

# Searching the Web of Objects



**Ricardo Baeza-Yates**

**VP of Yahoo! Research  
for EMEA & LatAm  
Barcelona, Spain**

## Content and Metadata trends

Content type	Amount of content produced per day
Published content	3-4 GB
Professional web content	~ 2 GB
User generated content	8-10 GB
Private text content	~ 3 TB (300x more)
Upper bound on typed content	~700 TB (~200x more)

Metadata type	Amount of metadata produced per day
Anchortext	100 MB
Tags	40 MB
Pageviews	180 GB
Reviews	Around 10 MB

[Ramakrishnan and Tomkins 2007]

## Examples

Wordnet

**Explicit**

Metadata

RDF

Wikipedia ODP

Y! Answers

Flickr

**UGC**

**Implicit**

Blogs,  
Groups

**Text**

**Scale**

**Anchors + links**

Queries+clicks

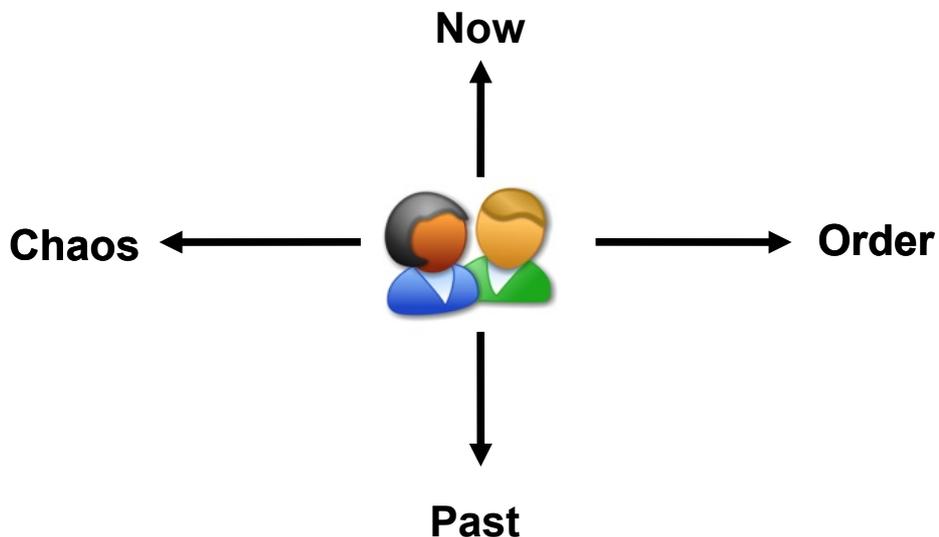
**Private**

**Quality?** <sub>6</sub>

## Trends

- User Generated Content
  - Massive (quality vs. quantity)
  - Social Networks
  - Real time (people + physical sensors)
- Impact
  - Fragmentation of ownership
  - Fragmentation of access (longer tail)
  - Fragmentation of right to access
- Viability
  - Business model based in advertising (?)

## What we really want?



## Search is Evolving

- Already, more than a list of docs
- Moving towards identifying a user's task
- Enabling means for task completion
- New experiences based on the Web 2.0
- Challenges: on-line, scalability

# More complete information in one search

The screenshot shows a Yahoo! search result for 'legal sea foods boston ma'. Three red boxes with arrows point to specific features:

- Shortcuts:** A map showing the location of Legal Sea Foods near Boston, with a red box around it.
- Deep Links:** A list of search results for 'Legal Sea Foods - The Standard for Quality and Freshness in the ...' with a red box around the text.
- Enhanced Results:** A detailed result for 'Legal Sea Foods - Waterfront - Boston, MA 02109' with a red box around the text and a small image of a dish.

The screenshot shows a Yahoo! Local search for 'legal sea foods' in Boston, MA 02108. The search results are sorted by 'Top Results | Distance | Highest Rated'. Two results are visible:

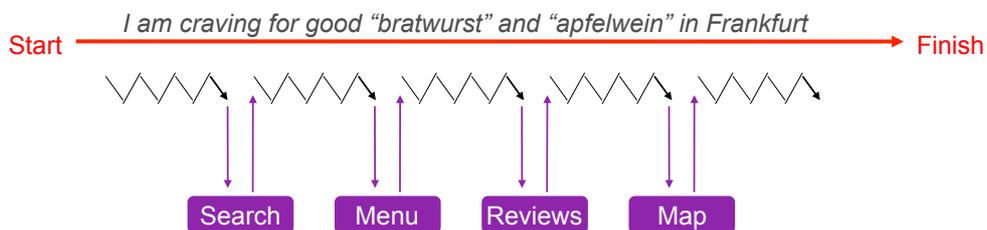
- Legal Sea Foods Restaurants:** 4.5 stars (15 reviews), 0.69 mi. Description: "Larry H... 'We ate at this restaurant last October and it was...' more".
- Legal Sea Foods - Prudential Center:** 4.0 stars (2 reviews), 1.07 mi. Description: "selconierge - 'Boston has many seafood restaurants to choose from in...' more".

A map on the right shows the location of the search results in Boston, MA. A yellow box highlights the 'See Map View' link and the map itself.

# Search: Content vs. Intent

## Premise:

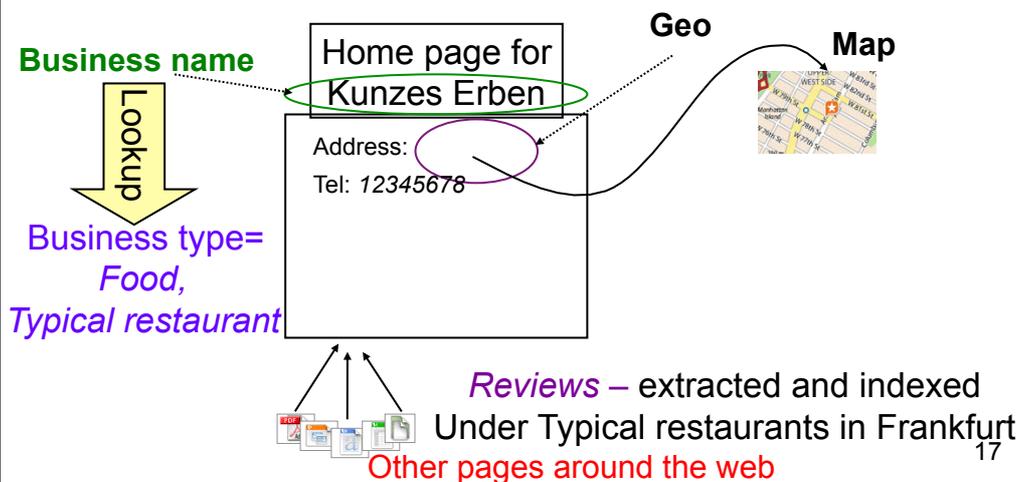
- People don't want to search
- People want to get tasks done and get straight to their answers



