



# Text Mining

Information Extraction &  
Named Entity Recognition

# | Outline

1. Information Extraction
2. Named Entity Recognition



# | #1 Information Extraction

- Information extraction (IE)
  - adalah pencarian otomatis pada informasi yang terstruktur seperti entitas, hubungan antar entitas, dan atribut yang menggambarkan entitas dari sumber yang tidak terstruktur
  - Mengumpulkan informasi dari banyak kumpulan teks
  - Bagian dari *natural language processing*
  - Menghasilkan sebuah representasi informasi yang relevan dan terstruktur:
    - Relasi (seperti dalam database)
    - Sebuah Basis Pengetahuan



# | #1 Information Extraction

- Information extraction (IE) systems
  - Tujuan:
    1. Membuat informasi menjadi lebih **terorganisir** dengan baik sehingga **berguna** untuk manusia
    2. Informasi ditampilkan dalam sebuah format yang tepat secara **semantic** sehingga memungkinkan dilakukan **inferensi** pada tahap selanjutnya oleh algoritma komputer



# | #1 Information Extraction

- IE systems mengekstrak informasi yang terstruktur, jelas dan faktual dari teks yang tidak terstruktur
- Singkatnya : Siapa melakukan apa ke siapa, kapan dan di mana?

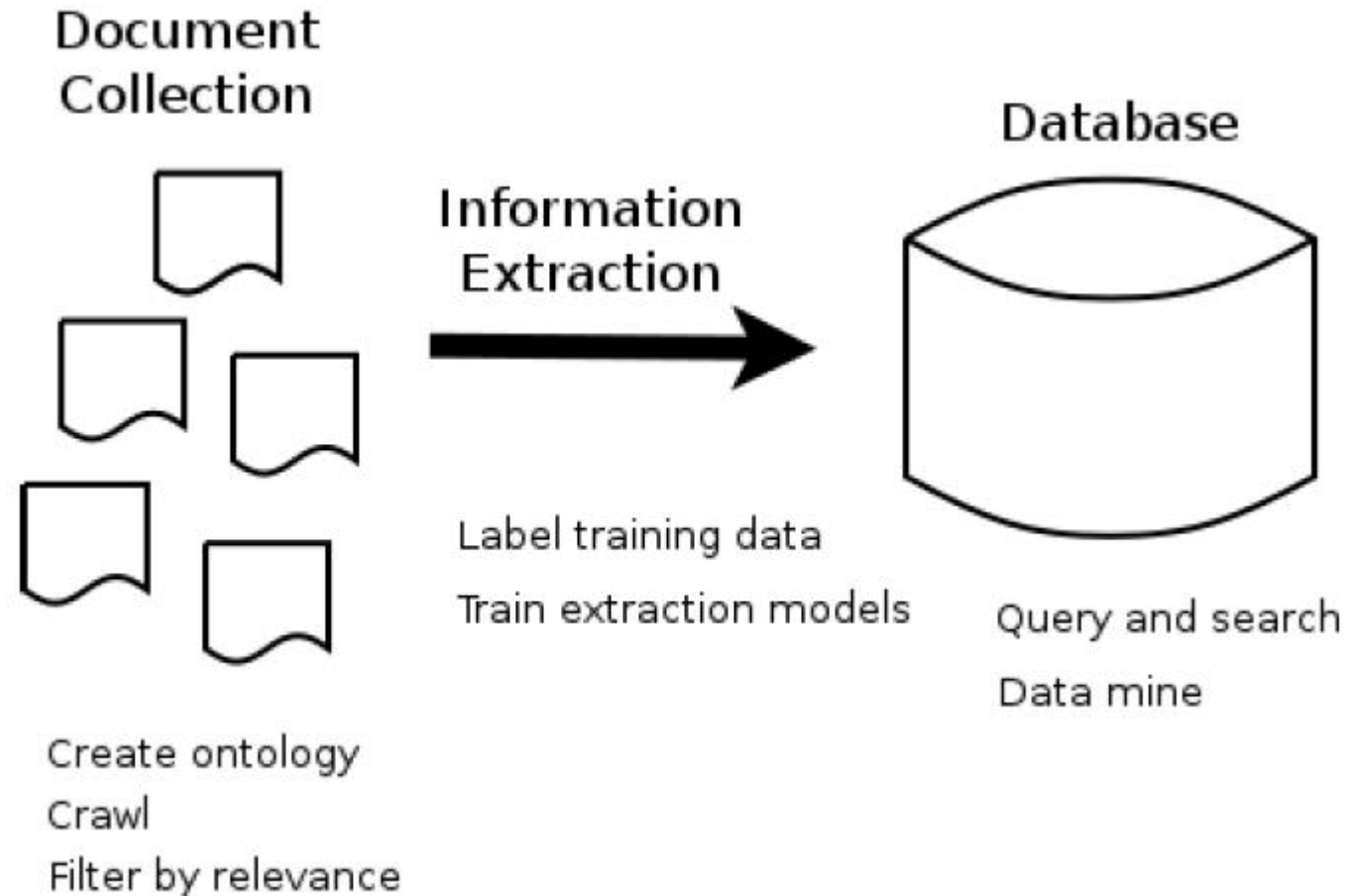


# | #1 Information Extraction

- Tugas yang sering dilakukan pada IE:
