

# The surplus value of semantic annotation

ESAIR

Maarten Marx

Universiteit van Amsterdam

October 2010

## Message

- Semantic annotation is costly,
- but has possibly many unforeseen beneficial exploitations.
- E.g., DBLP, DBpedia.
- The surplus value of semantic annotations is proportional to the value of these “unexpected” applications.

# Outline

- What is semantic annotation and what not?
- Types of Semantic Annotation
- Types of Applications
- How to create Surplus Value?

# What is Semantic Annotation? (in this talk)

- Data added to documents.
- Each added piece of data is qualified by a **semantic** category.
- Typically as **attribute-value** pairs.
- **Examples**
  - `<document dc:date='2010-10-30'>`
  - `<ne type='PER' normalized='Jimi_Hendrix'>Jimmy  
Hendricks</ne>`
- **Non-examples**
  - Tags added by users to Flickr.
  - Anchor text of inlinking pages.

## Two Kinds of Semantic Annotation

1. Add metadata to the document.
2. Split document into structural elements and add metadata to these elements.

## Add metadata to document

- Examples:
  - Language identification
  - Extract mentioned named entities.
  - Add controlled vocabulary terms.
- Document remains the same.

## Structure the document and annotate that structure

- Examples:
  - POS and named entity tagging
  - Chapter detection
  - Reported speech annotation.
- Ideally the old document can be obtained from the new document by removing the added structure and metadata.
- May redefine which “documents” will be indexed.

## Example: From text to nested XML

- **Notes of meetings** have a rich nested structure:

```
<meeting>
  <topic topic='...' page='..' ...>
    <block lectern='Wilders' ...
      <speaker name='Wilders' anchor='..' ...>
        <p>
          Mevrouw de voorzitter. Om te
          beginnen mijn oprechte dank aan u persoonlijk omdat u
          op mijn verjaardag vandaag een debat over de islam
          heeft gepland. Een mooier cadeau had ik mij niet kunnen
          wensen!
        </p>
        ...
      </speaker>
      <speaker name='Dijsselbloem' ...
      <speaker name='Wilders' anchor='..' ...
      ...
    </block>
```

## SA in practice

- Semantic annotation can be **expensive**.
- Bag of words approach is easy, robust, scalable and **cheap**.
- Thus SA is typically applied when
  - ★ heavy competition
  - ★ “document” does not fit smoothly in ad-hoc search scenario
  - ★ “document” has extractable implicit structure
- **Verticals**
  - eBay
  - Google Books
  - LinkedIn

# Exploiting Semantic Annotation for IR

- Three clusters of applications:
  1. Improve **retrieval performance** and search result presentation.
  2. Improve user's **interaction with the document**.
  3. Improve user's **interaction with and understanding** of the **search space**.

## Example: Google Book Search

- High quality document understanding is needed.
- [Vincent 07] lists the following SA tasks and applications:
  - Document repairing (2)
  - Language identification (1)
  - Chapter detection (2)
  - Content linking (2) turn implicit link-structure into hypertext
  - Summarization (1,2)
  - Metadata extraction and cross-validation (1,3) extract book title, author, publisher, edition, publication year, etc.
  - Topic identification (1,3)
  - Book clustering and linking (1,3)

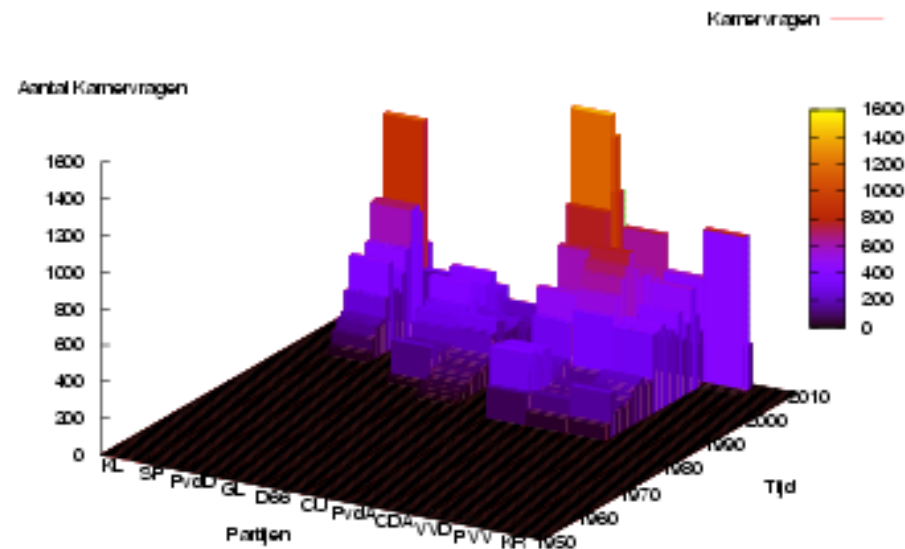
## Applications in 3 years of ESAIR

**Improve Ranking**  
**Ontology**  
**Faceted Search**  
**Query Expansion**  
**Focused Retrieval**

Novelty  
Summarization  
Result Grouping  
Classification  
Linked Data  
User Modelling

# Applications on top of a search-engine

- From one-dimensional ranked result list (10 blue lines)
- To high dimensional datacube



## Beyond retrieval

- From retrieval to exploration/browsing.
- From finding individual facts to understanding the whole.