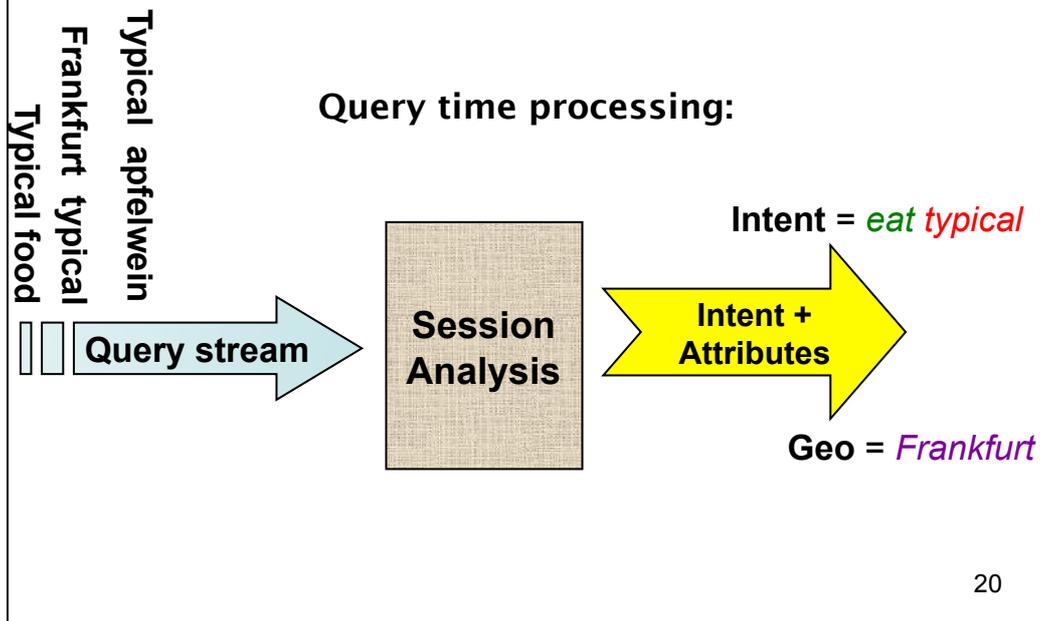
13

# How this might work – I

## Index time processing:



## How this might work – II



## Net

- We move from a web of pages to a **Web of Objects**
- Objects are **people, places, businesses, restaurants** ... (named entities)
- Objects have attributes
  - Missing, noisy, etc.
- Intents are satisfied by presenting **objects and attributes**
- Attributes define faceted search

## Research Challenges

- Crawling objects
- Object extraction (entities)
- Object disambiguation
- Object consolidation
- Object normalization
- Object indexing
- Object ranking
- Object visualization

22

## How do we get structured objects/attributes?

- Web Content
  - Metadata/Taxonomies/Folksonomies
  - Machine learning techniques
  - Classification/Extraction/Semantic Web
- Web Usage
  - Implicit relations
- Building out an open ecosystem
  - Publishers have incentives to contribute

23

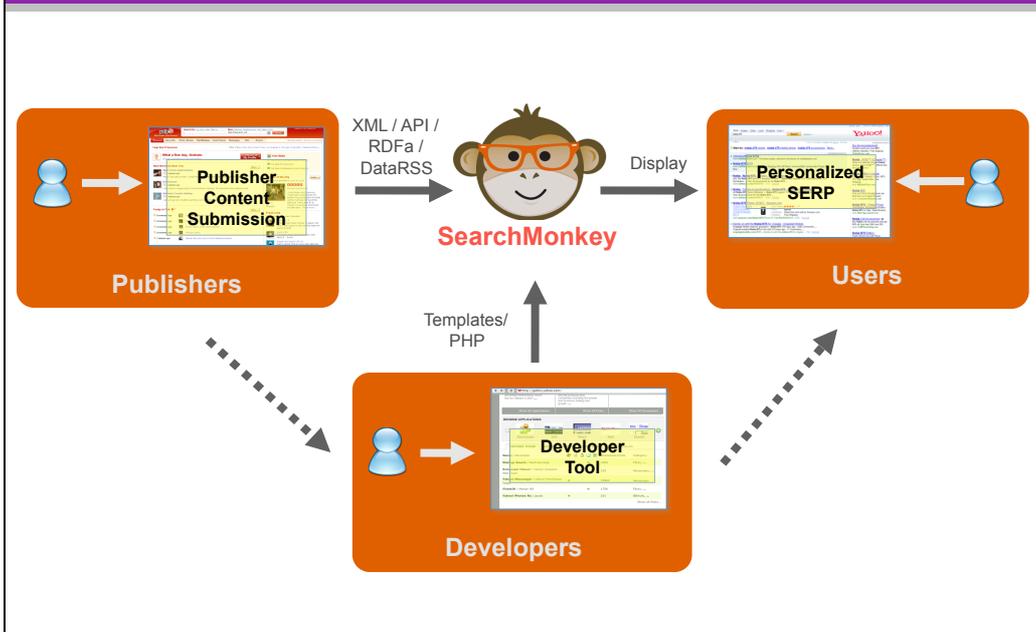
# Search Engine Result Page

The screenshot shows a Yahoo! search engine result page for the query "barcelona". The search bar at the top contains "barcelona" and the search button is labeled "Search". Below the search bar, there are navigation tabs for "Web", "Images", "Video", "Local", "Shopping", and "More".

