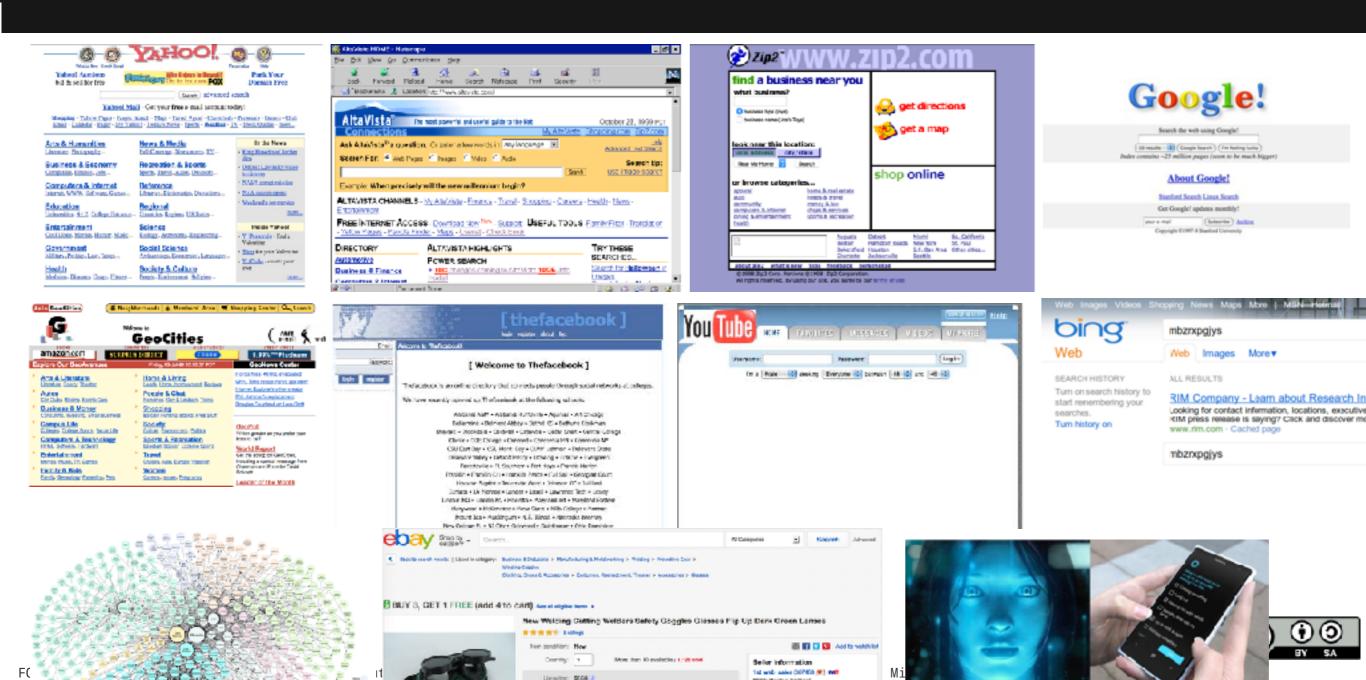
### Ask Me Anything about

# Search Engines History and Future

(altavista -> Google -> Bing -> Wolfram Alpha -> Social Graph -> ?)

Michael Christen M.Sc., mc@yacy.net, @0rb1t3r



#### Geocities 1994 - 2009

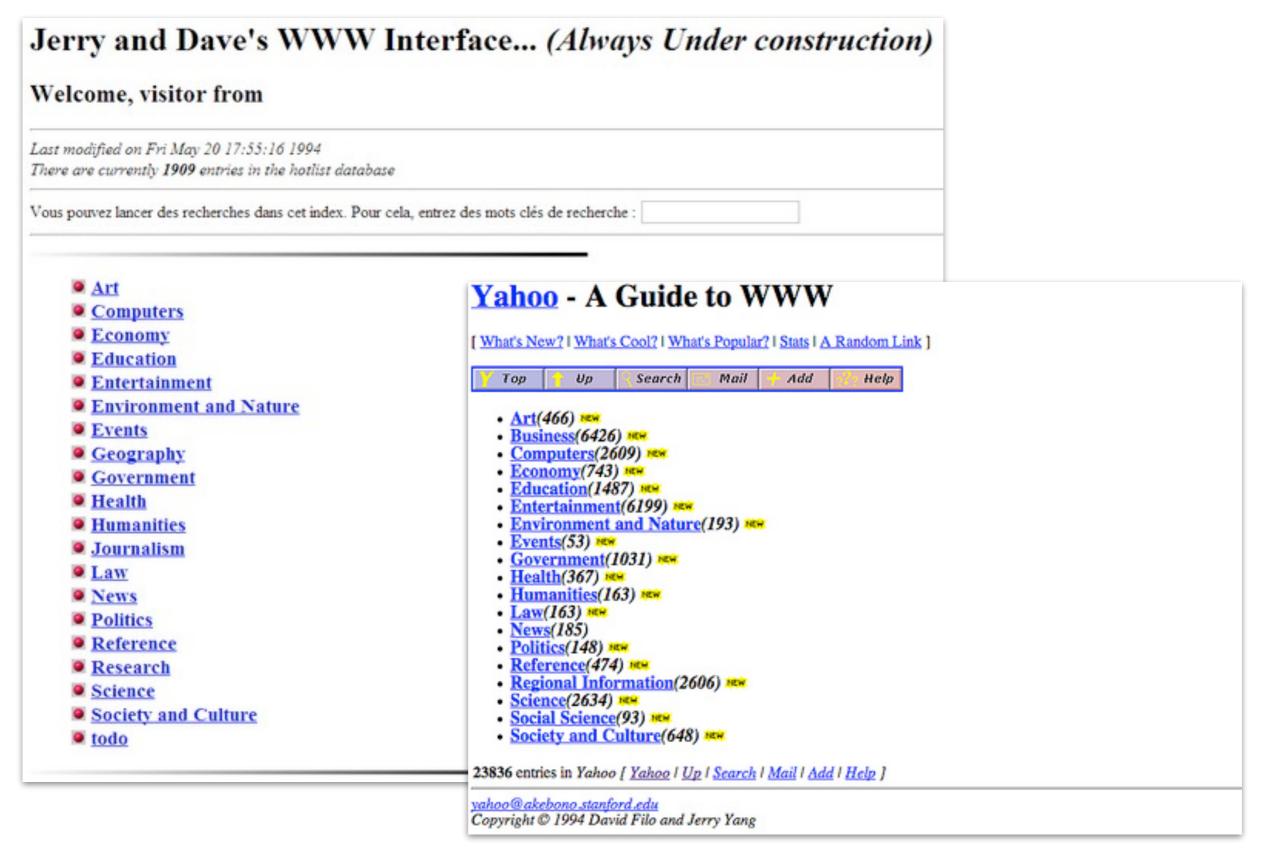


Yahoo bought Geocities 1999 and delete it 2009

38 Million personal home pages ("the first web") lost!

reocities.com has a copy, ca. 600GB torrent was there as well

### Yahoo 1994 - started as Jerry and Dave's Bookmarks





### **Yahoo** 1995









Yahoo! Auctions bid & sell for free



Park Your **Domain Free** 

advanced search Search

Yahoo! Mail - Get your free e-mail account today!

Shopping - Yellow Pages - People Search - Maps - Travel Agent - Classifieds - Personals - Games - Chat Email - Calendar - Pager - My Yahoo! - Today's News - Sports - Weather - TV - Stock Quotes - more ...

#### Arts & Humanities

Literature, Photography...

#### **Business & Economy**

Companies, Finance, Jobs...

#### Computers & Internet

Internet, WWW, Software, Games ...

