The Future of Semantic Search
Where We Are and Where We Are Going

Tamas Doszkocs, Ph.D.
Computer Scientist

tamas@weblib.com
What is semantics?
What is 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

“All our work, our whole life is a matter of semantics”
Felix Frankfurter
“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
Organizing and Searching the Web

search
Manual ... DMOZ ...
Yahoo ...
Algorithmic ...
InfoSeek ...
Google ...
Social ...
Facebook ...
Google+ ...
Mobile ...
Apple ...
Google ...
BIG DATA ...
Hadoop ...
NoSQL ...
Semantic ...
Bing ...
Google ...
... Watson
Meno

And how will you enquire, Socrates, into that which you do not know?

What will you put forth as the subject of enquiry?

Socrates

I know, Meno, what you mean; but just see what a tiresome dispute you are introducing.

You argue that man cannot enquire either about that which he knows, or about that which he does not know;

for if he knows, he has no need to enquire;

and if he does not, he cannot; for he does not know the very subject about which he is to enquire.
a few reminders of keyword search problems:

“MS”
- acronyms and abbreviations
- NIH Site Search vs. NIHSeek

“Big Brother”
- user intent and context; disambiguation

“inflamed testicles”
- Query interpretation
- NLM PubMed vs. WebLib PubMed Reviews Semantic Search

“who performed the first us heart transplant”
- Complex queries and search logic
- database scope and user expectations: PubMed vs. Google
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)"

Information from the Knowledge Graph in search results

This feature is being released gradually over the next few days to people using google.com in English. If you don’t see this section on your result pages yet, don’t worry – we’re working to get it to you as soon as we can.

When you search on Google for a person, place, or thing, you might see a section to the right of your search results that highlights facts, photos, and other snippets of information about your search. Use this section to find quick information and facts about the subject or to start exploring related subjects.

To give it a try, search for your favorite movie, landmark, historical figure, or try one of these: { Eiffel Tower }, { Hayne's World }, { dalmation }, { Galileo }. You can also search by image to see this section of facts and information.
Semantic Search Trends

... Bing Decision Engine ... Google Knowledge Graph ...

A database of more than 570 million real-world people, places and things with 18 billion attributes and connections among them.
Projections indicate that up to 100 billion uniquely identifiable objects will be connected to the Internet by 2020.
Google Knowledge Graph

drug information

Cozaar Information from Drugs.com
www.drugs.com/cozaar.html
Cozaar (losartan) is used to treat high blood pressure (hypertension). Includes Cozaar side effects, interactions and indications.
Cozaar Side Effects - Cozaar Dosage - Cozaar Drug Interactions - Losartan

PEP COZAAR - Merck
File Format: PDF/Adobe Acrobat - Quick View
is detected, COZAAR® should be discontinued as soon as possible. ... COZAAR® is available as tablets for oral administration containing either 25 mg, 50 mg or ...

Side Effects of Cozaar (Losartan Potassium) Drug Center - Rxlist
www.rxlist.com/cozaar-side-effects-drug-center.htm
Find a comprehensive guide to possible side effects when taking Cozaar (Losartan Potassium) for Professionals, Patients, and Caregivers.
Cozaar User Reviews - Cozaar - How should I take losartan ... - Precautions

Losartan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Losartan
Clinical data. Trade names, Cozaar, AHFS/Drugs.com - monograph ... Inc. under the trade name Cozaar. As of 2009, losartan is available in generic form.
Mechanism of action and ... - Pharmacokinetics - Research - Synthesis

Losartan (Cozaar) Side Effects, Drug Interactions - MedicineNet
www.medicinenet.com/losartan/index
by Dr. Omuchome Ogbru - in 108 Google+ circles - More by Dr. Omuchome Ogbru
Explains the medication losartan (Cozaar), a drug indicated for the treatment of hypertension. Side effects, drug interactions, warnings and precautions, and use ...

Cozaar (losartan)
“Scientists today cannot hope to manually track all of the published science relevant to their work. A cancer biologist, for instance, can find more than 2 million relevant papers in the PubMed archive, more than 200 million Web pages with a Google search, and databases holding results from experiments that produce millions of gigabytes of data.”