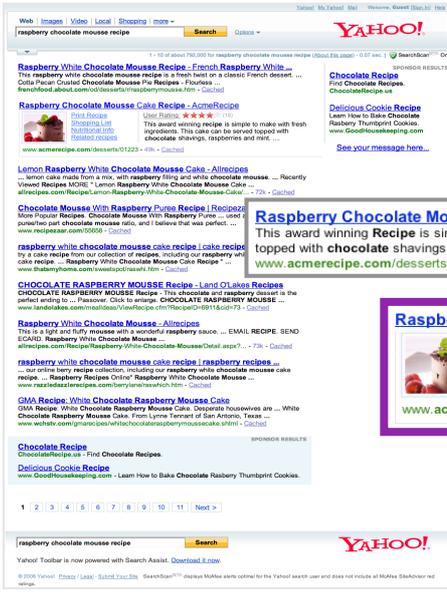
Red circles highlight several key areas:

- Navigation and Search Bar:** The "Web" tab, the search bar containing "barcelona", and the "Search" button.
- Autocomplete/Related Concepts:** A dropdown menu showing suggestions like "barcelona hotels", "barcelona weather", and "barcelona airport". To the right, "Explore related concepts:" lists "Catalonia", "Barcelona city", "population", "Catalan", "restaurants", "Barcelona tourism", "luxury hotels", and "museums".
- Sponsored Results:** A yellow box for "400 Hotels in Barcelona - Spain" with the text "Great rates, guest-reviews. No reservation fee, you pay at the hotel." and a link to "Booking.com/hotels-in-barcelona".
- Main Results:** A result for "Barcelona, Spain - Visitor Guide" from "travel.yahoo.com" with links for "Hotels", "Restaurant Guide", "Flights", and "Map". Below it, a "Top Rated Things To Do (314)" list includes "Sagrada Família", "Ramblas", and "Casa Milà (La Pedrera)".
- Wikipedia Snippet:** A snippet for "Barcelona - Wikipedia, the free encyclopedia" with a brief description and a link to "wikipedia.org/wiki/Barcelona".
- Left Sidebar:** A sidebar with "View Notes (1)", "SearchScan - On", and a list of related terms like "Wikipedia", "Yahoo! Travel", "Wikitravel", and "Answers.com".
- Right Sidebar:** A "Sponsored Results" box for "What to Do Barcelona" from Viator, and a "SearchMonkey" logo.

# The SearchMonkey Ecosystem

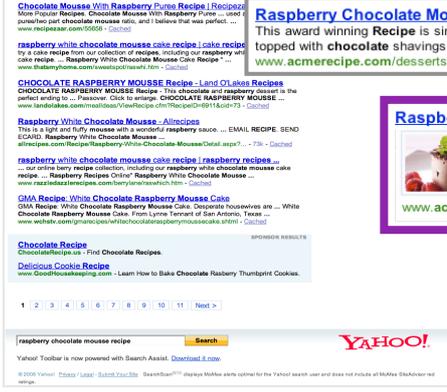


# Site Owners



- Straight to the answers
- Increase in quality of site traffic
- Fosters loyalty and engagement

**BEFORE**



- Straight to the answers
- Increase in quality of site traffic
- Fosters loyalty and engagement

**AFTER**

Enhanced Results and Infobars:  
**Ratings and reviews, images, deep links, and other name-value pairs.**

# The Wisdom of Crowds

- James Surowiecki, a *New Yorker* columnist, published this book in 2004
  - “Under the right circumstances, groups are remarkably intelligent”

- Importance of diversity, independence and decentralization **Aggregating data**

*“large groups of people are smarter than an elite few, no matter how brilliant—they are better at solving problems, fostering innovation, coming to wise decisions, even predicting the future”.*

# The Wisdom of Crowds

- Crucial for Search Ranking
- Text: Web Writers & Editors
  - not only for the Web!
- Links: Web Publishers
- Tags: Web Taggers
- Queries: All Web Users!
  - Queries and actions (or no action!)

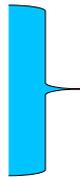
30

# Geo-tagged Photos in Flickr

The image shows a screenshot of the Flickr website interface. At the top, the Flickr logo is visible, along with navigation links like 'Home', 'The Tour', 'Sign Up', 'Explore', and 'Upload'. A search bar contains the text 'frankfurt'. Below the search bar, a map of Frankfurt, Germany, is displayed. The map is overlaid with numerous small thumbnail images, representing geo-tagged photos. A search bar at the bottom of the map area contains the text '142,032 geotagged items' and 'Sort by: Interesting • Recent'. The map shows various districts and landmarks of Frankfurt, with the Main River and Main Tower visible. The interface includes standard map controls like zoom in/out buttons and a 'Link to this map' button.

## The Wisdom of Crowds

- Popularity
- Diversity
- Quality
- Coverage



**Long tail**

## The Long Tail

Explore Flickr through tags

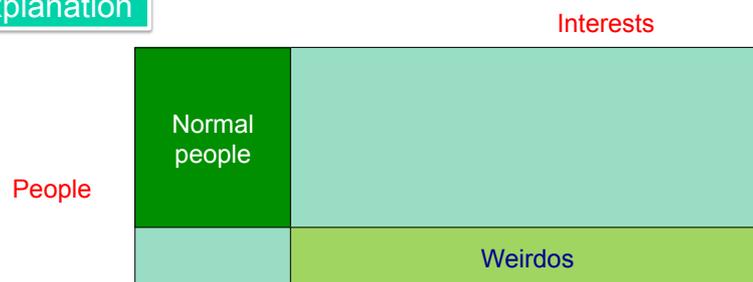
architecture **art** australia **beach** birthday blue bw **california** canada  
**canon** china christmas **city** concert england **europa** **family** festival flower  
flowers food **france** friends fun germany green **italy** **japan** london  
**music** **nature** new newyork night **nikon** nyc paris park **party**  
people portrait red sanfrancisco sky **snow** spain street **summer** sunset taiwan  
**travel** trip uk **usa** vacation water **wedding** white winter

## Heavy tail of user interests

Many queries, each asked very few times, make up a large fraction of all queries

Movies watched, blogs read, words used ...

