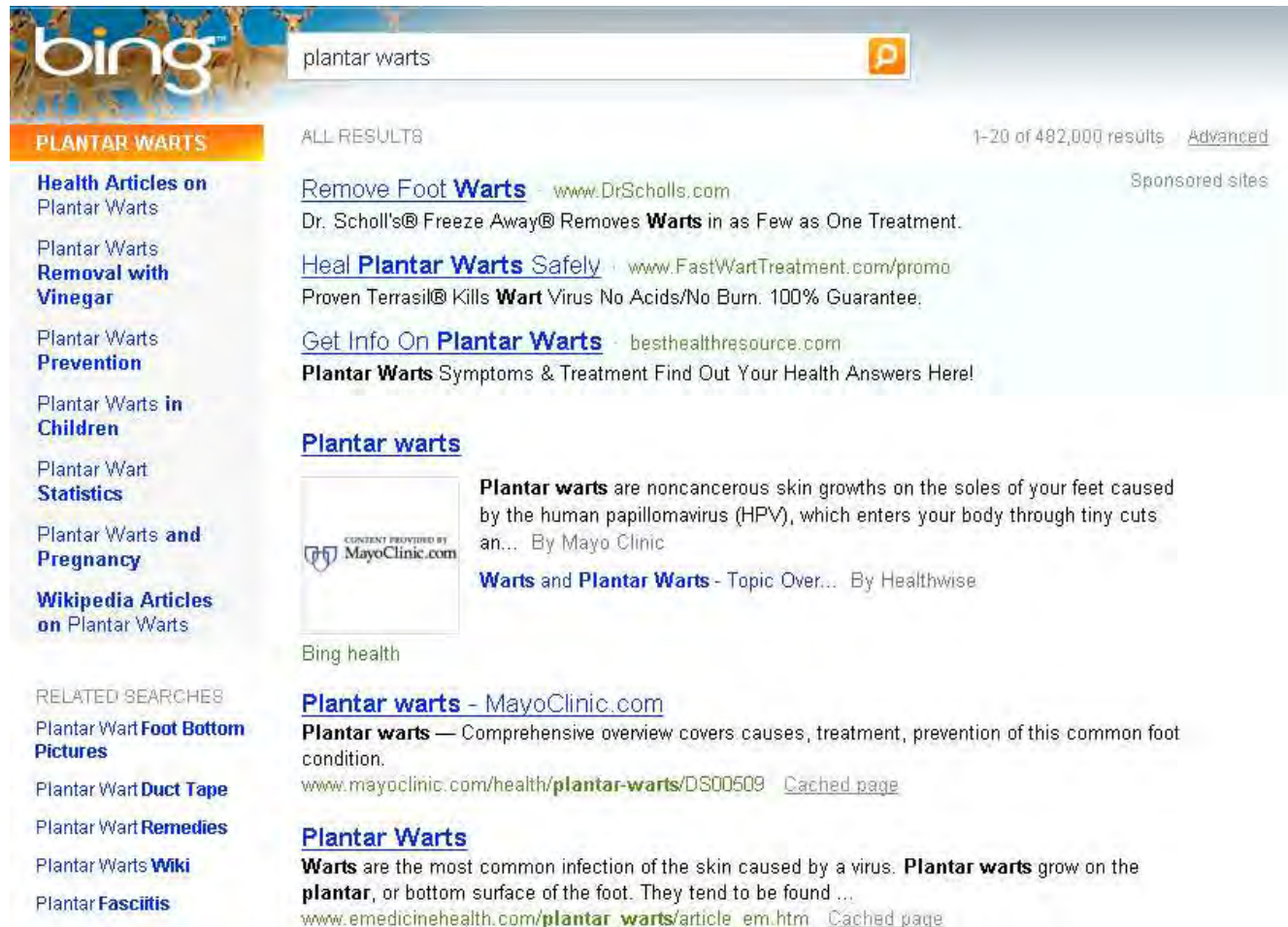
1. [genital warts](#)
2. [human papillomavirus infection](#)
3. [plantar warts](#)
4. [acute lymphocytic leukemia](#)
5. [pregnancy](#)
6. [moles](#)
7. [infection](#)
8. [acne](#)