#### Education

Universities, K-12, College Entrance...

#### Entertainment

Cool Links, Movies, Humor, Music...

#### Government

Military, Politics, Law, Taxes...

#### Health

Medicine, Diseases, Drugs, Fitness...

#### News & Media

Full Coverage, Newspapers, TV ...

#### Recreation & Sports

Sports, Travel, Autos, Outdoors...

#### Reference

Libraries, Dictionaries, Quotations...

#### Regional

Countries, Regions, US States...

#### Science

Biology, Astronomy, Engineering...

#### Social Science

Archaeology, Economics, Languages...

#### Society & Culture

People, Environment, Religion...

#### In the News

- King Hussein of Jordan dies
- Online: Lewinsky video testimony
- NASA comet mission
- NBA season opens
- Weekend's top movies

more...

#### Inside Yahoo!

- Y! Personals find a Valentine
- Shop for your Valentine
- Y! Clubs create your own

more...

Yahoo Inc. (styled as Yahoo!) is an American multinational technology company headquartered in Sunnyvale, California. It is globally known for its Web portal, search engine Yahoo! Search, and related services, including Yahoo! Directory, Yahoo! Mail, Yahoo! News, Yahoo! Finance, Yahoo! Groups, Yahoo! Answers, advertising, online mapping, video sharing, fantasy sports and its social media website. It is one of the most popular sites in the United States. [9] According to third-party web

Yahoo Inc.

Type Public

NASDAQ: YHOO & Traded as

NASDAQ-100 Component

S&P 500 Component

January 1994 (as Jerry and Founded

David's guide to the World

Wide Web) March 1995 (as Yahoo!)

Headquarters Sunnyvale, California, U.S.

Area served Worldwide

Founder(s) Jerry Yang, David Filo

Chairman Maynard Webb

CEO Marissa Mayer

David Filo Key people

(Chief Yahoo)[1]

Industry Internet

Computer software

**Products** Yahoo News

> Yahoo Mail Yahoo Finance Yahoo Sports Yahoo Search Yahoo Messenger

Yahoo! Answers

Tumblr Flickr

See Yahoo products

▲ US\$4.96 billion (2015)[2] Revenue

Operating income

▼ US\$-4.74 billion (2015)[3]

Net income

▼ US\$-4.35 billion (2015)[4]

Total assets Total equity

∵isten

▼ US\$45.20 billion (2015)[5] ▼ US\$29.04 billion (2015)[6]

12.500 (2015)[7] **Employees** 

#### AltaVista 1995-2013



AltaVista was an early web sear search engines, but it lost ground brand but based all AltaVista sea down by Yahoo! and since then, Email from early January, 1996:

Thank you for your comments. As we are starting up the Beta phase of the Alta Vista project, we are trying to respond individually to some of the messages we have received.

We are putting the following hardware information on the site.

Alta Vista is a very large project, requiring the cooperation of at least 5 servers, configured for searching huge indices and handling a huge Internet traffic load. The initial hardware configuration for Alta Vista is as follows:

Alta Vista -- AlphaStation 250 4/266

4 GB disk 196 MB memory

Primary web server for gotcha.com

Queries directed to WebIndexer or NewsIndexer

NewsServer -- AlphaStation 400 4/233

24 GB of RAID disks

160 MB memory

News spool from which news index is generated

Serves articles (via http) to those without news server

NewsIndexer -- AlphaStation 250 4/266

13 GB disk 196 MB memory

Builds news index using articles from NewsServer

Answers news index queries from Alta Vista

Spider -- DEC 3000 Model 900 (replacement for Model 500)

30 GB of RAID disk

1GB memory

Collects pages from the web for WebIndexer

WebIndexer -- Alpha Server 8400 5/300

210 GB RAID disk (expandable)

4 GB memory (expandable) 4 processors (expandable)

Builds the web index using pages sent by Spider.

Answers web index queries from Alta Vista

Thank you,

Alta Vista Technical Support

#### **AltaVista**







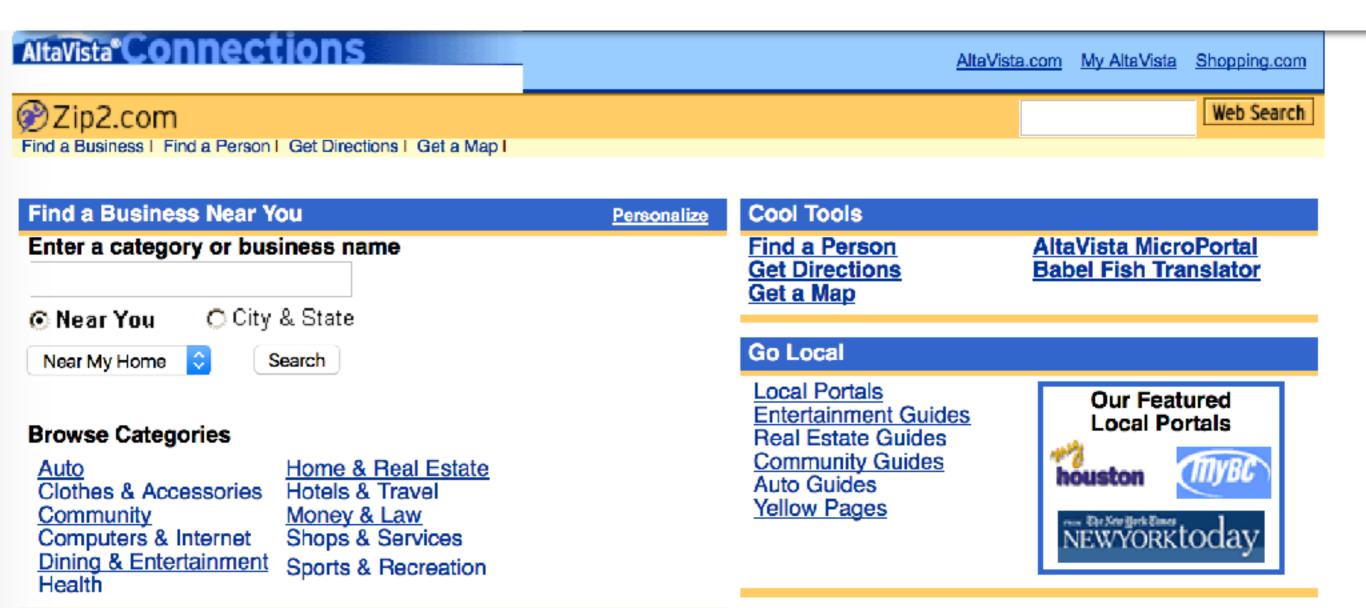
#### find a business near you what business? get directions ubusiness type (toys) business name (Joe's Toys) get a map look near this location: local address city/state Search Near My Home shop online or browse categories... home & real estate apparel auto hotels & travel money & law community shops & services computers & internet dining & entertainment sports & recreation health So. California Detroit Augusta Miami Boston Hampton Roads New York St. Paul Bakersfield Houston S.F. Bay Area Other cities... Charlotte Jacksonville Seattle

about zip2 what's new jobs feedback personalize

©1998 Zip2 Corp. Portions ©1998 Zip2 Corporation.

All rights reserved. By using our site, you agree to our terms of use.

# Zip2 1999 - bought by AltaVista



AltaVista/Zip2 provides the Internet's leading local portal services to get you locally relevant results -- faster.

Products | Services | Affiliates | Partners | News | Company | Jobs

Feedback Help

©1999 Zip2 Corp. Portions ©1999 Zip2 Corporation.

All rights reserved. By using our site, you agree to our terms of use.