One explanation



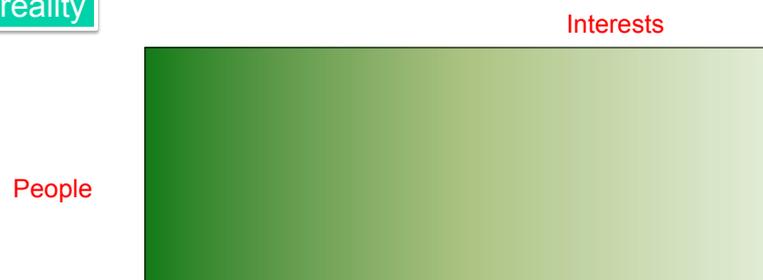
## Personal distribution has a heavy tail

Many queries, each asked very few times, make up a large fraction of all queries

Applies to word usage, web page access ...

We are all partially eclectic

The reality



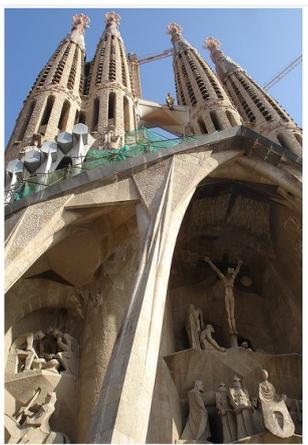
Broder, Gabrilovich, Goel, Pang; WSDM 2009

## Why the heavy tail matters

Not because the worst-sellers make a lot of money

But because they matter to a lot of people

## Tag Mining - Collective Knowledge



- Many users annotate photos of “La Sagrada Familia”:
  - Sagrada Familia, Barcelona
  - Sagrada Familia, Gaudi, architecture, church
  - church, Sagrada Familia
  - Sagrada Familia, Barcelona, Spain
- Derived collective knowledge:
  - Barcelona, Gaudi, church, architecture

Van Zwol et al, 2008 - 2010

# Tag Explorer

**TagExplorer**  
Powered by Flickr

Query: paris ✕

locations

europe + france + ile de france +

london + montmartre +

subjects

arc de triomphe + art +

eiffel tower + louvre +

museum + notre dame + seine +

activities

travel + trip +

time

2005 + 2006 + 2007 + vacation +

**Photo Results**

**Photo Details**

Paris - Gare du Pont Cardinet - 28-07-2007 - 9h03

Taken by: [Panoramas](#)

[View photo on Flickr](#)

Tags: bridge sky panorama paris france reflection rain de pluie pod gare perspective pluie ponte reflet most ciel zebra pont brug angkor passage brücke chemin petite 1925 1920 fer ptassembler lpburi rer panneaux köprü narai ceinture piétons angkorian moct cardinet batgnolles yéppa smartblend indicateurs étennecazin angkorien

Copyright © 2008 Yahoo! All Rights reserved. [Privacy Policy](#) - [Terms of Service](#) - [Copyright/IP Policy](#)

**Endless image browsing**

# Could suggest tags: nice but ....

**London Eye**

London Eye and Golden Jubilee Bridge seen from Westminster Bridge.

**Tag list**

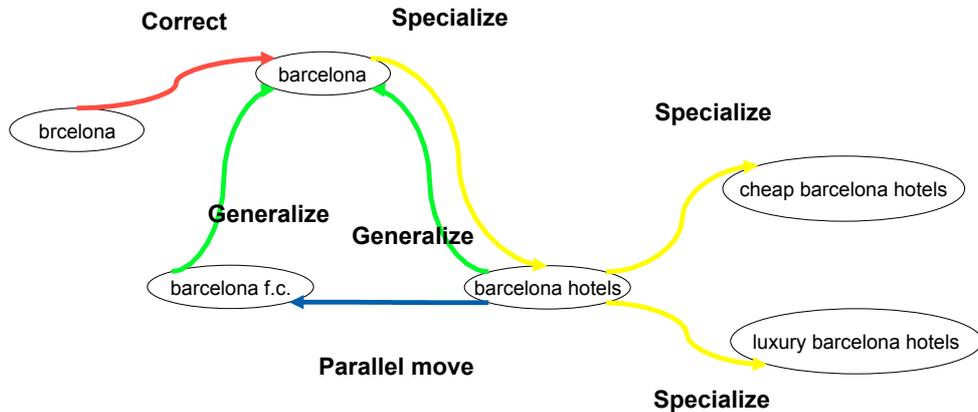
london eye, thames,

**Suggested tags**

- london
- england
- uk
- river
- eye
- south bank
- big ben
- night
- bridge
- 2006



# Query-reformulation types

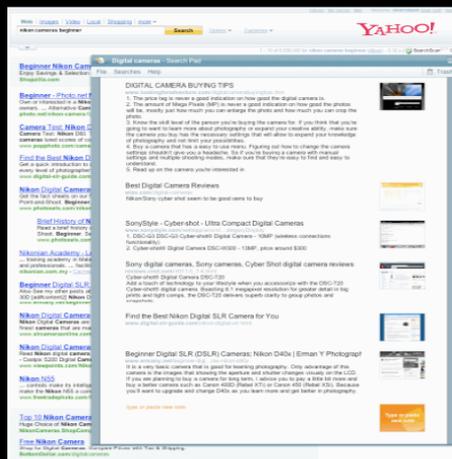


Bonchi, Donato, Castill, etc. Query flow model, 2009

# Search Pad

“... keeps track of search query terms ... when it **detects a trend**, offers to save the result in an online document.”