  - Mengidentifikasi entitas nama
  - Mengidentifikasi relasi antar entitas
  - Memasukkan ke database dengan format tertentu
- Tugas yang lain:
  - Mengekstraksi Kejadian
  - Membuat Template Otomatis dari informasi dalam teks
  - dsb
- Manfaat dan Aplikasi: natural language understanding, question-answering, summarization, dsb.



# | #1 Information Extraction



# | #1 Information Extraction

- Contoh,
  - Mengumpulkan data pemasukan, laba, pimpinan, kantor dsb dari laporan perusahaan
    - Kantor Pusat BHP Billiton Limited, dan kantor utama BHP Billiton Group, terletak di Melbourne, Australia.
    - Kantor Pusat(“BHP Biliton Limited”, “Melbourne, Australia”)
  - Mempelajari interaksi produk obat dari literatur penelitian medis





# | #1 Information Extraction

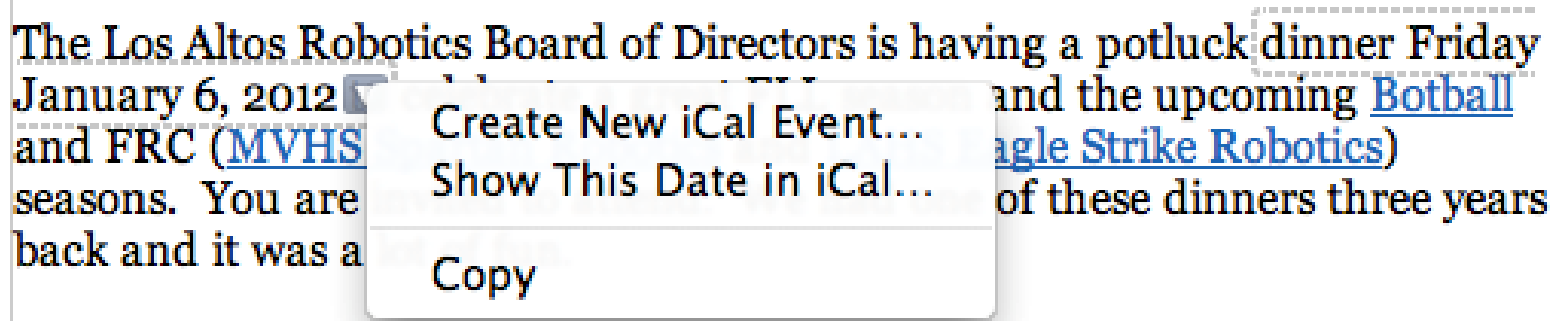
Target relation: `HeadquarteredIn(<company>, <city>)`

Headquartered in Portland, OR, AFMS was founded by Mike Erickson in 1992.	→ HeadquarteredIn(AFMS, Portland)
The files, kept at Boy Scout headquarters in Irving, Texas, consist of....	→ HeadquarteredIn(Boy Scout, Irving)
Yes to both, at the Nike HQ in Beaverton.	→ HeadquarteredIn(Nike, Beaverton)
AAPH is moving our headquarters and member fulfillment services to Portland, Oregon!	→ HeadquarteredIn(AAPH, Portland)