### Google 1998



Google is an American multinational technology company specializing in Internet-related services and products. These include online advertising technologies, search, cloud computing, and software.<sup>[5]</sup>

Most of its profits are derived from AdWords,<sup>[6][7]</sup> an online advertising service that places advertising near the list of search results.

#### Google Inc.



Type Subsidiary

Industry Internet

Computer software

Founded September 4, 1998; 17 years

ago

Menlo Park, California[1][2]

Founders Larry Page

Sergey Brin

Headquarters Googleplex, Mountain View,

California, U.S.[3]

Coordinates @ 37.422°N 122.084058°W

Area served Worldwide

Key people Sundar Pichai (CEO)

Products List of Google products

Parent Independent (1998–2015)

Alphabet Inc. (2015-present)

Subsidiaries List of subsidiaries

Footnotes / references

[4]



### Web 2.0 since 2000-

burst of the .com bubble

user-generated content forums, wikis, etc

unsuccessful attempts to create alternatives to Google



### Wikipedia 2001

#### **HomePage**

HomePage I RecentChanges I Freferences I Random Page
You can edit this page right now! It's a free, community project

Welcome to Wikipedia, a cotaborative project to produce a complete encyclopedia from scratch. We started in January 2001 and sheady have over 6,000 articles. We want to me so let's get to work—anyone can edit any page—copyedit, write a little, write a lot. See the Wikipedia FAQ for information on how to edit pages and other questions. If you're visiting is first time, welcome! The content of Wikipedia is covered by the GNU Free Documentation License.

#### Philosophy, Mathematics, and Natural Science

Astronomy and Astrophysics - Biotics - Chemistry - Earth Sciences - Mathematics - Philosophy - Physics - Science - Statistics

#### Social Sciences

Anomalous Phenomena - Anthropology - Archaeology - Countries of the world - Economics - Geography - History of Science and Technology - Language - Lin. - Etachology - Sociology

#### Applied Arts and Sciences

Addiculture - Architecture - Business and Industry - Communication - Computing - Education - Engineering - Earnly and Consumer Science - Health Sciences - Law - Uniformation Science - Endisc Affairs - Technology - Transport

#### Culture

Classics - Critical Theory - Dance - Entertainment - Film - Cames - Hobbies - Literature - Music - Opera - Parting - Performing Arts - Recreation - Fieldor - Soulds.

Theater and Drame - Tourism - Visual Arts and Design

#### Other Category Schemes

About Wilipedia category schemes — Library of Congress catalog scheme — Dewey Decimal System — Wilipedia arranged by topic — Year in Review — Historical answersar tables — Biographical Listing

#### International Wikipedias

About the International Whipedian — ICutatan (Catalati — IChnese Printwill — ICertain (Catalati — ICertain (Catalati

#### Wikipedia

FAQ - Policy - Article news - Amountements - Requested articles - Help desk - New topics - Britisht prose - Meling list - Feature requests - Bug reports - Commontes Domain Resources - Wilepedia and Nupedia - Sandbox - License - Wilepedians

Today is Finday, [July 27] 2001, servertime (Pacific Standard Time)

(Tek



#### Facebook 2004



Facebook is a corporation and online social networking service headquartered in Menlo Park, California, in the United States. Its website was launched on February 4, 2004, by Mark Zuckerberg with his Harvard College roommates and fellow students Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes. [8][9][10] The founders had initially limited the website's membership to Harvard students, but later expanded it to colleges in the Boston area, the Ivy League, and Stanford University. It gradually added support for students at various other universities and later to high-school students. Since 2006, anyone who is at least 13 years old was allowed to become a registered user of the website, though the age requirement may be higher depending on applicable local laws. [11] Its name comes from the face book directories often given to American university students. [12]

Facebook, Inc.



### Youtube 2005



first video was the visit of the founder at the zoo



#### Twitter 2006

#### Twitter.com

A global community of friends and strangers answering one simple question: What are you doing?



Tim Huegdon Just got up. Not in work this



Twitter (/twxter/) is an online social networking service that enables users to send and read short 140-character messages called "tweets".

Registered users can read and post tweets, but those who are unregistered can only read them. Users access Twitter through the website interface, SMS or mobile device app.<sup>[10]</sup> Twitter Inc. is based in San Francisco and has more than 25 offices around the world.<sup>[11]</sup>

#### Twitter, Inc.



Twitter's homepage as of August 2015 (only in

select countries)

Type Public

Traded as NYSE: TWTR ₽

Founded March 21, 2006; 9 years ago<sup>[1]</sup>

Headquarters San Francisco, California,

United States<sup>[2]</sup>

Area served Worldwide

Founder(s) Jack Dorsey, Noah Glass, Biz

Stone, Evan Williams

Chairman Omid Kordestani

CEO Jack Dorsey

Industry Internet

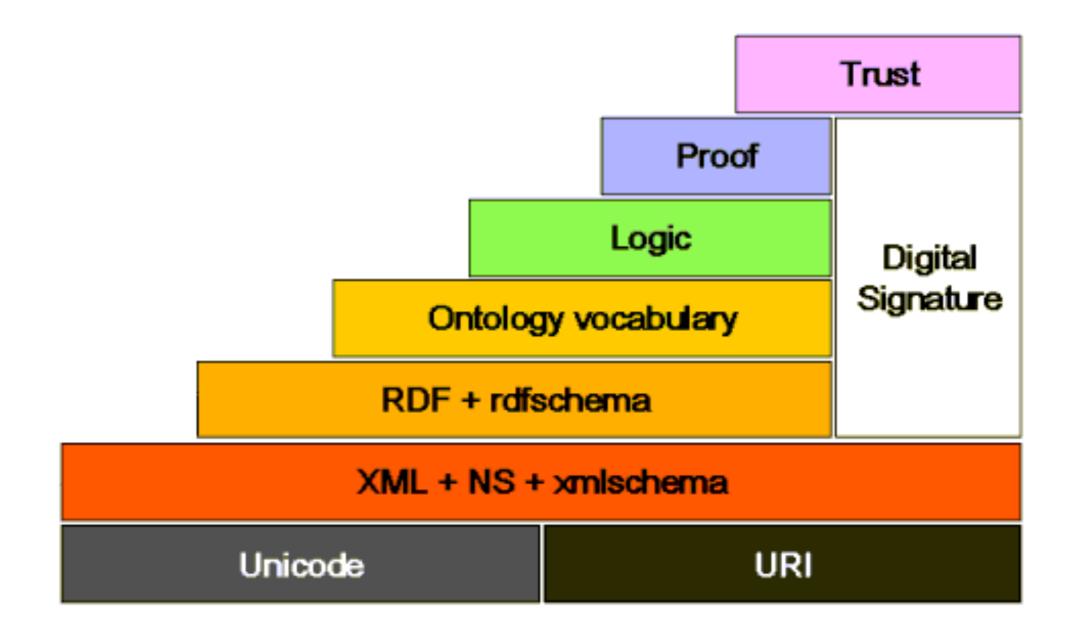
Revenue ▲ US\$2.21 billion (2015)[3]

Net income VS\$521 million (2015)[3]

Employees 3,638 (2015)[4]

Subsidiaries Vine

### Semantic Web mid-2000 .. kind of first version



a lot of money was wasted to find out what one can do with this. Quaero 2005 - 2008 The European Union spend 400 million Euro for nothing NOBODY had the slightest idea how a search interface would look like!

