Sistemi per il recupero delle informazioni

Gabriele Pozzani

A.A. 2013/2014

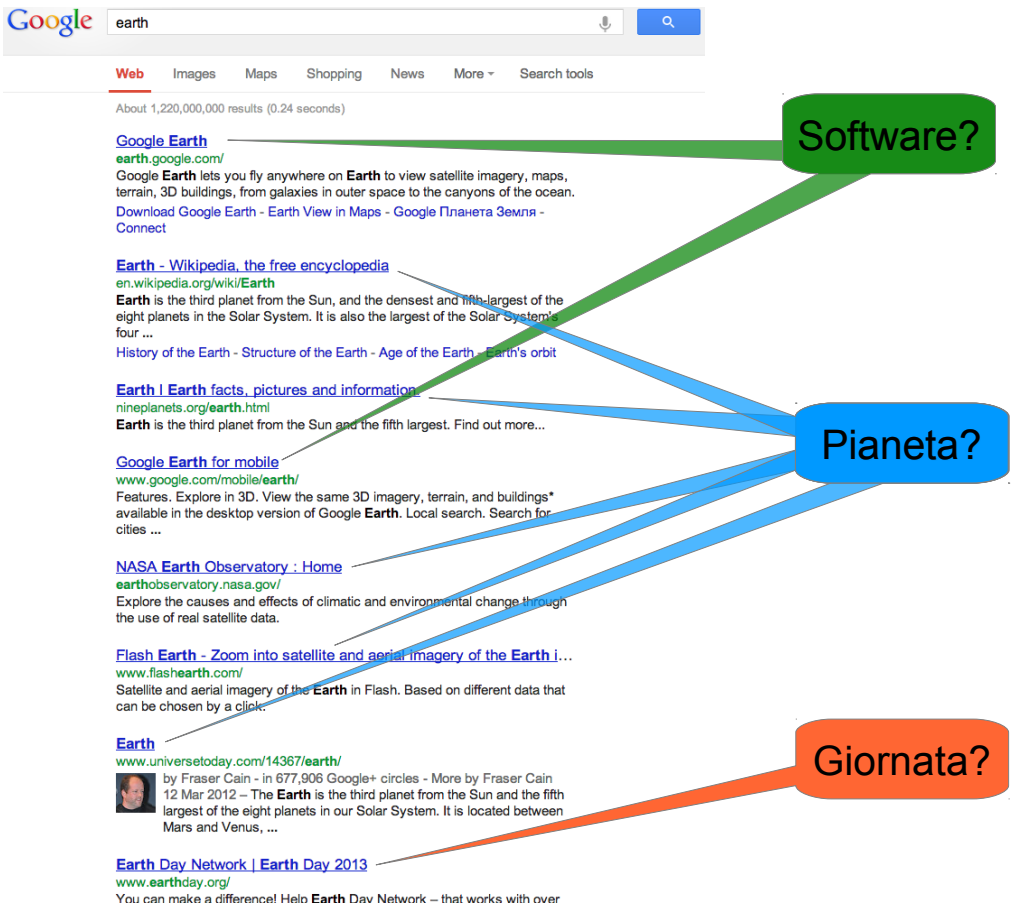
Corso di Laurea Magistrale in Editoria e Giornalismo

Ricerca semantica

Problema (I)

- In ogni lingua, una parola può avere più significati
- Quando si esegue una ricerca con una tale parola, i risultati più rilevanti dipendono da quale significato era inteso dall'utente
- Attualmente (stato in evoluzione), vengono ritornate le pagine più rilevanti per tutti i possibili significati