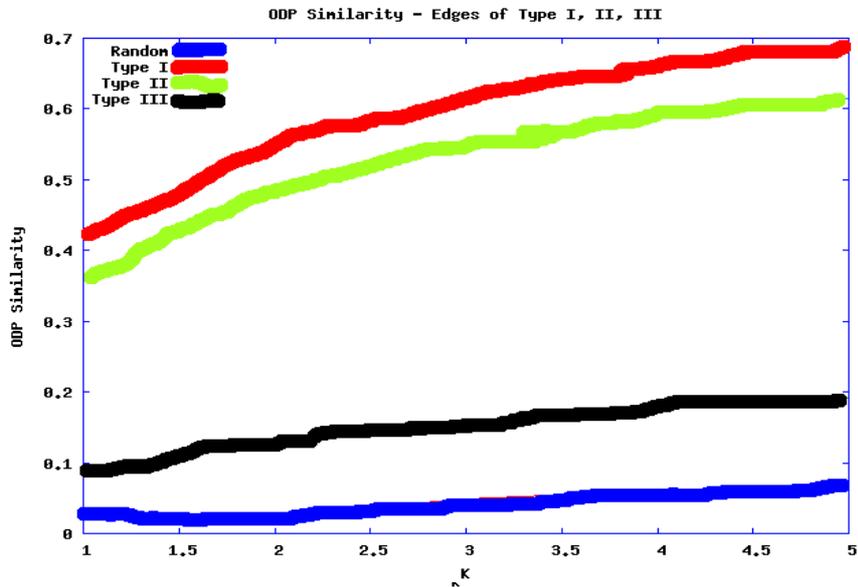
CNET



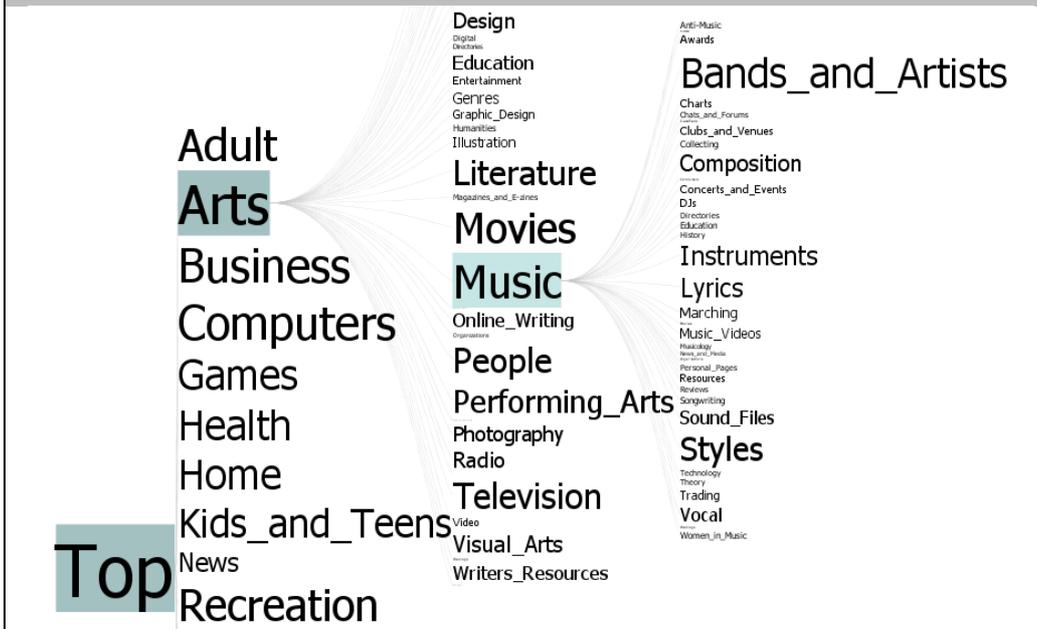


# Experimental Evaluation

Baeza-Yates & Tiberi, KDD 2007



# Hierarchical Clustering





# Exploiting Microformats

**Geolocation**

**Rich abstract**

**Related pages based on metadata**

**Events from personal calendar, Conferences, and bio from LinkedIn**

# Bridging implicit and explicit metadata

**Pablo Ruiz Picasso** (October 25, 1881 – April 8, 1973), often referred to simply as **Picasso**, was a Spanish painter and sculptor. His full name is **Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Clito Ruiz y Picasso**.<sup>[1]</sup> One of the most recognized figures in 20th century art, he is best known as the co-founder, along with Georges Braque, of cubism.

**Biography** [edit]

Pablo Picasso was born in Málaga, Spain, the first child of José Ruiz y Blasco and María Picasso y López. He was christened with the names Pablo, Diego, José, Francisco de Paula, Juan Nepomuceno, María de los Remedios, and Cipriano de la Santísima Trinidad.<sup>[2]</sup> Picasso's father was a painter whose specialty was the naturalistic depiction of birds and who for most of his life was also a professor of art at the School of Crafts and a curator of a local museum. The young Picasso showed a passion and a skill for drawing from an early age; according to his mother,<sup>[3]</sup> his first word was "piz," a shortening of *lápiz*, the Spanish word for pencil.<sup>[4]</sup> It was from his father that Picasso had his first formal academic art training, such as figure drawing and painting in oil. Although Picasso attended art schools throughout his childhood, often those where his father taught, he never finished his college-level course of study at the Academy of Arts