# | #1 Information Extraction

## Low Level IE

- Contoh Low Level IE, deteksi event pada Apple mail atau Gmail

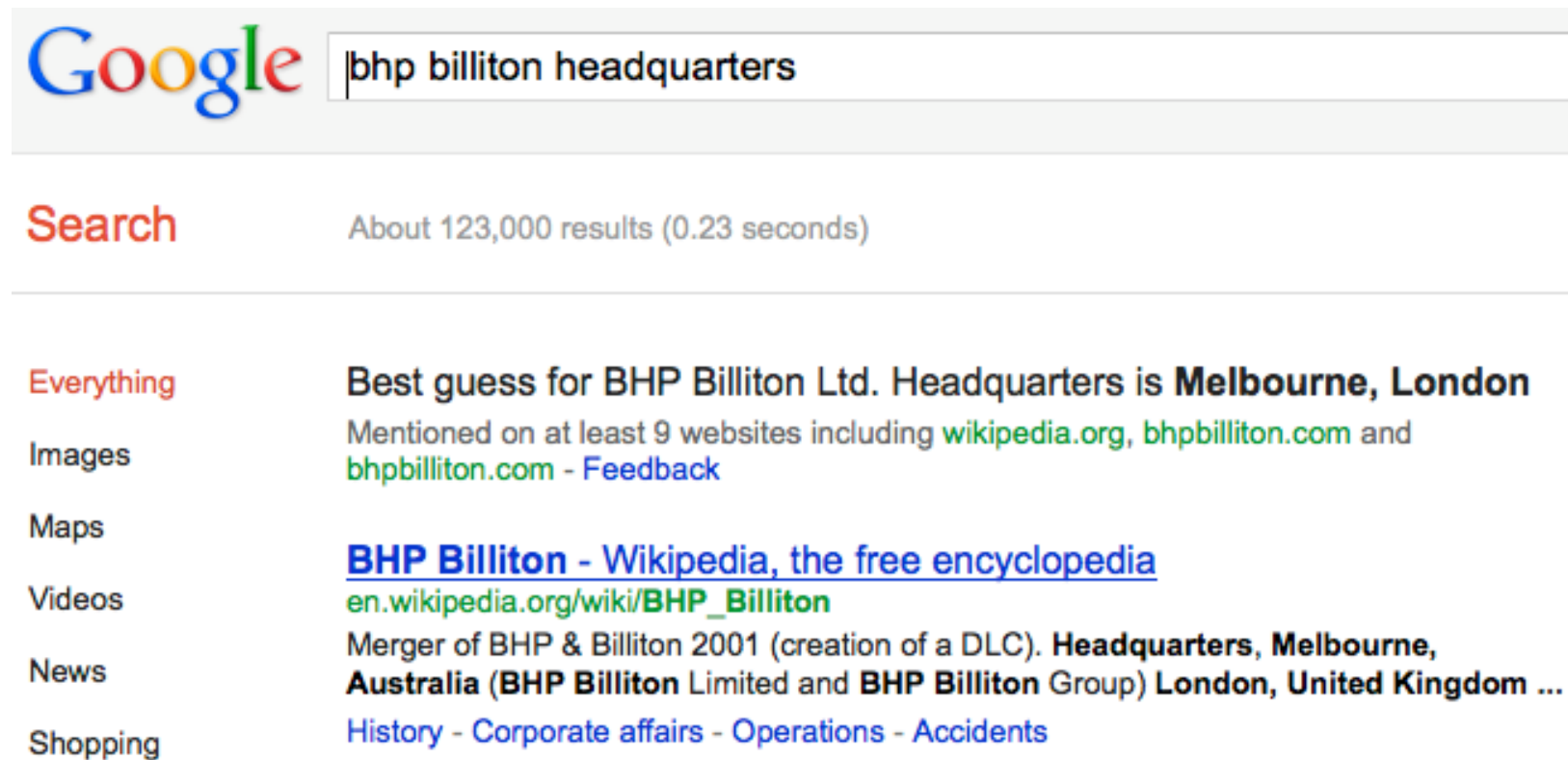


- Biasanya berbasis pada regular expressions and name lists



# | #1 Information Extraction

## Low Level IE



The screenshot shows a Google search interface. The search bar contains the text "bhp billiton headquarters". Below the search bar, the word "Search" is displayed in red, followed by the text "About 123,000 results (0.23 seconds)". On the left side, there is a vertical list of search filters: "Everything", "Images", "Maps", "Videos", "News", and "Shopping". The "Everything" filter is selected. The main search results area displays the following information:

- Best guess for BHP Billiton Ltd. Headquarters is Melbourne, London**
- Mentioned on at least 9 websites including [wikipedia.org](#), [bhpbilliton.com](#) and [bhpbilliton.com](#) - [Feedback](#)
- [BHP Billiton - Wikipedia, the free encyclopedia](#)
- [en.wikipedia.org/wiki/BHP\\_Billiton](#)
- Merger of BHP & Billiton 2001 (creation of a DLC). **Headquarters, Melbourne, Australia (BHP Billiton Limited and BHP Billiton Group) London, United Kingdom ...**
- [History](#) - [Corporate affairs](#) - [Operations](#) - [Accidents](#)



# | #1 Information Extraction

Why is IE hard on the web?

Established Phoenix 1994  
**NetStoreUSA.com**

Luckys Collectors Guide To 20th Century Yo-Yos:  
History And Values

EMAIL THIS PAGE TO A FRIEND

- English Books
- German Books
- Spanish Books
- Sheet Music
- Musical Supplies
- US/World Maps
- Sports Memorabilia
- Videos/Posters

[English Books](#) > [Antiques/Collectibles](#) > [Toys](#) > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

<< PREVIOUS TITLE | NEXT TITLE >> << NEW RELEASES >>