Problema, esempio



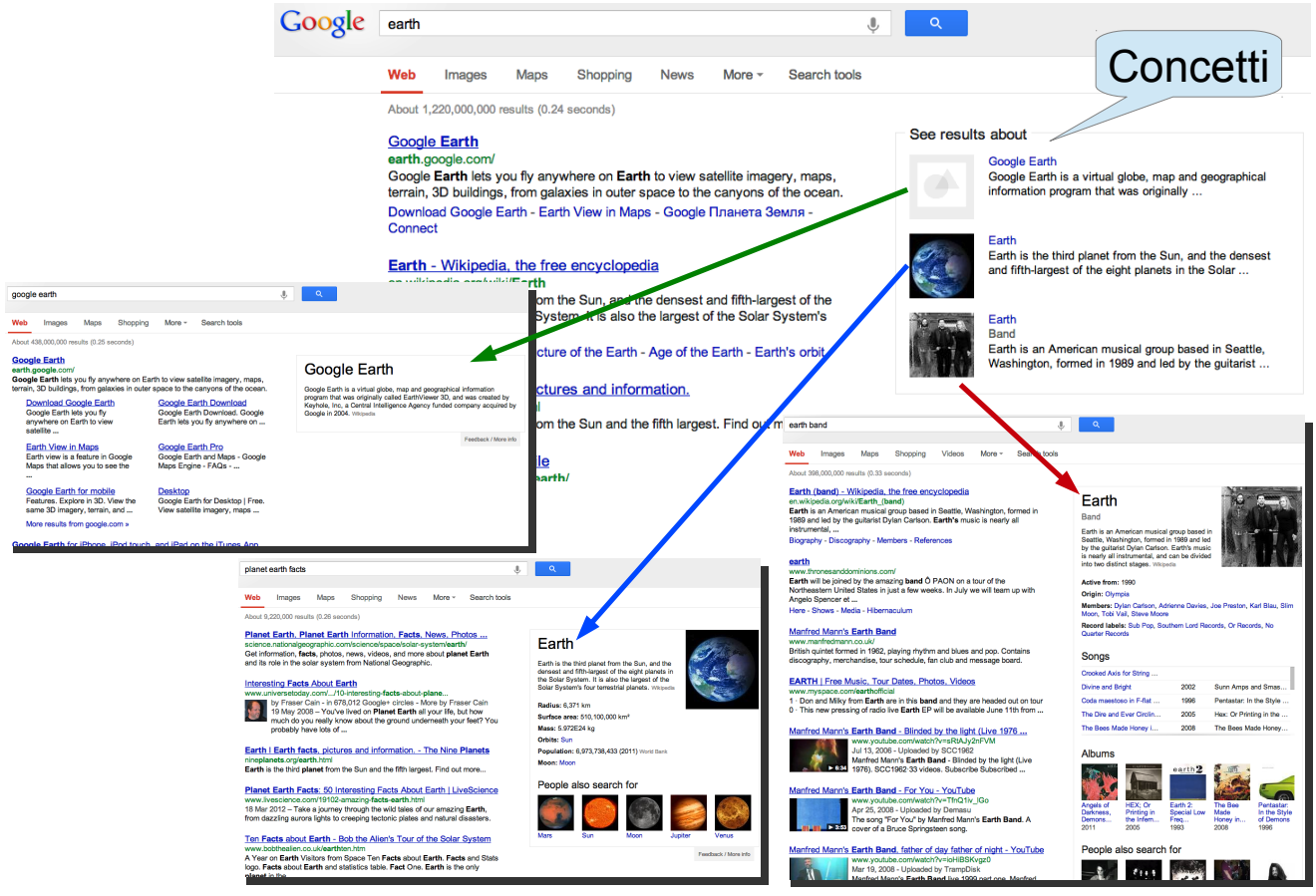
Problema (II)

- “earth” è solo una parola, una stringa, una sequenza di caratteri
- I risultati dei diversi significati sono mescolati tra loro
  - Non sono organizzati in base al significato