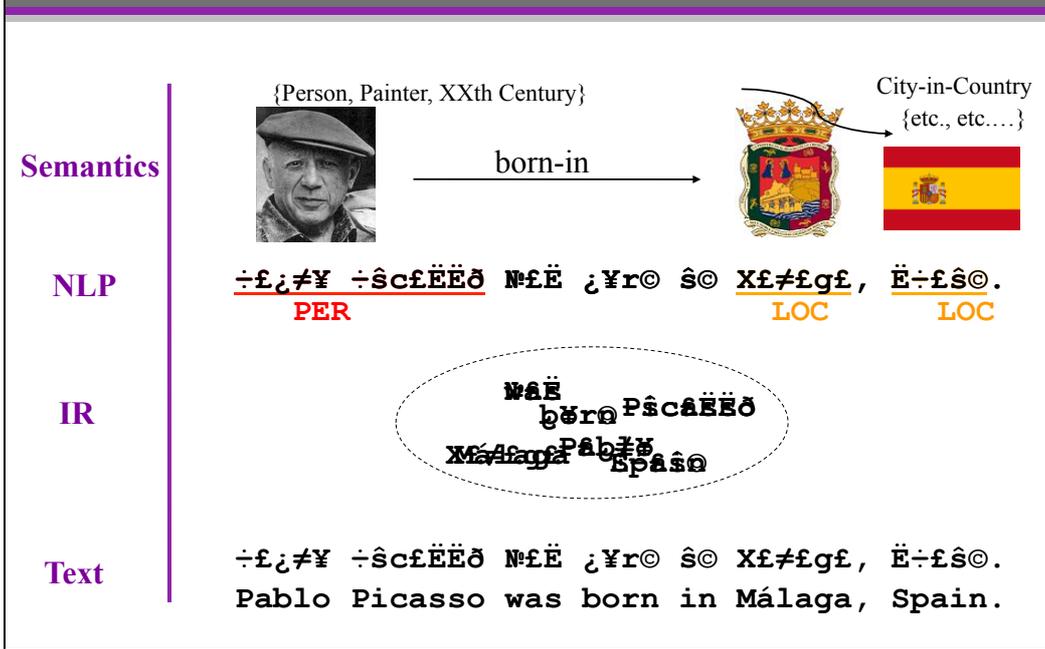
**Pablo Picasso**

**Birth name** Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Martyr Patricio Clito Ruiz y Picasso

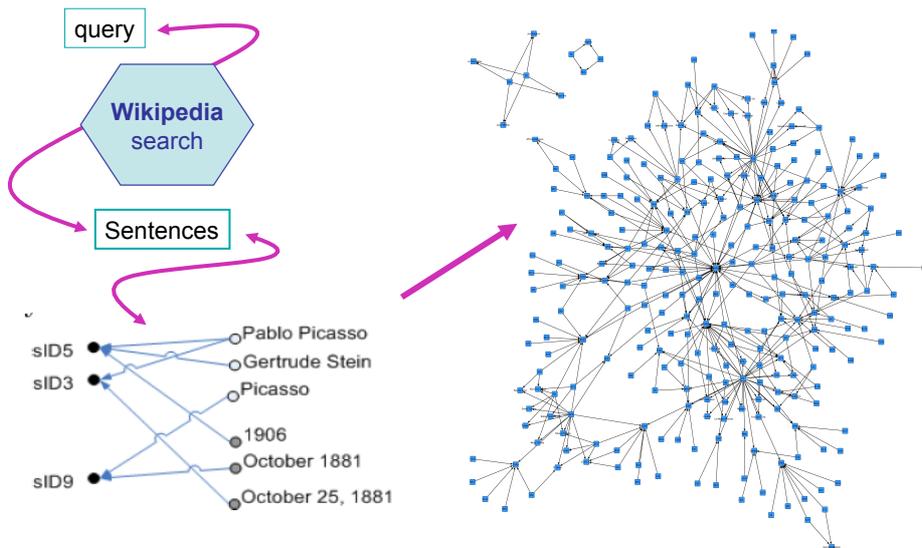
**Born** October 25, 1881  
Málaga, Spain

**Died** April 8, 1973 (aged 91)  
Mougins, France

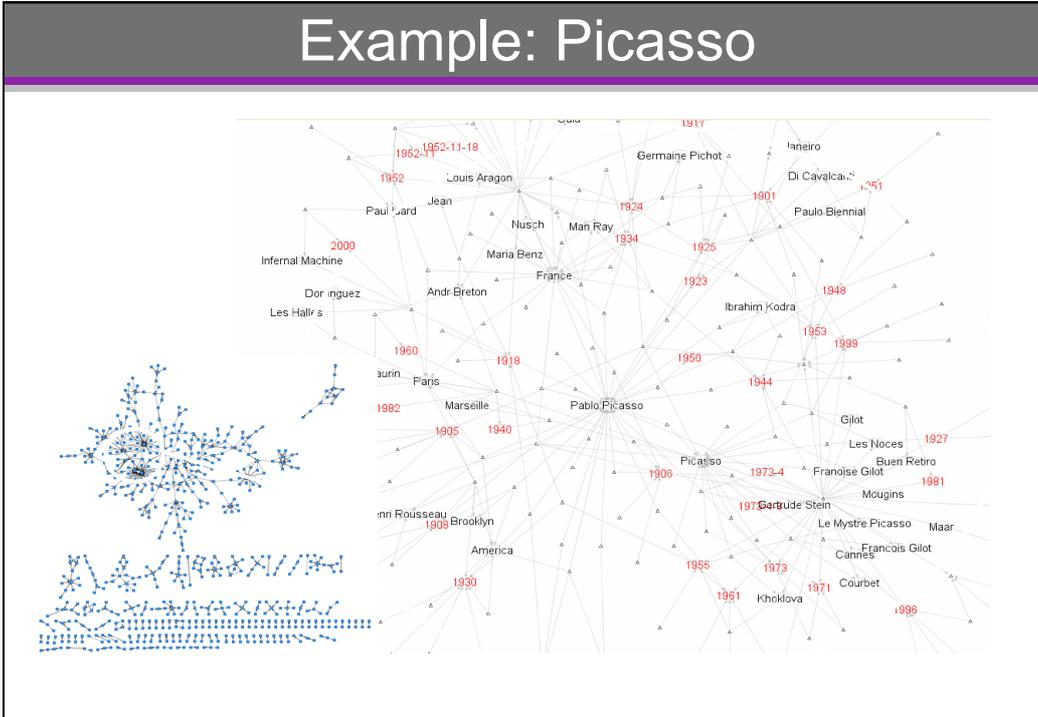
# Language, Text, Search & "Semantics"



## Entity Graph (linear time entity extraction)



# Example: Picasso



# Correlator: Relations in Wikipedia

**1905: While spacetime can be viewed as a consequence of Albert Einstein &#**

While spacetime can be viewed as a consequence of Albert Einstein's 1905 theory of special relativity, it was first explicitly proposed mathematically by one of his teachers, the mathematician Hermann Minkowski, in a 1908 essay building on and extending Einstein's work.