**Luckys Collectors Guide To 20th Century Yo-Yos: History And Values**  
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates  
Paperback  
Published: October 1999  
Lucky J's Swim & Surf  
ISBN: 0966761200

CHECK THE AVAILABILITY OF THIS PRODUCT

+ ADD TO CART

→ VIEW CART CHECKOUT

**PRODUCT CODE: 0966761200**

- USA/Canada: US\$ 43.40
- Australia/NZ: A\$ 124.50
- Other Countries: US\$ 80.90

[convert to your currency](#)

ADVANCED SEARCH >>

Home  
To Order  
Privacy  
Affiliates Coop  
Education  
Government  
About us  
Contact

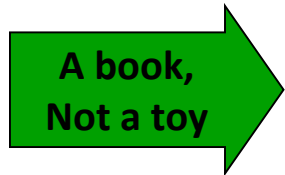
Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it

Kalau manusia pasti tahu kalau ini adalah buku, tapi bisa jadi komputer mengenalinya sebagai Toys, karena berada di bawah menu Toys



# | #1 Information Extraction

Why is IE hard on the web?



NetStoreUSA.com

Established Phoenix 1994

Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

English Books  
German Books  
Spanish Books

Sheet Music  
Musical Supplies

US/World Maps  
Sports Memorabilia  
Videos/Posters

English Books > Antiques/Collectibles > Toys > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

<< PREVIOUS TITLE | NEXT TITLE >> << NEW RELEASES >>

**Luckys Collectors Guide To 20th Century Yo-Yos: History And Values**  
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates  
Paperback  
Published: October 1999  
Lucky J's Swim & Surf  
ISBN: 0966761200

PRODUCT CODE: 0966761200

- USA/Canada: US\$ 43.40
- Australia/NZ: A\$ 124.50
- Other Countries: US\$ 80.90

[convert to your currency](#)

CHECK THE AVAILABILITY OF THIS PRODUCT

ADD TO CART

VIEW CART CHECKOUT

Home  
To Order  
Privacy  
Affiliates Coop  
Education  
Government  
About us  
Contact