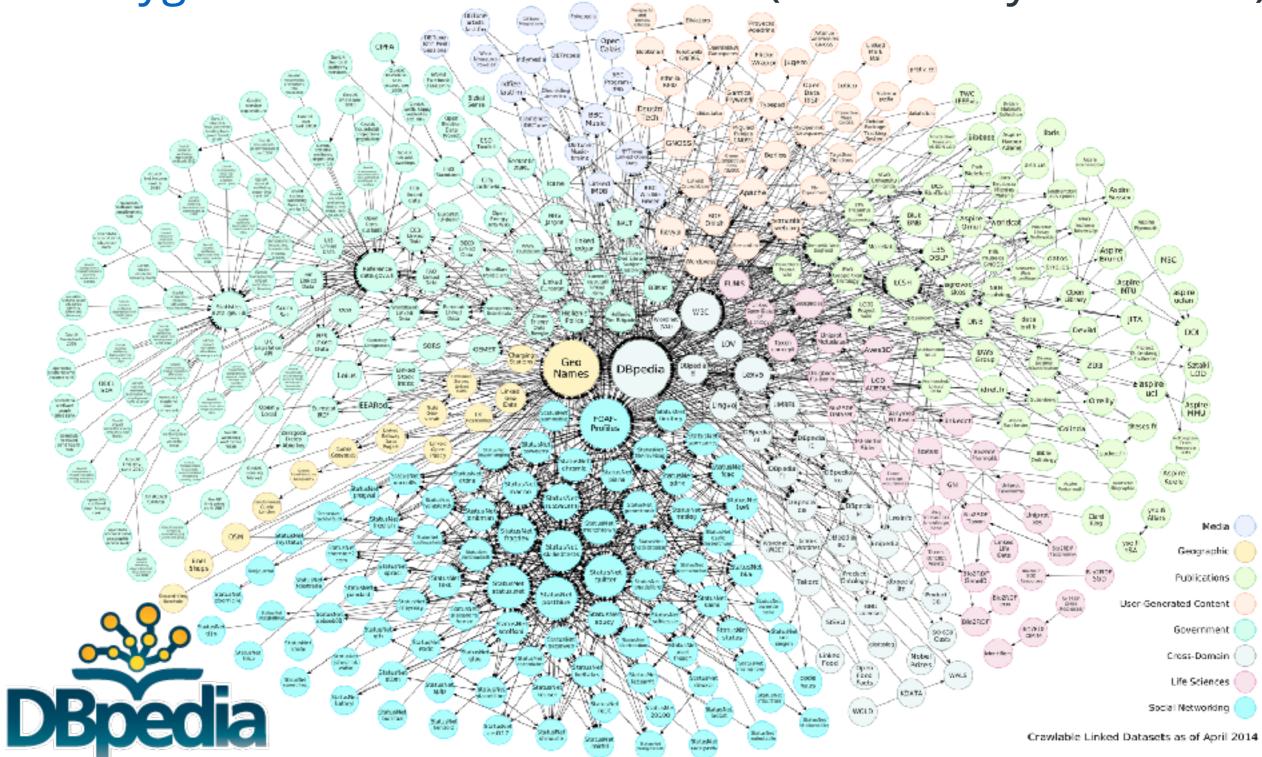


# Semantic Web re-named to Giant Global Graph



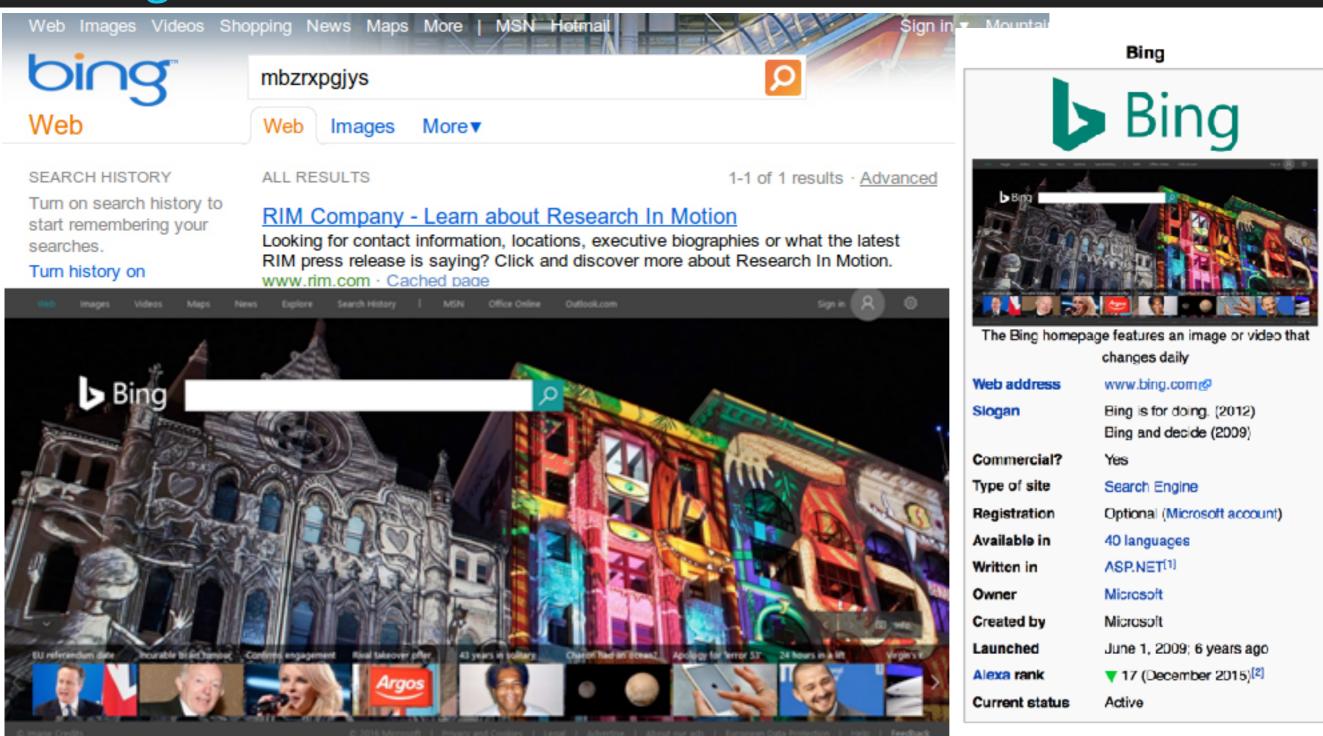
### Giant Global Graph, second version' of Semantic Web

Playground of scientific interest (but nobody else cares)