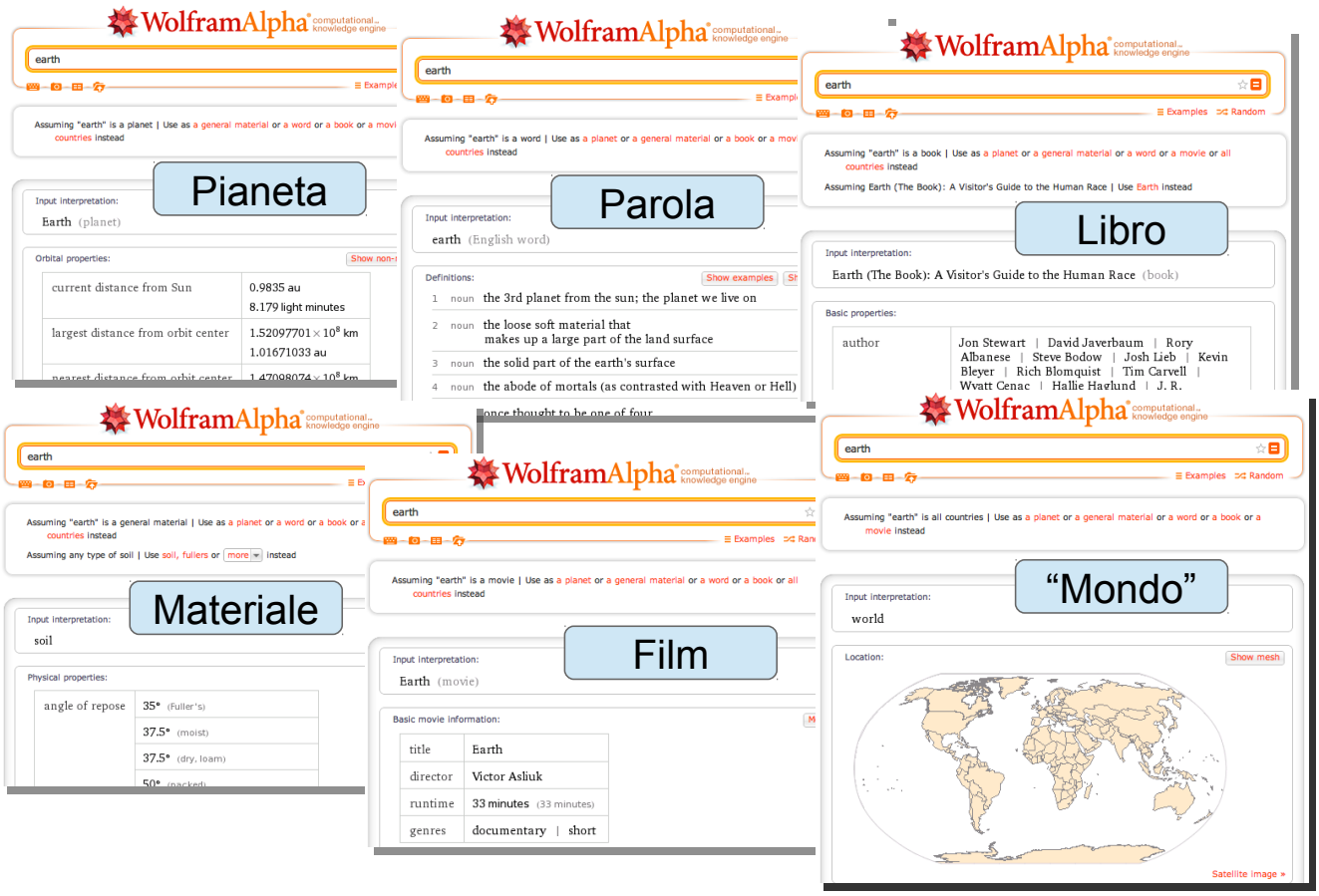
La ricerca semantica

- Obiettivo della ricerca semantica è di
  - aggiungere semantica ai termini
    - Non più stringhe, ma concetti/entità
  - Correlare i documenti non a parole ma ai concetti
  - Organizzare i risultati in base ai concetti
  - Cercare di comprendere quale concetto era sottinteso dall'utente nella propria ricerca (ed esplicitarlo)