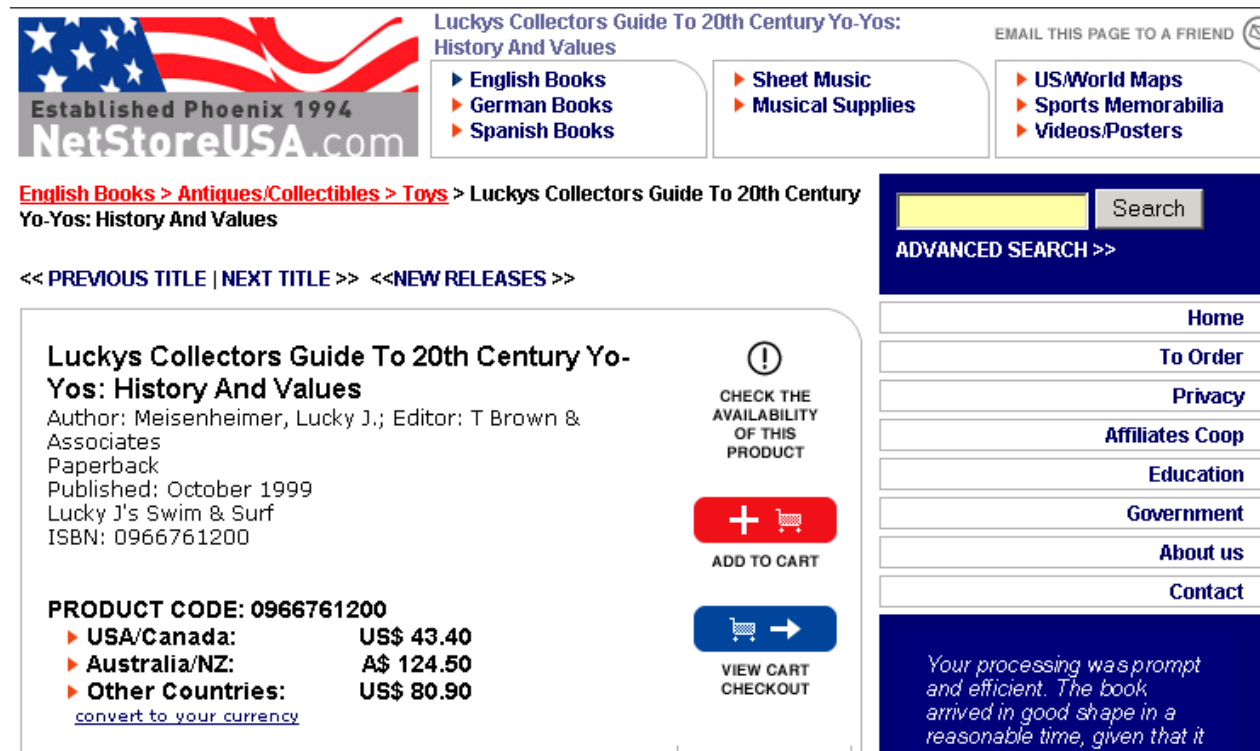
Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it

Beberapa produk ditempatkan tidak *directly* sesuai dengan kategorinya



# | #1 Information Extraction

Why is IE hard on the web?



Established Phoenix 1994  
**NetStoreUSA.com**

Luckys Collectors Guide To 20th Century Yo-Yos:  
History And Values

EMAIL THIS PAGE TO A FRIEND

- English Books
- German Books
- Spanish Books
- Sheet Music
- Musical Supplies
- US/World Maps
- Sports Memorabilia
- Videos/Posters

[English Books](#) > [Antiques/Collectibles](#) > [Toys](#) > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

<< PREVIOUS TITLE | NEXT TITLE >> <<NEW RELEASES >>

**Luckys Collectors Guide To 20th Century Yo-Yos: History And Values**  
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates  
Paperback  
Published: October 1999  
Lucky J's Swim & Surf  
ISBN: 0966761200

CHECK THE AVAILABILITY OF THIS PRODUCT

ADD TO CART

VIEW CART CHECKOUT

PRODUCT CODE: 0966761200

- USA/Canada: US\$ 43.40
- Australia/NZ: A\$ 124.50
- Other Countries: US\$ 80.90