### Symptoms

1. [headache](#)
2. [weight loss](#)
3. [fever](#)
4. [cough](#)
5. [fatigue](#)
6. [back pain](#)
7. [sore throat](#)
8. [diarrhea](#)



# Semantic Knowledge: Bing



The image shows a Bing search results page for the query "plantar warts". The search bar at the top contains the text "plantar warts" and a magnifying glass icon. Below the search bar, the Bing logo is visible on the left. The search results are displayed in a list format. The first result is a sponsored link from Dr. Scholl's, titled "Remove Foot Warts" and "Dr. Scholl's® Freeze Away® Removes Warts in as Few as One Treatment." The second result is from FastWartTreatment.com, titled "Heal Plantar Warts Safely" and "Proven Terrasil® Kills Wart Virus No Acids/No Burn. 100% Guarantee." The third result is from besthealthresource.com, titled "Get Info On Plantar Warts" and "Plantar Warts Symptoms & Treatment Find Out Your Health Answers Here!". Below the search results, there is a section for "Plantar warts" with a small image placeholder and a description: "Plantar warts are noncancerous skin growths on the soles of your feet caused by the human papillomavirus (HPV), which enters your body through tiny cuts an... By Mayo Clinic". There is also a link to "Warts and Plantar Warts - Topic Over... By Healthwise". At the bottom, there is a section for "Plantar warts - MayoClinic.com" with a description: "Plantar warts — Comprehensive overview covers causes, treatment, prevention of this common foot condition." and a link to "www.mayoclinic.com/health/plantar-warts/DS00509" with a "Cached page" link. On the left side of the page, there is a sidebar with various links related to plantar warts, including "Health Articles on Plantar Warts", "Plantar Warts Removal with Vinegar", "Plantar Warts Prevention", "Plantar Warts in Children", "Plantar Wart Statistics", "Plantar Warts and Pregnancy", "Wikipedia Articles on Plantar Warts", "RELATED SEARCHES", "Plantar Wart Foot Bottom Pictures", "Plantar Wart Duct Tape", "Plantar Wart Remedies", "Plantar Warts Wiki", and "Plantar Fasciitis".

bing

plantar warts

PLANTAR WARTS


ALL RESULTS 1-20 of 482,000 results [Advanced](#)

[Remove Foot Warts](#) - [www.DrScholls.com](#) Sponsored sites  
Dr. Scholl's® Freeze Away® Removes **Warts** in as Few as One Treatment.

[Heal Plantar Warts Safely](#) - [www.FastWartTreatment.com/promo](#)  
Proven Terrasil® Kills **Wart** Virus No Acids/No Burn. 100% Guarantee.

[Get Info On Plantar Warts](#) - [besthealthresource.com](#)  
**Plantar Warts** Symptoms & Treatment Find Out Your Health Answers Here!

[Plantar warts](#)

 **Plantar warts** are noncancerous skin growths on the soles of your feet caused by the human papillomavirus (HPV), which enters your body through tiny cuts an... By Mayo Clinic

[Warts and Plantar Warts](#) - Topic Over... By Healthwise

Bing health

[Plantar warts - MayoClinic.com](#)  
**Plantar warts** — Comprehensive overview covers causes, treatment, prevention of this common foot condition.  
[www.mayoclinic.com/health/plantar-warts/DS00509](#) [Cached page](#)

[Plantar Warts](#)  
**Warts** are the most common infection of the skin caused by a virus. **Plantar warts** grow on the **plantar**, or bottom surface of the foot. They tend to be found ...  
[www.emedicinehealth.com/plantar\\_warts/article\\_em.htm](#) [Cached page](#)

Health Articles on Plantar Warts

Plantar Warts Removal with Vinegar

Plantar Warts Prevention

Plantar Warts in Children

Plantar Wart Statistics

Plantar Warts and Pregnancy

Wikipedia Articles on Plantar Warts

RELATED SEARCHES

Plantar Wart **Foot Bottom Pictures**

Plantar Wart **Duct Tape**

Plantar Wart **Remedies**

Plantar Warts **Wiki**

Plantar **Fasciitis**

# Integrating the Health Knowledge Base via a **web service** into a **search** or **federated search** application

The screenshot displays the NIH Library search interface. At the top, the NIH logo and the tagline "First Step - Where Searching Begins" are visible. Below this is a navigation bar with tabs for "Basic Search", "News", "Video", "Evidence-based Medicine", "Academic Web", "Alternative Medicine", "Animal Studies", and "Drugs and Chemicals". A search bar contains the text "bipolar disorder" and a "Search" button. To the right of the search bar are links for "Select Sources", "Show hits", and "Feedback".