*Machine Science*

http://www.sciencemag.org/content/329/5990/399.summary
Overview of Semantic Search Approaches
WebLib’s Semantic Search Technology
WebLib Semantic Search and Discovery Apps

federated search, distributed search, discovery search, enterprise search, semantic search

- PolyMeta
- HealthMash
- NLMplus
- WebLib Green Energy Semantic Search
  - OSTI Green Energy Semantic Search
  - OntoFind
Sciatica is a symptom of a problem with the sciatic nerve, a large nerve that runs from the lower back down the back of each leg. It controls muscles in the back of your knee and lower leg and provides feeling to the back of your thigh, part of your lower leg and the sole of your foot. When you have sciatica, you have pain, weakness, numbness or tingling. It can start in the lower back and extend down your leg to your calf, foot, or even your toes. It's usually on only one side of your body. Sciatica...

OntoFind and the Web Knowledge Base

Explore and Discover:

**sciatica**

- Sciatica
  - Sciatic nerve
  - Pain
    - Low back pain
    - Back pain
    - Neuralgia
    - Referred pain
  - Spinal disc herniation
  - Degenerative disc disease
  - Back pain
  - Lumbar spinal stenosis
- Vertebral column
  - Spinal disc herniation
  - Radiculopathy
  - Lumbar spinal stenosis
  - Piriformis syndrome
- Neurological disorder
  - Piriformis syndrome
  - Radicular pain
  - Sacroiliac joint dysfunction
  - Spinal decompression

Google: Google-Keyword 50 - 10,100,000

49 Results for: sciatica...

Web Results

1. **Sciatica - Wikipedia, the free encyclopedia**
   - [Sciatica](https://en.wikipedia.org/wiki/Sciatica)
   - 2,342 people like this.
   - **Sciatica** is a set of symptoms including pain that may be caused by general nerve compression.

2. **Sciatica Symptoms, Causes, Treatments, Exercise**
   - [Sciatica Symptoms](https://www.webmd.com/back-pain/guide/sciatica-symptoms)
   - 619 people like this.
   - Learn about the symptoms and causes of **sciatica**, a painful nerve condition, from [webmd.com](https://www.webmd.com)

3. **Sciatica - MayoClinic.com**
   - [Sciatica](https://www.mayoclinic.org/diseases-conditions/sciatica/symptoms-causes/syc-20358722)
   - 346 people like this.
What is OntoFind?

OntoFind is a customizable semantic search engine powered by WebLib’s Web Knowledge Base.

OntoFind facilitates knowledge discovery from heterogeneous information sources.

OntoFind produces more meaningful search results than conventional keyword searches.

The Web Knowledge Base is available as a web service.
Key Components of WebLib’s Semantic Search and Knowledge Discovery Technology
Taxonomy Powered Content

- Content with applied taxonomy
- Content today with structured XML
- Smart Content: Human Consumption, Machine Organization & Display, Machine Interpretation
- Structured Content: Human Consumption, Machine Organization & Display
- Unstructured Content: Human Consumption
- Manual Processes vs. Partially Automated vs. Highly Automated
- Semantic Meaning vs. Logical Structure vs. Flat Text
- Higher Value vs. Medium Value vs. Lower Value
Semantic Search and Smart Content

Smart Content At Elsevier

- Better discovery through semantic search & navigation
  - Faceted search & browse
  - Ontology-driven navigation
  - Task-specific results
  - Personalized/localized results
  - Link to evidenced-based content

- Better understanding through analysis and visualization
  - Question & Answer
  - Actionable Content & Alerts
  - Tag clouds
  - Heatmaps
  - Animations

- New knowledge through aggregation and synthesis
  - Topic pages
  - Social network maps
  - Geolocation maps
  - Data integration and mashups
  - Text mining
  - Inference and Reasoning
Where are we and where are we going?