[convert to your currency](#)

Home

To Order

Privacy

Affiliates Coop

Education

Government

About us

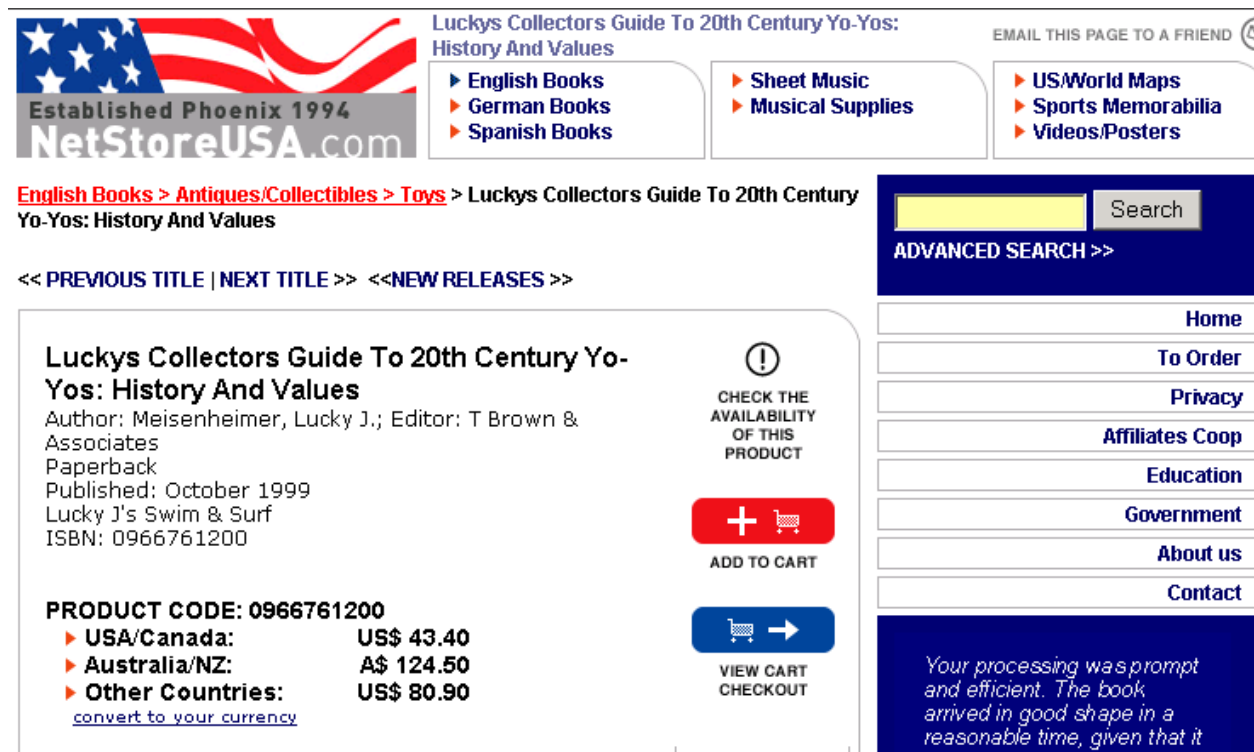
Contact

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it

Judul tidak dituliskan secara eksplisit sebagai judul, hanya dibold dan memakai besar huruf

# | #1 Information Extraction

Why is IE hard on the web?



Established Phoenix 1994  
**NetStoreUSA.com**

Luckys Collectors Guide To 20th Century Yo-Yos:  
History And Values

EMAIL THIS PAGE TO A FRIEND

- English Books
- German Books
- Spanish Books
- Sheet Music
- Musical Supplies
- US/World Maps
- Sports Memorabilia
- Videos/Posters

English Books > Antiques/Collectibles > Toys > Luckys Collectors Guide To 20th Century Yo-Yos: History And Values

<< PREVIOUS TITLE | NEXT TITLE >> <<NEW RELEASES >>

**Luckys Collectors Guide To 20th Century Yo-Yos: History And Values**  
Author: Meisenheimer, Lucky J.; Editor: T Brown & Associates  
Paperback  
Published: October 1999  
Lucky J's Swim & Surf  
ISBN: 0966761200