Attuale evoluzione



Non solo google



Altri motori di ricerca

- Esistono altri motori di ricerca semantici, anche meno conosciuti
  - Bing
  - GoPubMed
    - Motore di ricerca semantico biomedico
  - Iglue
    - Motore di ricerca semantico con capacità di annotazione/arricchimento semantico dei siti web in realtime
  - Yummly
    - Motore di ricerca semantico per ricette e cibo

Come funziona (I)

- L'idea di base è simile a quella vista per i thesauri
  - I termini riconducibili allo stesso significato e concetto vengono raggruppati in un unico insieme (Synset)
  - I synset vengono collegati tra loro in base alle relazioni semantiche (le più disparate) esistenti tra i concetti che rappresentano
  - Si costruisce così un grafo (rete semantica) in cui
    - I nodi rappresentano i concetti/synset
    - Gli archi rappresentano le relazioni semantiche tra i concetti
  - L'informazione non strutturata (classiche pagine web) deve essere strutturata (grafo) per essere fruita al meglio
    - Vi ricorda qualcosa?

Come funziona (II)

- Si entra nell'ambito della “rappresentazione della conoscenza” e dei “sistemi esperti”
- In questo campo non si parla di “thesauro” ma di “ontologia” e “knowledge base”

Come funziona (III)

- I termini ambigui di una ricerca vengono disambiguati vedendo se appartengono a synset/concetti legati tra loro
  - Termini utilizzati in una stessa ricerca, molto probabilmente, sono correlati tra loro nel significato
    - “red apple” è molto più probabile che abbia il significato di “mela rossa” che non di “Apple® rosso”

Google Knowledge Graph (I)

- Il nome deriva proprio dalla sua rappresentazione a grafo dei concetti
- Una ricerca
  - oltre ad essere eseguita sulle pagine web

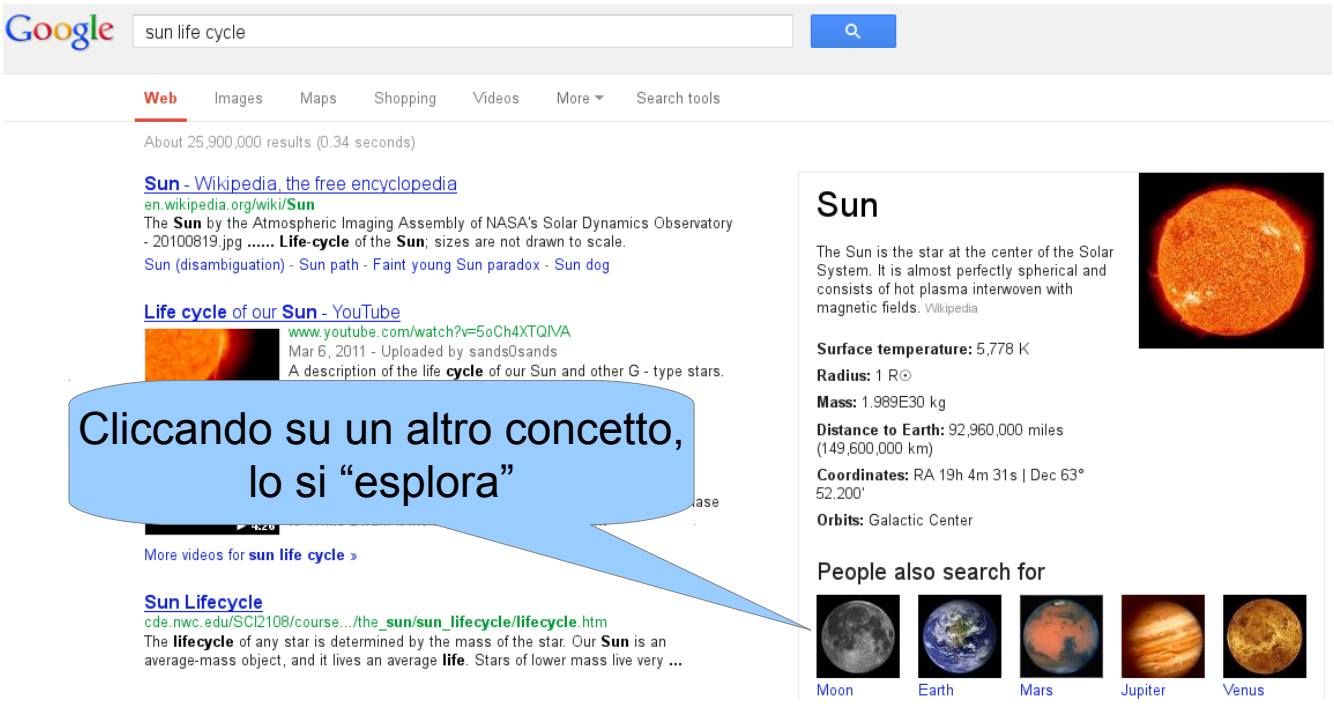
  viene effettuata anche sul knowledge graph
- Quando un termine non può essere disambiguato, tutti i concetti ad esso riconducibili nel grafo vengono mostrati