The Wild Wild Web

where everything is getting meaningfully connected
The Tip of the Iceberg

- The Open Web
- The Hidden Web
Keyword Search vs. Semantic Search

**Symbols, Strings and Operators**
- manipulating query words
- limited utilization of morphology and syntax
- displaying links and text snippets
- Placing the burden of information digestion on the user

**Real world Entities and Relationships**
- Find the meaning of queries and content
- Identify Intent and context
- Nuance facts, data, answers, information
- Discover, dig, digest, deliver
• **Semantic Searching on the Web** vs. **Searching on the Semantic Web**

• **Semantic University**

• **The Semantic Web Landscape**
  – W3C Standards
    • RDF (Resource Description Framework)
    • RDFa (for marking up data inside Web pages)
    • SPARQL (SPARQL Protocol and RDF Query Language)
    • OWL (Web Ontology Language)
Semantic Web Technologies are a family of very specific technology standards from the World Wide Web Consortium (W3C) that are designed to describe and relate data on the Semantic Web and inside enterprises, e.g.

- **Ontologies**
- **Linked Open Data**
- Semantic Web Technologies complement other Semantic Technologies on the Web
- Tim Berners-Lee’s vision of an entire Web of interoperable data has not yet been realized
Semantic Technologies

a diverse family of technologies that seek to derive meaning and knowledge from data

- **Natural-language processing (NLP).** is a field of information science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. **UIMA, Stanford NLP, Open NLP…**

- **Data and Text Mining, Big Data and Big Analytics.** Data mining technologies employ pattern-matching algorithms to tease out trends and correlations within large sets of data. **Hadoop, MongoDB, Oracle, IBM…**

- **Artificial intelligence and Expert Systems.** AI or expert systems technologies use elaborate reasoning models to answer complex questions automatically. These systems often include machine-learning algorithms that can improve the system's decision-making capabilities over time.

- **Database Management Systems (DBMS), NoSQL, MarkLogic…**

- **Information Retrieval, Search Engines and Semantic Search Engines, Solr/Lucene, SIRE …**

- **Cloud Computing, Amazon EC2, Microsoft and Google…**

- **Visualization, Prefuse, KNALIJ…**
Semantic Search and Discovery Engines

Albert Einstein's Brain  Google Server Farm  Timeline of Computable Knowledge  Watson
The State of the Semantic Arts

• The Web is becoming “semantified”
• Search engines, which have achieved remarkable success *without semantics*, are morphing into *semantic search engines* in order to produce *better results* and meet *more complex information needs*
• “There are no companies or products in this field that have eclipsed all others offering *universal semantic processing* or *semantic search*, yet”
Big Data, Big Analytics, Big Challenges, Big Insights, Big Bang for the Buck, Big Mistakes, Big Abuses, Big Crises

- We are in a crisis of knowledge

- Every day, we create 2.5 quintillion bytes of data ($10^{18}$)

- 90% of the data in the world today has been created in the last two years alone

- Tens of billions of web link graph nodes indexed and over one trillion nodes discovered by the major search engines

- Access to a large amount of knowledge is critical for success at answering open-domain questions (David Ferrucci, Lead Researcher for IBM's Watson Project)

- Statistical techniques that use simple models and use lots of data trump approaches that use complex models, deep algorithms or hand-coded rules (Kaushik Chakrabarti, Microsoft)
Information Visualization
worth watching and hearing

semantic search behind the scenes

• Breakfast with Google’s Search Team
• Microsoft Speech Recognition Breakthrough for the Spoken, Translated Word
semantic search is what semantic technologies can do today but what the .... is semantics?

Professor Irwin Corey explains
Semantic Search Means Well

Seek and Ye Shall Find

Tamas Doszkocs, Ph.D.
Computer Scientist

Breakfast with Google’s Search Team

tamas@weblib.com
References

http://youtu.be/FawygFT5Brs

Introducing the Knowledge Graph: things, not strings
http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Breakfast with Google’s Search Team
Published on Aug 20, 2012 by GoogleVideos
http://www.youtube.com/watch?feature=player_embedded&v=8a2VmxqFg8A

Microsoft Speech Recognition Breakthrough for the Spoken, Translated Word
October 2012 (Speaking in English >> Chinese in the speaker’s own voice)
http://www.youtube.com/watch?v=Nu-nlQqFCKg