Example http://dbpedia.org/page/Open\_Data\_Institute\_

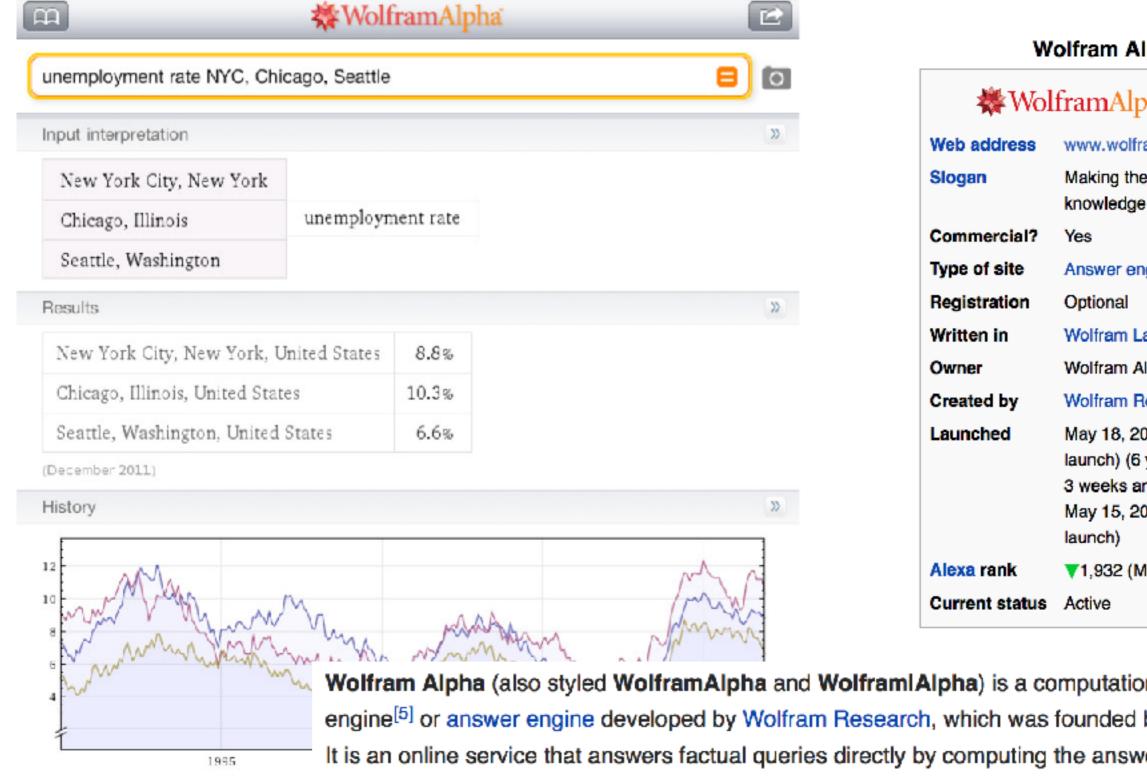
## **Bing** 2009



Bing is a web search engine owned and operated by Microsoft. The service has its origins in Microsoft's previous search engines: MSN Search, Windows Live Search and later Live Search. Bing provides a variety of search services, including web, video, image and map search products. It uses the ASP.NET programming language and follows the design principles of Microsoft's "Metro" design language.



# WolframAlpha 2009 "Knowledge Engine", not a SE



#### Wolfram Alpha



Wolfram Alpha (also styled WolframAlpha and WolframIAlpha) is a computational knowledge engine<sup>[5]</sup> or answer engine developed by Wolfram Research, which was founded by Stephen Wolfran It is an online service that answers factual queries directly by computing the answer from externally sourced "curated data", [6] rather than providing a list of documents or web pages that might contain the answer as a search engine might.[7]



F0S

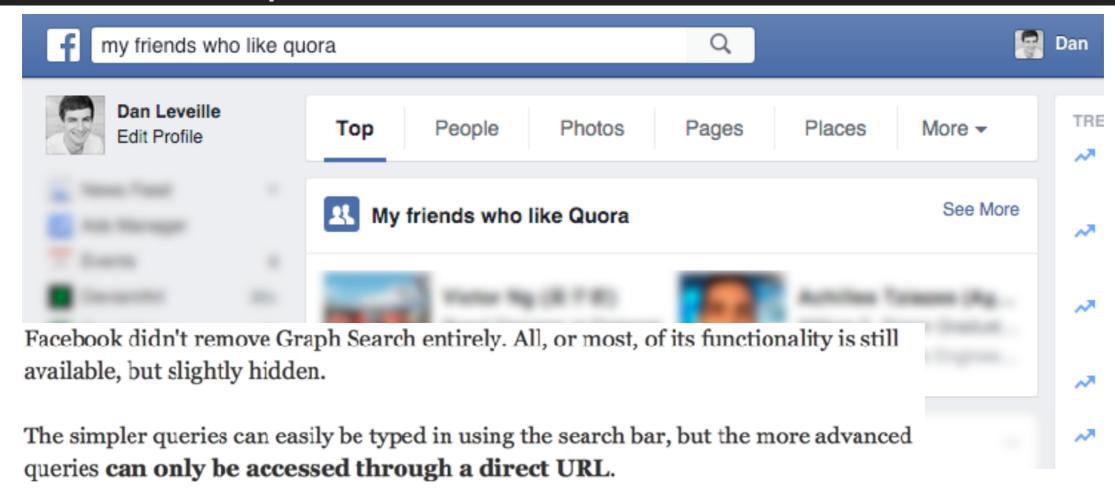
(not seasonally adjusted)

New York City, New York, United States

(from Jan 1990 to Dec 2011) (in percent)

Chicago, Illinois, United States

### Facebook Graph Search 2012-2014



My friends who like Quora:

facebook.com/search/255232486973/likers/me/friends/intersect

My male friends:

facebook.com/search/males/me/friends/intersect

My friends who live near Los Angeles:

facebook.com/search/110970792260960/residents-near/present/friends/intersect

My male friends who live near Los Angeles that like Quora:

facebook.com/search/255232486973/likers/males/110970792260960/residents-near/present/me/friends/intersect

### Semantic Web "v3.0" - but we do not call it like that

schema.org			Search
	Home	Schemas	Documentation

#### Organization of Schemas

The schemas are a set of 'types', each associated with a set of properties. The types are arranged in a hierarchy. The core vocabulary currently consists of 642 Types, 992 Properties, and 219 Enumeration values.

Browse the full hierarchy:

### Google suggested an RDF schema...

- One page per type
- Full list of types, shown on one page

Or you can jump directly to a commonly used type:

- Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Embedded non-text objects: AudioObject, ImageObject, VideoObject
- Event
- Health and medical types: notes on the health and medical types under MedicalEntity.
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating
- Action

Using the extension mechanism the core vocabulary is extended by the following hosted extensions:

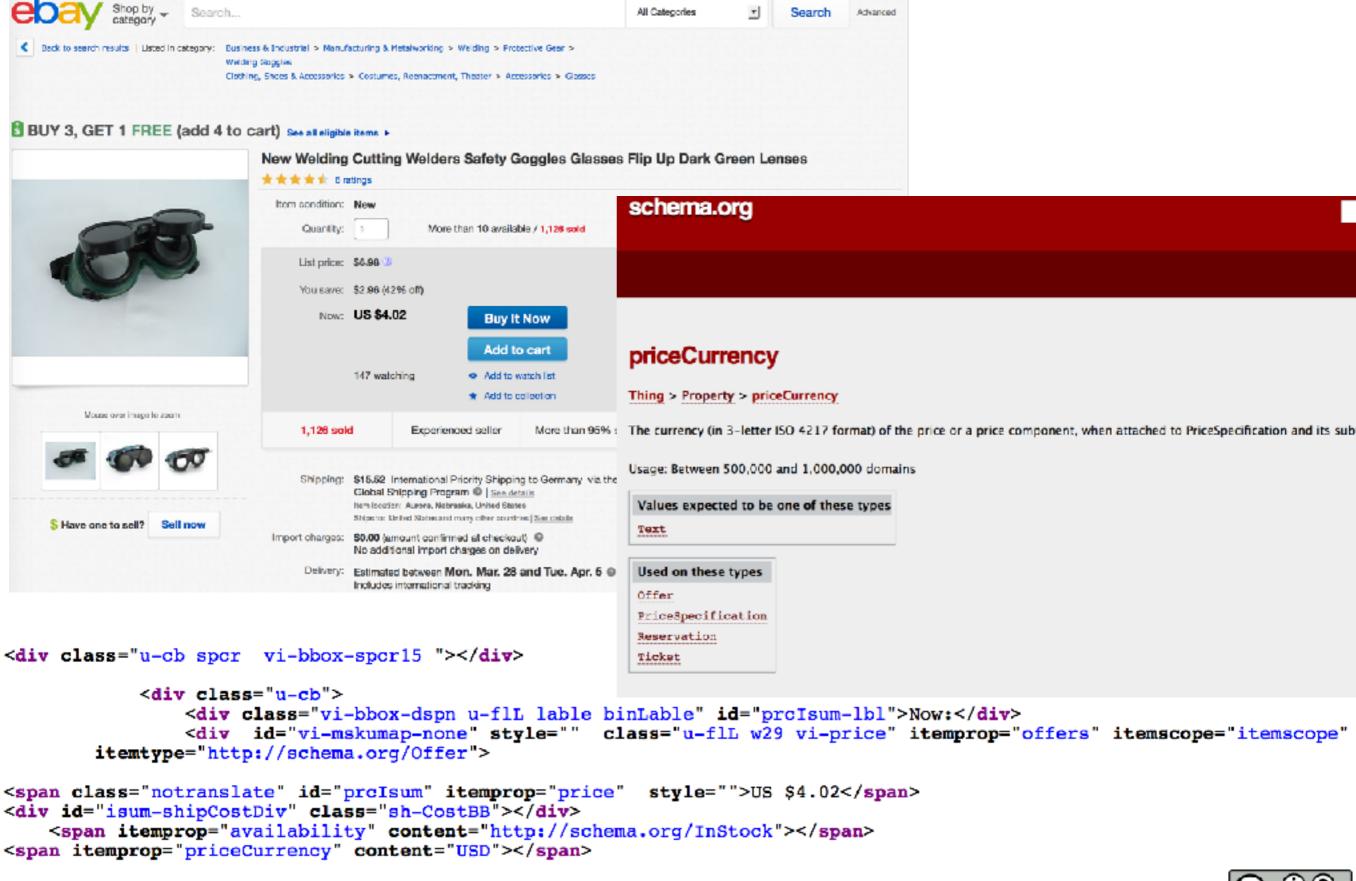
- auto.schema.org
- · bib.schema.org

See also the releases page for recent updates and project history.

We also have a small set of primitive data types for numbers, text, etc. More details about the data model, etc. are available here.



# Semantic Web is actually widely used through schema.org



# Library:

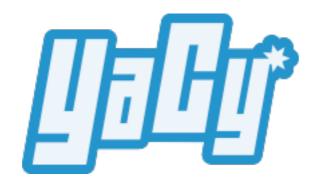


with
XML/JSON
API:





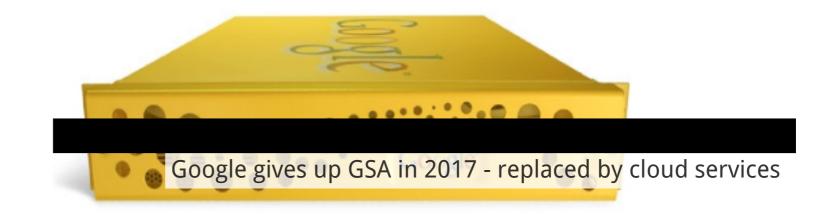
with API+ Harvester:







### on Hardware:



### **Software:**



### **Software:**





# The Future from the past



Star Trek IV 1986 - nobody uses a keyboard in the future



### Voice Search Smart Phone becomes Search Interface

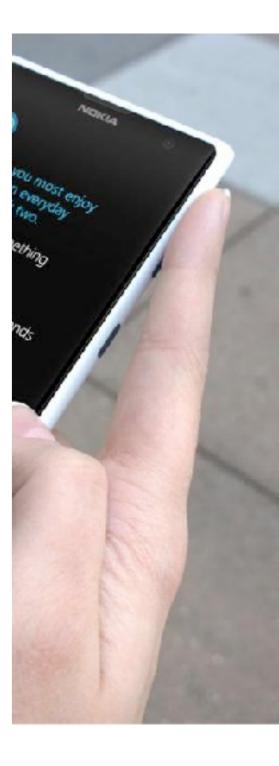


# "Sing Me a Song"











It will always be about answering questions!



maybe future search engines fill the gaps in the currently notanswered questions.



most common questions are completely answered:

weather?

what to do next?

how can I?

sports?

how can i connect to someone?

world knowledge

where can I buy?

what should I buy?



### Recommendation Engines are based on Search Engine Tec

what should I buy?





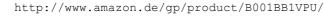
Total price: \$15.97

Add both to Cart

Add both to List

http://www.amazon.com/Black-Welding-Cup-Goggles-50mm/dp/B002DLYR8A/







Which
Question do
People Ask?

who am i
who is perfect
who let the dogs out
who club augsburg

Weitere Informationen





where

where is the love

where is my mind

where to invade next

where are you now

Weitere Informationen

Zum Start der Suche Eingabetaste drücken



Which Question do People Ask?

what do you mean
what the font
what is my ip
what does the fox say
Weitere Informationen



Which
Question do
People Ask?

how i met your mother
how deep is your love
how hot
how to be single

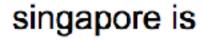
Weitere Informationen



Which Question do People Ask?

singapore is a fine city

singapore is the best



singapore is in what country singapore is located where

if a search engine gets a lot of attention...

it learns... and gets stronger!



Weitere Informationen

future search engines will answer to all questions!

the more we search, the more the search engine operator knows about us

maybe even before we ask



### Ask Me Anything about

## Search Engines History and Future

(altavista -> Google -> Bing -> Wolfram Alpha -> Social Graph -> ?)

Michael Christen M.Sc., mc@yacy.net, @0rb1t3r

# Why not trying to run a search engine yourself using free software?!



Web Search Engine http://yacy.net



Twitter Search Engine http://loklak.org



### Ask me anything about Search Engines

What can I ask about Search Engines?

SEO (= Google Optimization Ranking / Relevancy

Online Marketing

Big Data / Scaling

Microservices (Solr, Elasticsearch)