Sun, 01 Jan 1905 00:00:00 GMT

Search

Wikipedia Names Places Events Concepts News Answers

"Albert Einstein"

was named Person of the Century

- 1905: While spacetime can be viewed as a consequence of Albert Einstein &#
- 1907: on the heat capacities of solids with quantized energy levels by Albert
- 1915: Pick in 1911 introduced Albert Einstein to the work of Italian math
- 1916: He married Esther Einstein ( b. ca. 1888, Austria/Galicia ; imm. ca. 1891
- 1917: Albert Einstein, in his paper Zur Quantentheorie der Strah
- 1921: He was fascinated by the theories
- 1923: Einstein, Albert (
- 1924: , together

Timeline © SHALE

Events in the timeline

1879 - 1955

(From [W List of mathematicians \(E\)](#)) "Einstein, Albert (Germany/USA, 1879 - 1955)"

(From [W List of Swiss Americans](#)) "Albert Einstein (1879 - 1955) theoretical physicist widely regarded as the most important scientist of the 20th century and one of the greatest physicists of all time"

(From [W Einstein High School](#)) "Albert Einstein High School in Montgomery County, Maryland"

(From [W Albert Einstein High School](#)) "Albert Einstein High School is known to be one of the

State

New Jersey

# New Pages

- For topics without a Wikipedia page, Correlator creates a “synthetic page” with an overview of the topic
- Query:
  - art deco chicao
- Synthetic page:
  - Defines Art Deco
  - Defines Chicago
  - Shows relations between Art Deco and Chicago

The screenshot shows the Correlator interface with a search bar containing 'art deco chicao'. Below the search bar are icons for Wikipedia, Names, Places, Events, Concepts, News, and Answers. The main content area displays search results for 'Art Deco' and 'Chicago'. The 'Art Deco' section includes a paragraph about the movement and a link to 'W Art Deco: View full article'. The 'Chicago' section includes a paragraph about the city and a link to 'W Chicago: View full article'. Below these are category-based results for '1930 architecture' and 'Skyscrapers in Chicago', each with a list of related articles and a 'View more entries' link.

# Synthetic page - example

## Category: Dinosaurs of South America

**W Buitreraptor**: It was found in **Argentina** and was described in 2005 .The fossilised bones were found in 2005 in sandstone in Patagonia , **Argentina** - by an excavation lead by Petr Májovický , curator of **dinosaurs** at the Field Museum in Chicago ). Buitreraptor was discovered in the same fossil site that had earlier yielded Giganotosaurus , one of the largest known carnivorous **dinosaurs** .

**W Herrererasaurus**: Herrererasaurus ( meaning " Herrera 's lizard , " after the name of the rancher who discovered the first fossil of the animal ) was one of the earliest **dinosaurs** . This view is further supported by ichnological records showing large tridactyl footprints that can be attributed only to a theropod **dinosaur** , dating from the Ladinian ( Middle Triassic ) of the Los Rastros Formation in **Argentina** and predating Herrererasaurus by 3 to 5 million years . The importance of Herrererasaurus and Eoraptor lies in the fact that their remains allow for directly testing the idea of **dinosaurs** being a monophyletic group , i.e. all **dinosaurs** have a common ancestor .

**W Unaysaurus**: It was recovered from the red beds of the Santa Maria Formation ( also known as the Caturrita Formation ) , which is the geologic formation where similarly old **dinosaurs** like Saturnalia have been found .The oldest **dinosaurs** in the world are from here and nearby in **Argentina** ( like the Eoraptor ) , which suggests that the first **dinosaurs** may have originated in the area .

**W Carnotaurus**: Carnotaurus ( pronounced /,karnot'ɔrɛs/ KAHR-noh-TAWR-us ; meaning " meat-bull " , referring to its distinct bull-like horns ( Latin carne = flesh + Greek ταύρος = Bull ) was a large predatory **dinosaur** , with horns vaguely resembling a bull's . Carnotaurus lived in Patagonia , **Argentina** during the Maastrichtian stage of the Late Cretaceous , and was discovered by José F. Bonaparte , who has uncovered many other bizarre South American **dinosaurs** . ... Together , these **dinosaurs** form the subfamily Carnotaurinae in the family Abelisauridae .

**W Eoraptor**: Eoraptor was one of the world 's earliest **dinosaurs** . ... Early **dinosaur**The bones of this primitive **dinosaur** were first discovered in 1991 , by University of Chicago paleontologist Paul Sereno , in the Ischigualasto Basin of **Argentina** .

**W Argyrosaurus**: Argyrosaurus ( pronounced /,ɑrʒɪ'rɔʊ'sɔrɛs/ AHR-j-ro-SAWR-us ) meaning " Silver lizard " , because it was discovered in **Argentina** , which is sometimes known as " Silver land " ( Greek argyros meaning " silver " and sauros meaning " lizard ) was a genus of herbivorous titanosaurid **dinosaur** that lived about 70 million years ago , during the Late Cretaceous Period of what is now South America ( **Argentina** and Uruguay ) . It was one of the largest **dinosaurs** , having a height of 8 metres , a length of up to 20-30 metres and a weight of up to 80 tonnes . It was a herbivore .

**W Argentinosaurus**: Argentinosaurus ( meaning " **Argentina** lizard " ) was a herbivorous sauropod **dinosaur** genus that was among the largest land animals that ever lived . ... Argentinosaurus is featured prominently in the permanent exhibition G ants of the Mesozoic at Fernbank Museum of Natural History in Atlanta , Georgia , USA . This display depicts a hypothetical encounter between Argentinosaurus and the carnivorous theropod **dinosaur** Giganotosaurus . ... At 123 feet long , this skeletal reconstruction represents the largest **dinosaur** mount ever to be assembled .