! CHECK THE AVAILABILITY OF THIS PRODUCT

+ ADD TO CART

VIEW CART CHECKOUT

PRODUCT CODE: 0966761200

- USA/Canada: US\$ 43.40
- Australia/NZ: A\$ 124.50
- Other Countries: US\$ 80.90

[convert to your currency](#)

Home  
To Order  
Privacy  
Affiliates Coop  
Education  
Government  
About us  
Contact

Your processing was prompt and efficient. The book arrived in good shape in a reasonable time, given that it

Need this price

Untuk harga – ada banyak harga yang ditampilkan, harus pilih yang mana?



# | #1 Information Extraction

How is IE useful?

Background:

- Plain text advertisements
- Mengestrak informasi dari iklan di koran lalu diformat menjadi seperti di samping.
- Lalu bisa ditulis lagi dengan menggunakan format lain tapi informasinya tetap sama



Vente Villa 4 pièces Nice (06000)  
Réf. 12390: Sur les Hauteurs de Nice. Superbe villa moderne (190m2), 2 chambres et 1 suite parentale, 3 salles de bain. Très grand salon/salle à manger, cuisine américaine équipée. Prestations de haut standing. Vue panoramique sur la mer. Cette villa a été construite en 2005. 1 270 000 euros. Si vous êtes intéressés, contactez vite Mimi LASOURIS 06.43.43.43. 43

## REAL ESTATE TEMPLATE

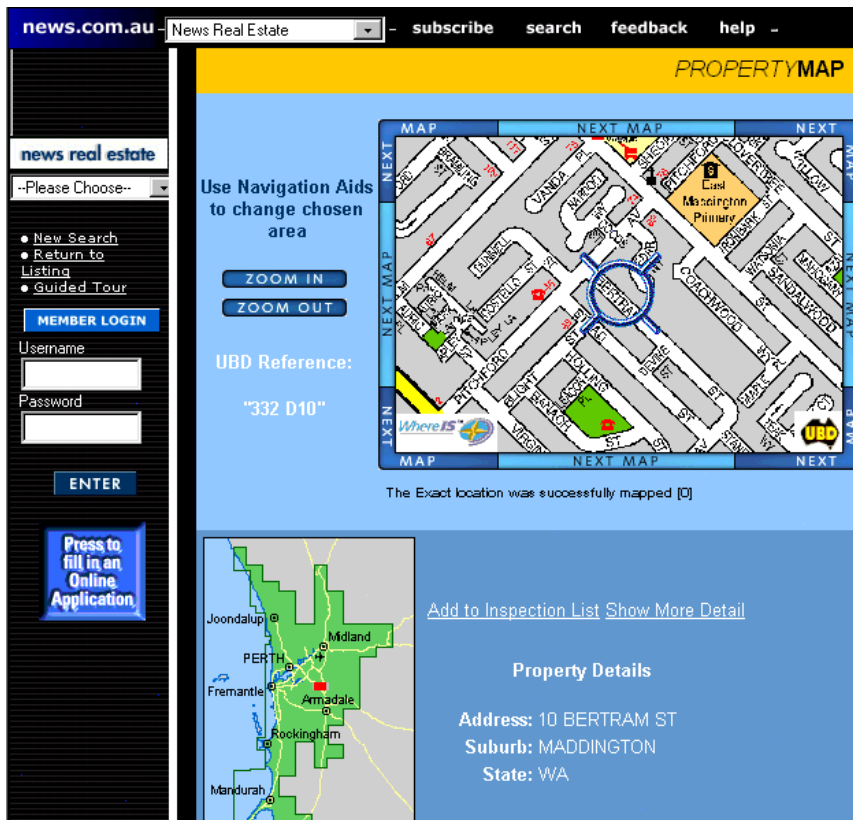
Reference: 12390  
Prize: 1 270 000  
Surface: 190 m2  
Year Built: 2005  
Rooms: 4  
Owner: Mimi LASOURIS  
Telephone: 06.43.43.43. 43