Google Knowledge Graph, esempio (I)

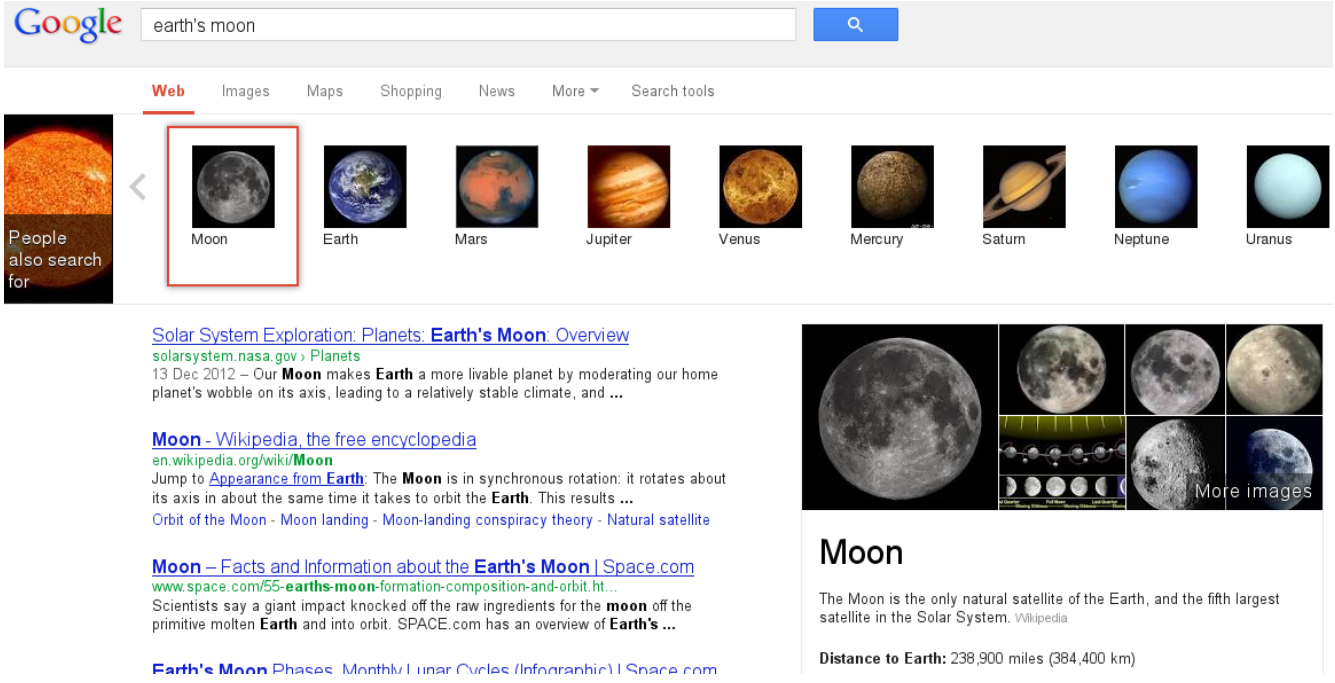
Google Knowledge Graph (II)