Below the search bar is an "Explore and Discover" section with five categories, each with a plus sign icon:

- Related Concepts**: bipolar disorder, mania, manic, manic depression
- Health Concerns**: depression, major depressive disorder, suicide, schizophrenia
- Tests and Treatments**: psychiatric status rating, psychotherapy, psychological tests, family therapy
- Drugs and Substances**: valproate, valproic acid, carbamazepine, epilol
- Alternative medicine**: omega-3 fatty acids, melatonin, flaxseed and flaxseed oil

Below this section, it states "115 Results for bipolar disorder...".

The main search results area is titled "Basic Search Results" and includes a pagination control "1 | 2 | 3 | 4 | 5 | 6 | Next >". On the left, there are filters for "Topics", "Date", "Author", and "Source". The "All Results (115)" section shows a list of results, including "bipolar disorder... (59)", "Searching for New Genetic Risk Factors ... (6)", "patients with bipolar disorder (5)", "concurrent symptoms in inter-episode bipolar disorder (4)", "More", "patients with bipolar... (20)", and "Searching for New Genetic Risk Factors...".

The main results list includes:

- Check all results - Send checked results
- [Support for a bipolar affective disorder susceptibility locus on chromosome 12q24.3](#)  
Psychiatric genetics - Apr 20, 2010  
PubMed
- [Sleep, illness course, and concurrent symptoms in inter-episode bipolar disorder](#)  
Eidelman, Polina; Talbot, Lisa S; Gruber, June, et al. - J Behav Ther Exp Psychiatry - 2010  
 [Show more with this title] Web of Science - NIH
- [Hippocampal volumes in adults with bipolar disorder](#)  
Sachdev, P., Ivanovski, B, Chen, X, Malhi, G.S, Javadapour, A, Wen, W - Journal of Neuropsychiatry and Clinical Neurosciences 22 (1), pp. 55-62 - 2010

On the right side, there are two sections: "News" and "Video". The "News" section features a link to "Psychiatric Drugging of Infants and Toddlers in the US - Part 1" with a summary: "The number of children in this age group diagnosed with bipolar disorder also doubled over the last decade, Reuters reported. ...". The "Video" section contains two empty video player boxes and a "More" link.

# Distributing content, search and librarians

- From **ad hoc medicine** to **evidence based medicine** to **translational medicine** to **personalized medicine**
- From **point of need** to **point of care**
- The **NIH Clinical Center WebLib** Federated Search application
- The **Welsh Medical Library** at Johns Hopkins University

# Next Steps for HealthMash

and the

## Health Knowledge Base

- implement comprehensive semantic search
- expand the Health Knowledge Base
  - approximately 60,000 concepts in consumer health
  - millions of relationships in context
- into a **Biomedical Knowledge Base**
  - more 4,000,000 concepts
  - tens of millions of relationships **in context**
  - tens of millions of **semantic triples** and **n-tuples**
  - **Inferencing** and **question answering**
- from data to information to knowledge to wisdom
- leverage the wisdom of health professionals, the wisdom of patients, the wisdom of the masses and the wisdom of the ages

# Improved health information services

- *Libraries as Knowledge Bases*
- ***Librarians as Search Engines***
- *The Web as the Knowledge Base*
- ***Search Engines as Librarians***
- ***Librarians as Personal Information Trainers and Team Coaches***

so **semantic search** is what semantic technologies  
can do today

but

what the ...

is

**semantics**  
?

[Professor Irwin Corey explains](#)



Professor Irwin Corey at the Cutting Room NYC



UNITED STATES

National Library of Medicine

Integrating Semantic Search,  
Federated Search and  
Biomedical Knowledge Base  
Technologies for  
Improved Services by Medical Libraries

EAHIL 2010

Tamas Doszkocs, Ph.D.

Computer Scientist

[doszkocs@nlm.nih.gov](mailto:doszkocs@nlm.nih.gov)