# | #1 Information Extraction

## How is IE useful?



Mengambil informasi alamat rumah di web secara otomatis



# | #1 Information Extraction

## IE vs Information Retrieval

- **Information Retrieval**

- User Query -> Teks/dokumen yang relevan
- *Pendekatan*: keyword matching
- *Query generality* : full

- **Information Extraction**

- Analisis Linguistic yang ditargetkan pada informasi yang relevan
- *User Query -> Informasi yang relevan*
- *Pendekatan*: Analisis linguistic
- *Query generality*: Dibatasi pada target informasi



## | #2 Named Entity Recognition (NER)

- Named Entity Recognition (NER) adalah salah satu Subtask yang sangat penting dalam IE : **Menemukan** dan **Mengklasifikasi** nama-nama Entitas dalam teks
- Nama Entitas apa saja? Tergantung pada Aplikasinya.
  - People, places, organizations, times, amounts, etc.
  - Names of genes and proteins (Settles 05)
  - Names of college courses (McCallum 05)



# | #2 Named Entity Recognition (NER)

- NER untuk menemukan, contoh :
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.



# | #2 Named Entity Recognition (NER)

- NER untuk menemukan dan mengklasifikasi, contoh :
  - The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person  
Date  
Location  
Organization



# | #2 Named Entity Recognition (NER)

- Manfaat:
  - Melakukan Indeksi Entitas dsb.
  - Sentiment bisa disematkan pada perusahaan atau produk
  - NER juga bermanfaat dalam banyak aplikasi NLP (*Natural Language Processing*) seperti *question-answering*, rangkuman dan sistem dialog. NER juga berkaitan task information extraction lainnya seperti dengan relation detection, event detection dan temporal analysis.



# | #2 Named Entity Recognition (NER)

Ambiguitas pada NER :

Ada dua jenis ambiguitas yang dapat ditemui NER.

- Pertama, **kata yang sama dapat berarti dua entitas yang berbeda**. Misalnya kata Soekarno dapat berarti presiden pertama Indonesia, atau nama belakang seorang seniman (Enrico Soekarno), keduanya entitas berbeda walaupun tipenya sama (orang/person).
- Jenis ambiguitas kedua adalah **nama yang sama tapi tipe berbeda**. Contohnya adalah Bung Karno sebagai stadion dengan Bung Karno sebagai orang. Ambiguitas umumnya ditangani dengan menggunakan kamus.

**Solusi:** Deteksi named entity dapat dilakukan dengan melihat pola kata disekitarnya.

Misalnya frasa yang didahului oleh kalimat “pergi ke ... “ atau “datang dari ... “ kemungkinan besar adalah named entity bertipe lokasi.

machine learning dapat digunakan untuk mempelajari pola secara otomatis dan melakukan prediksi label kategori.



# | #2 Named Entity Recognition (NER)

- **Tiga pendekatan Standart untuk NER (dan IE)**
  1. Hand-written regular expressions
  2. Using classifiers
    - Generative: Naïve Bayes
    - Discriminative: Maxent models
  3. Sequence models
    - HMMs
    - CMMs/MEMMs
    - CRFs





# | #2 Named Entity Recognition (NER)

## NER vs POS Tag

- POS Tag mengidentifikasi kelompok tata bahasa mana kata itu berasal, jadi apakah itu merupakan NOUN, ADJECTIVE, VERB, ADVERBS dll berdasarkan pada konteks. Mencari hubungan di dalam kalimat dan memberi setiap kata dalam kalimat tag yang sesuai. Sedangkan NER mencoba mencari tahu apakah suatu kata adalah entitas bernama atau tidak. Entitas yang dinamai adalah orang, lokasi, organisasi, ekspresi waktu dll melalui klasifikasi.
- Setiap tag POS dilampirkan ke satu kata, sedangkan tag NER dapat dilampirkan ke beberapa kata. Jadi NER melibatkan tidak hanya mendeteksi jenis Entitas Bernama, tetapi juga kata batas. Deteksi NER adalah tugas "tingkat lebih tinggi" daripada penandaan POS, dan umumnya menggunakan POS sebagai fitur input.



# Thank you