Search Suggestions

Semantic Search in Practice Social Impact of Search Engines

Information Retrieval Theory

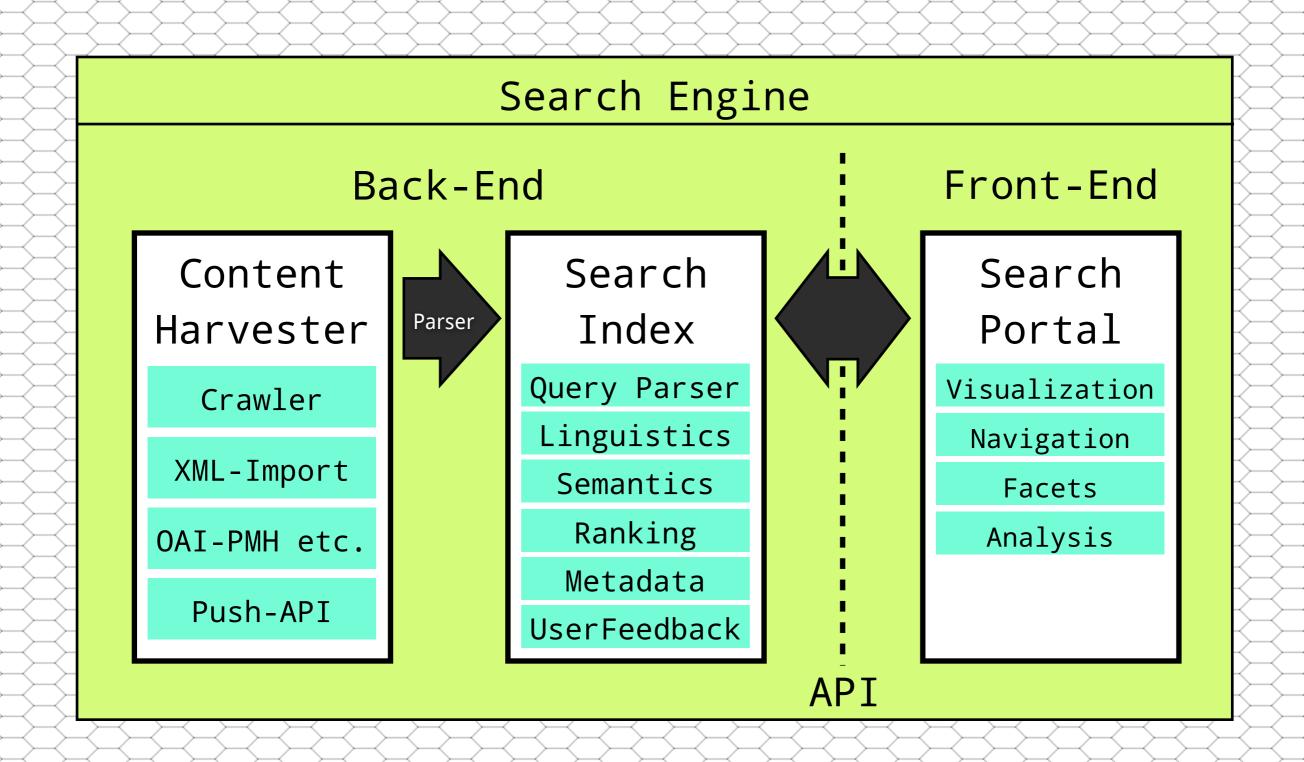


### Add-on Information Retrieval vs. Data Retrieval

	Information Retrieval i.e. Search Engines	<b>Data Retrieval</b> i.e. Databases	
Purpose	retrieve information based on query keywords	retrieve records based on well-formed terms	
Query Language	human language, limited formal syntax, i.e. search modifier	formal language, logic, regular expressions	
Expectation	information about the content, fuzzyness accepted	exact evaluation, single error means failure	
Order	a result ranking based on user expectation, implicit order	none or explicit order may be given in query	
Users	Humans (sometimes Software)	Machines / Software	
Semantics	enrichment about result domain using facets by default	result grouping possible on demand	