Building a Web-Scale Semantic Search Engine
John Shafer, Search Labs, Microsoft Research
http://km.aifb.kit.edu/ws/jiws2012/papers/Lincoln%20Project%20-%20JIWES12%20Keynote.pptx
References

Semantic Search Over the Web
Roberto De Virgilio, Francesco Guerra and Yannis Velegrakis
Springer, 2012

Semantic Search Tutorial at SemTech 2012
http://www.slideshare.net/pmika/semantic-search-tutorial-at-semtech-2012

1st Joint International Workshop on Entity-oriented and Semantic Search (JIWES) 2012
http://km.aifb.kit.edu/ws/jiwes2012/

Machine Learning for Query-Document Matching in Web Search
Hang Li and Jun Xu, Microsoft Research Asia

Large-Scale Graph Mining and Learning for Information Retrieval
Bin Gao, Taifeng Wang, and Tie-Yan Liu, Microsoft Research Asia
Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. A Case for Semantic Full-Text Search (position paper) (paper/slides)

This is Watson

Inside the mind of Watson
Chris Welty, July 4, 2012
http://videolectures.net/eswc2012_welty_watson/

Timeline of Systematic Data and the development of Computable Knowledge
http://www.wolframalpha.com/docs/timeline/

Is Google Afraid of the Big Bad Wolfram?
Simon Penson, August 27, 2012
References


On the foundations of data interoperability and semantic search on the web Haidarian Shahri, dissertation

Wikidata
“Wikidata aims to create a free knowledge base about the world that can be read and edited by humans and machines alike”
http://meta.wikimedia.org/wiki/Wikidata

Schema.org
http://blog.schema.org/

Linked Data: A Personal View from Jerry Persons
18 June 2012
http://www.diglib.org/archives/3167/
References

Elsevier Smart Content LDR, SemTech NYC Oct-17-2012
by Alan Yagoda on Oct 18, 2012

The once-hyped Semantic Web still bubbles along
November 13, 2012

Semantic Search in 2025
Doc Sheldon, November 6, 2012

A Data 7th Wave Approaching
By Richard Wallis on February 16, 2012
http://dataliberate.com/2012/02/a-data-7th-wave-approaching/

What is Semantic Search?
Tony John, March 15, 2012
References

The Restructuring of Knowledge
David Weinberger
http://semtechbiznyc2012.semanticweb.com/sessionPop.cfm?confid=68&proposalid=4989

Web 3.0 – The Dawn of Semantic Search
James Hendler, IEEE Computer, Jan. 2010

Breakthrough Analysis: Two + Nine Types of Semantic Search
By Seth Grimes InformationWeek
January 21, 2010


Improving large-scale search engines with semantic annotations
Damaris Fuentes-Lorenzo Corresponding et al.
References

Semantic University
http://www.cambridgesemantics.com/semantic-university

Semantic Search and the Semantic Web
Semantic University

SemSearchPro – Using Semantics throughout the Search Process
Thanh Tran, Daniel M. Herzig, Gunter Ladwig

SIREn: Semantic Information Retrieval Engine
“SIREn extends the widely recognized open-source search products Lucene and Solr in order to offer powerful semistructured search features”
http://siren.sindice.com

SindiceTech
“SindiceTech specializes in "Big Data" infrastructure that deals with semistructured, semantic data”…
http://www.sindicetech.com/
Tamas Doszkoccs

Computer Scientist, CEO WebLib LLC

Received his Ph.D. from the University of Maryland where he was also granted an M.S. degree in Computer Science and an M.L.S. degree in Library and Information Science.

Current R & D interests focus on semantic technologies, bioinformatics, natural language processing, and web science.

In prior positions at the U.S. National Library of Medicine, National Institutes of Health, Tamas served as a senior computer scientist, Chief of the Biomedical Files Implementation Branch, and Chief of the Technical Services Division.

Tamas also taught at the University of Maryland and Indiana University and worked in private industry as a systems architect, software engineer and entrepreneur.