**W Neuquensaurus**: Neuquensaurus ( meaning " Neuquén lizard " ) was a titanosaur sauropod **dinosaur** that appeared in the Late Cretaceous , 71 million years ago in **Argentina** and Uruguay in South America . This **dinosaur** was 10-15 meters ( 34-51 feet ) long , and is believed to have possessed armor-like osteoderms .

**W Genyodectes**: Genyodectes ( Woodward , 1901 ) is a genus of ceratosaurian theropod **dinosaur** from the Lower Cretaceous of South America . The holotype material ( MLP 26-39 , Museo de La Plata , La Plata , **Argentina** ) was collected from the Cañadón Grande , Departamento Paso de Indios in the Chubut Province of **Argentina** and consists of an incomplete snout , including the premaxillae , portions of both maxillae , the right and left dentary , many teeth , a fragment of the left splenial , and parts of the supradentaries .

# Time Explorer

- **Finding Relations among Entities in News**

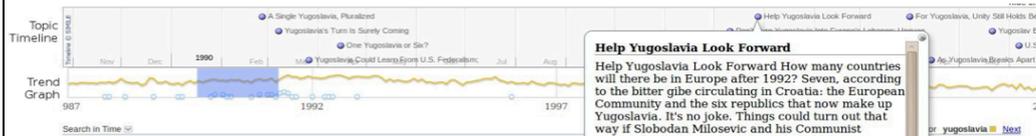
- Past, present or future!
  - Baeza-Yates, Searching the Future, 2005.
- The clue is the interface
- Part of the Living Knowledge EU project

- **Winner of the HCIR 2010 Challenge**

- New York Times collection (1987-2007)
- Found many interesting examples
- Generates new NLP research problems

64

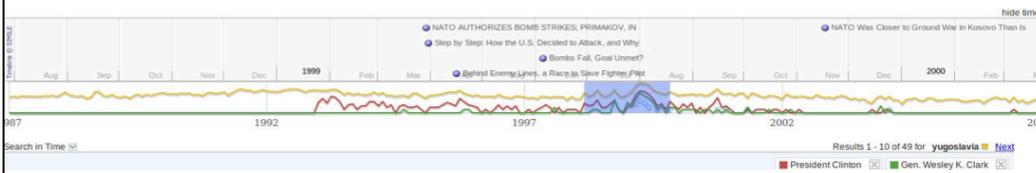
# Time Explorer



(a) Basic Timeline



(b) Time period Selection



(c) Timelijnne with entity trends

# Time Explorer



Figure 4: Searching the Future

**Judges at The Hague Refuse To Halt the NATO Bombing**  
Judges at the International Court of Justice in The Hague refused today to order NATO countries to halt their bombing of Yugoslavia. The judges... veiled language to say the bombing was breaking international law. "The Court is profoundly concerned about the use of force in Yugoslavia," the acting president, Judge Christopher Weeramaney, said.

**Yugoslavia Agrees to Visit by U.N. Team**  
Yugoslavia's representative here, delivered a letter welcoming the mission and assuring that Belgrade would facilitate the trip. The advance team is to discuss arrangements for the arrival.

The New York Times is an American daily newspaper founded and continuously published in New York City since 1851. Although it remains both the largest local metropolitan newspaper in the United States as well as third largest overall behind The Wall Street Journal and USA Today... [Read article at Wikipedia](#)

popularity:

Bias information: The Times has been variously described as having a liberal bias or described as being a liberal newspaper, or of having a conservative bias on certain issues by some writers. [Read article at Wikipedia](#)

Figure 5: Results

# Time Explorer

Living Knowledge Time Explorer

spain Future Search

YAHOO! RESEARCH

The screenshot shows the Time Explorer interface with a search bar containing 'spain'. A timeline from 2010 to 2050 is displayed, with a blue highlight on the year 2010. A list of events is shown on the right side of the timeline:

- Tense Time For Cities Vying to Stage 2010 Games
- Spain Tries Electricity Shaken, Not Stirred
- Eastern Europe, Post Communism: Five Years Later - A
- Spain's Ageless Beauties

Search in Time: Results 1 - 10 of 11 for Spain

**World Briefing | Europe: Plan For North-South Cooperation**  
The participants, meeting in Valencia, Spain, plan to create a free-trade area by 2010 and to push for greater investment in the Mediterranean economies.

**Spain's Ageless Beauties**  
By 2010, all of the more than 90 paradores in Spain's network, which is growing, will have been refreshed and modernized, making them as enticing as any city hotel.

## Implicit Search

### Search as a back end process

- **Trigger the right search depending on the context**
  - Writing email
  - Browsing news
- **Research challenges**
  - More on query intent prediction
  - Whole page layout optimization
  - Exploit the social layer
  - Interactive manipulation of the answer
  - Measure user engagement in any context
  - Compare user satisfaction across Web sites

68

## So what's next?

We are far from being done with innovation in search engines

- **Possible future**
  - The new frontiers: front-end and user experience
    - The most probable reason for users to switch between quasi-equivalent engines is a better user experience
  - We still don't understand well information needs (will we ever? brain electrodes don't work 😊)
  - New search: contextual content delivery
- **Large scale usage data is key to getting there BUT**

69

## Three major conflicting factors

Usage data at a very large scale	
More data over longer periods of time brings more insights	
Contextualization	Privacy
More data via larger communities, makes data less personalized  <i>wisdom of crowds does not work well on small corpora</i>	Over contextualization endangers privacy  Long-term logs endanger privacy

**Contextualize the task: query intent detection**

70

## Conclusions

- Web search is no longer about document retrieval
  - Means for web-mediated goals
  - New breed of search experiences
  - Demands search ecosystem combining content with intent
  - Exploiting the Wisdom of Crowds behind the Web 2.0
  - User aggregation versus personalization
    - Optimize common tasks

**Second edition  
coming next month**

**Modern  
Information Retrieval**  
the concepts and technology behind search  
Second edition



Ricardo Baeza-Yates  
Berthier Ribeiro-Neto



**Questions?**

[rbaeza@acm.org](mailto:rbaeza@acm.org)

<http://search.yahoo.com>

<http://labs.yahoo.com>

<http://sandbox.yahoo.com>