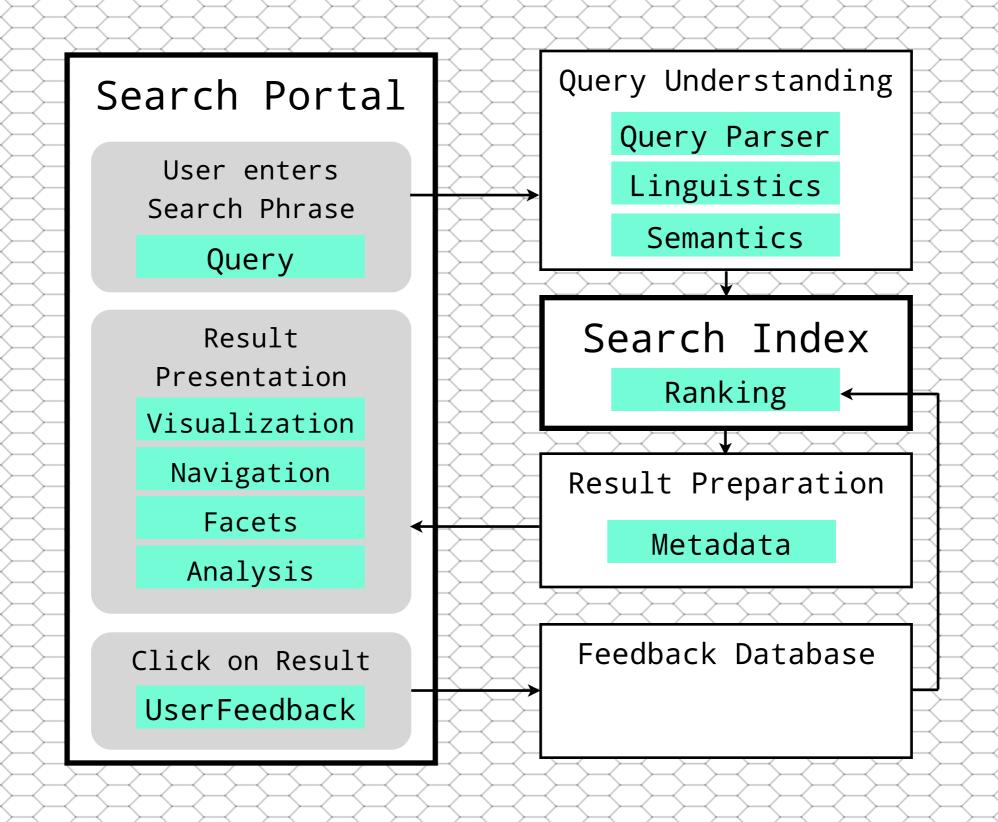


# Add-on Search Engine Portal Prototype Concept



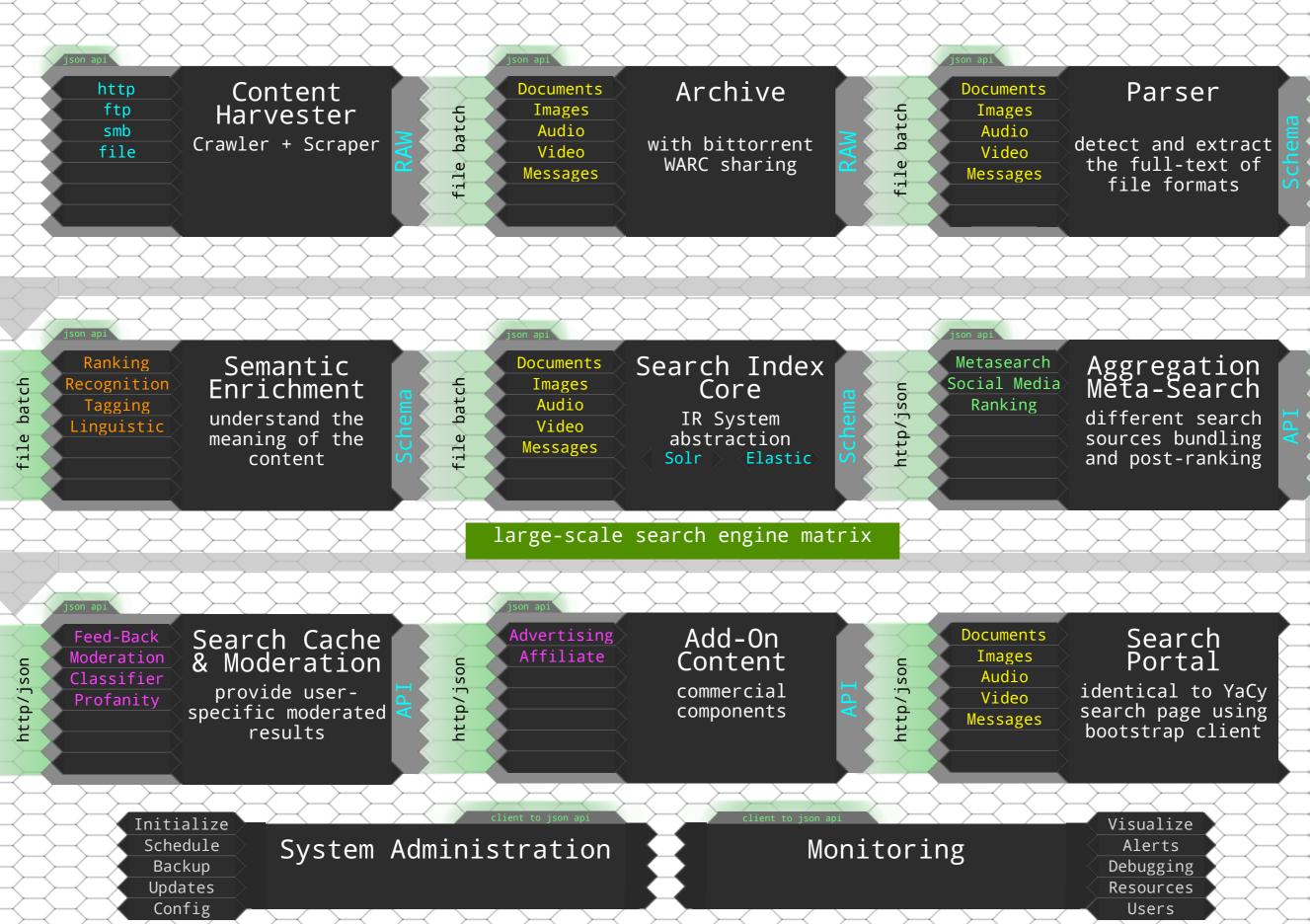


# Add-on Search Engine Portal Prototype Data Flow



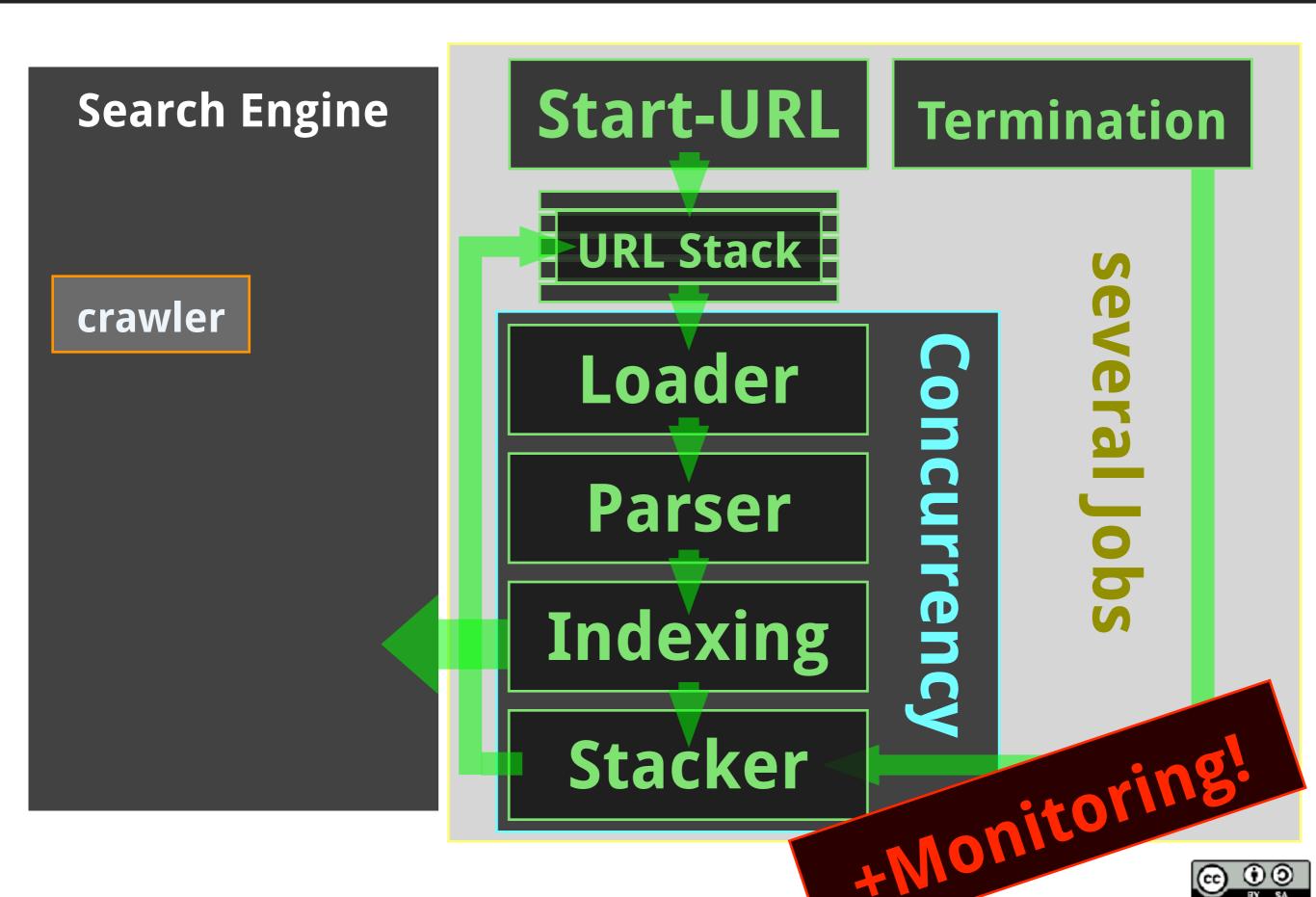


# Add-on Search Engine Portal Prototype Components



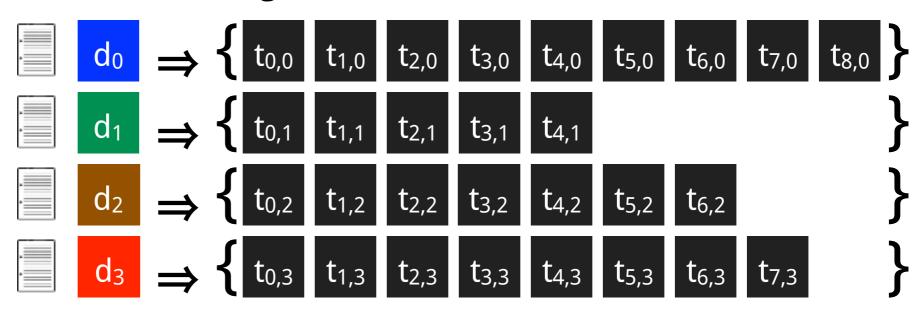
FOSSASIA 2016 Singapore: Ask me Anything about Search Engine History and Future (C) 2016-03-18 [CC-by-SA] by Michael Christen, mc@yacy.net; @Orb1t3r

# Add-on Search Engine Crawler Concept



### Add-on Information Retrieval Indexing Process

Corpus: every document is a ,bag of words'



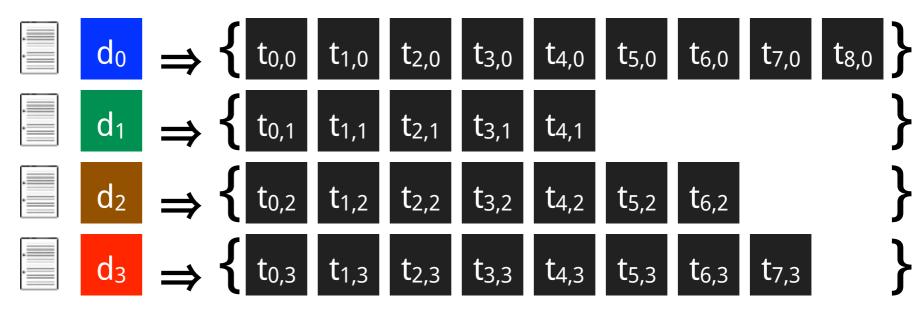
**Index:** all same words are referenced by a list of doc-ids

• *Indexing* is the process to create an *inverted index* from a (set of) document(s).



### Add-on Information Retrieval Indexing Process

Corpus: every document is a ,bag of words'



**Index:** all same words are referenced by a list of doc-ids

• *Indexing* is the process to create an *inverted index* from a (set of) document(s).

Terms are equal, ≃' in the context of stemming rules, synonyms, localization techniques, phonetic rewrites, partial match algorithms, stopword rules, and other techniques

An inverted index consist of inverted lists sorted by document identifier.