- La struttura a grafo permette di navigare tra i concetti
  - Ci si sposta da un nodo/synset/concetto ad altro ad esso collegato
  - La ricerca (Web e/o su KG) corrispondente viene eseguita

Google Knowledge Graph, esempio (II)



Google Knowledge Graph, esempio (III)



Google Knowledge Graph (III)

- Knowledge Graph ha attualmente affiancato e integrato la “classica” ricerca web
- Si può
  - Navigare nel grafo
  - Navigare i risultati della ricerca testuale web
- Inoltre eventuali termini “base” nella ricerca effettuata sul KG vengono ricondotti ad apposite ricerche web classiche, prestabilite anche dai termini che identificano il concetto “in cui ci si trova”

Come viene costruito il KG? (I)

- Informazioni sui concetti vengono tratte da database pubblici, inclusi:
  - Freebase
    - database gratuito e aperto di oltre 24 milioni di cose, tra cui film, libri, spettacoli televisivi, personaggi famosi, luoghi, aziende
  - Wikipedia
  - CIA World Factbook
    - pubblicazione annuale della CIA che riporta i dati statistici fondamentali e una sintesi di informazioni riguardanti tutti i paesi del mondo
  - Risorse specifiche per argomento
    - e.g., Weather Underground per informazioni meteo e la Banca Mondiale per statistiche economiche
  - I dati di ricerca di Google
    - usati per misurare la popolarità di un argomento e contribuire a decidere quali sono le informazioni più richieste

Come viene costruito il KG? (II)

- I concetti e le loro relazioni vengono inferiti
  - Utilizzando thesauri e ontologie già esistenti
  - Automaticamente analizzando i termini chiave delle pagine web e come le pagine web sono collegate dai link
- Secondo alcuni dati disponibili, nel 2012 Knowledge Graph contiene dati su
  - 570M di oggetti
  - 18G di fatti

Google Knowledge Graph (IV)

- Il motto del progetto è “Things, not strings” (“cose, non parole”)
  - Il focus delle ricerche si sposta dai termini/keyword agli oggetti/entità che esse rappresentano
- Google: non più “motore di ricerca”, ma “motore di conoscenza”

Conseguenze sul SEO

- Anche le tecniche di SEO risentono del cambiamento di focus
- Attualmente Google indicizza nel KG solo alcune informazioni delle pagine web
  - persone, prodotti, aziende, ricette, eventi, musica, video
  - Questa informazione deve essere strutturata
    - microdata

Microdata

- Estensione di HTML
- Permette di annotare/arricchire le pagine web (HTML) con nuovi marcatori che includono informazioni semantiche

Microdata, esempio (I)

```
<section> Hello, my name is John Doe, I am a graduate research
assistant at the University of Dreams. My friends call me Johnny.

You can visit my homepage at <a
href="http://www.JohnnyD.com">www.JohnnyD.com</a>.

I live at 1234 Peach Drive Warner Robins, Georgia.</section>
```

```
<section itemscope itemtype="http://schema.org/Person">

   Hello, my name is <span itemprop="name">John Doe</span>,
   I am a <span itemprop="jobTitle">graduate research assistant</span>
   at the <span itemprop="affiliation">University of Dreams</span>.
   My friends call me <span itemprop="additionalName">Johnny</span>.
   You can visit my homepage at
   <a href="http://www.JohnnyD.com" itemprop="url">www.JohnnyD.com</a>.

   <section itemprop="address" itemscope
            itemtype="http://schema.org/PostalAddress">

      I live at <span itemprop="streetAddress">1234 Peach Drive</span>
      <span itemprop="addressLocality">Warner Robins</span>,
      <span itemprop="addressRegion">Georgia</span>.

   </section>

</section>
```


Microdata, esempio (I)

Un web engine comprenderà il microdata come

```
Item
   Type: http://schema.org/Person
   name = John Doe
   jobTitle = graduate research assistant
   affiliation = University of Dreams
   additionalName = Johnny
   url = http://www.johnnyd.com/
   address = Item(1)
Item 1
   Type: http://schema.org/PostalAddress
   streetAddress = 1234 Peach Drive
   addressLocality = Warner Robins
   addressRegion = Georgia
```

Semantica di microdata

- Gli oggetti, con i loro attributi, definibili in microdata sono definiti in http://schema.org

Thing: additionalType, description, image, name, url
- CreativeWork: about, accountablePerson, aggregateRating, alternativeHeadline, associatedMedia, audience, audio, author, award, awards, comment, contentLocation, contentRating, contributor, copyrightHolder, copyrightYear, creator, dateCreated, dateModified, datePublished, discussionUrl, editor, encoding, encodings, genre, headline, inLanguage, interactionCount, isFamilyFriendly, keywords, mentions, offers, provider, publisher, publishingPrinciples, review, reviews, sourceOrganization, text, thumbnailUrl, version, video
  - Article: articleBody, articleSection, wordCount
    - BlogPosting
    - NewsArticle: dateline, printColumn, printEdition, printPage, printSection
    - ScholarlyArticle
      - MedicalScholarlyArticle: citation, publicationType
  - Blog: blogPost, blogPosts
  - Book: bookEdition, bookFormat, illustrator, isbn, numberOfPages
  - Comment
  - Diet: dietFeatures, endorsers, expertConsiderations, overview, physiologicalBenefits, proprietaryName, risks
  - ExercisePlan: activityDuration, activityFrequency, additionalVariable, exerciseType, intensity, repetitions, restPeriods, workload
  - ItemList: itemListElement, itemListOrder
  - Map
  - MediaObject: associatedArticle, bitrate, contentSize, contentUrl, duration, embedUrl, encodesCreativeWork, encodingFormat, expires, height, interactionCount, offers, playerType, regionsAllowed, requiresSubscription, uploadDate, width
    - AudioObject: transcript
    - ImageObject: caption, exifData, representativeOfPage, thumbnail
    - MusicVideoObject
    - VideoObject: caption, productionCompany, thumbnail, transcript, videoFrameSize, videoQuality
  - Movie: actor, actors, director, duration, musicBy, producer, productionCompany, trailer
  - MusicPlaylist: numTracks, track, tracks
    - MusicAlbum: byArtist
  - MusicRecording: byArtist, duration, inAlbum, inPlaylist
  - Painting
  - Photograph
  - Recipe: cookingMethod, cookTime, ingredients, nutrition, prepTime, recipeCategory, recipeCuisine, recipeInstructions, recipeYield

Riferimenti

- Ricerca semantica
  - http://en.wikipedia.org/wiki/Semantic_search
  - http://searchenginewatch.com/
- Google Knowledge Graph
  - http://www.google.com/intl/it/insidesearch/features/search/knowledge.html
  - http://googleblog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html
  - http://support.google.com/websearch/bin/answer.py?hl=it&answer=2620861
  - http://searchenginewatch.com/search?q=%22knowledge+graph%22
- Ricerca semantica e SEO
  - http://searchenginewatch.com/article/2214849/Googles-Knowledge-Graph-Implications-for-Search-SEO
  - http://searchenginewatch.com/article/2234448/Keywords-Are-Dead-Long-Live-User-Intent