## Mashup's of semantically annotated data

<http://www.congressspeaks.com/>

## Mashup's of semantically annotated data

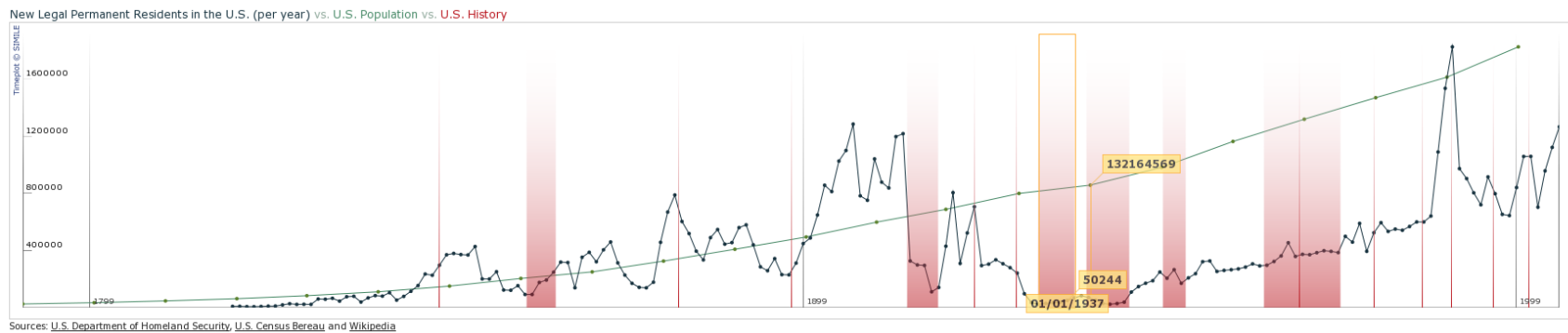
<http://www.congressspeaks.com/>

How can SA lead to such mashups?

- 
- Follow the Linked Open Data rules [Berners Lee 06].
- Use shared identifiers for semantic categories and instances.
- Publish in open formats.

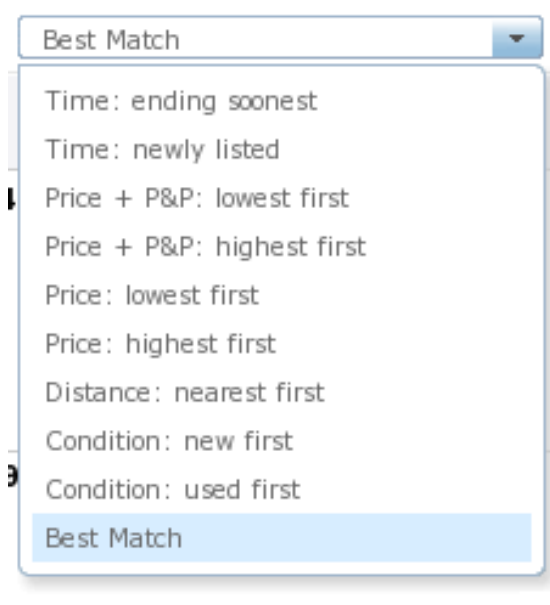
# Result grouping

- Temporal [Simile]
- Spatial [Google Maps]
- Controlled Vocabulary [ebay]



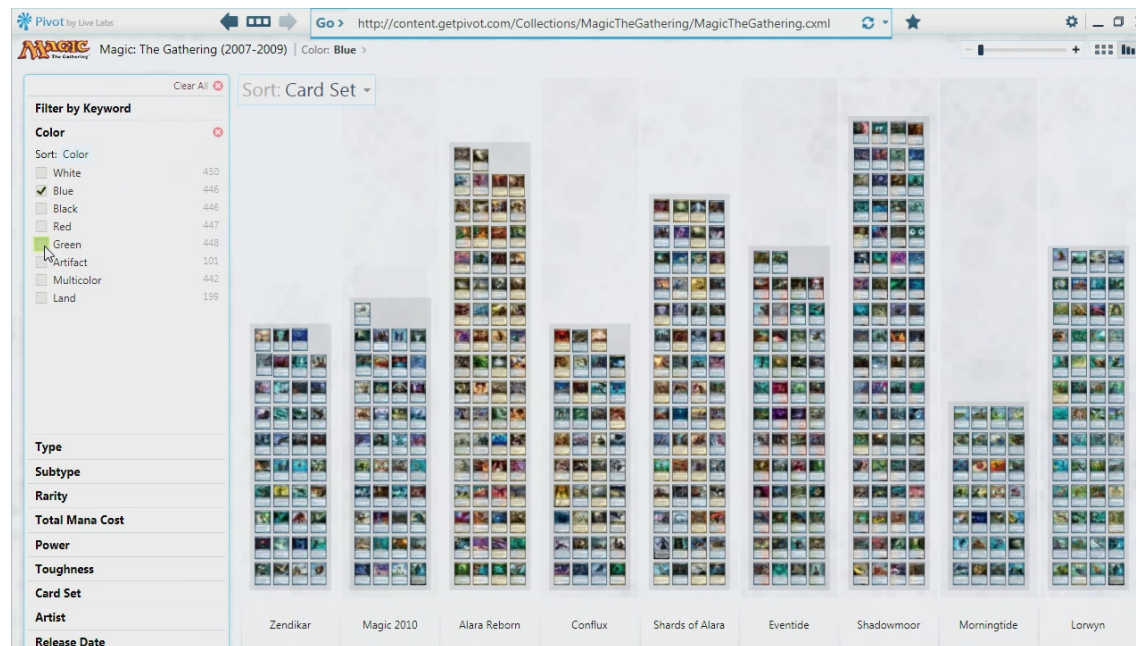
## Alternative sortings

- Personalized search (e.g distance between user and document)
- Economic ranking involving (expected) costs and utility



# Faceted Search ++

- drill down, slicing in high dimensional space
- Interactive visual histograms (Microsoft Pivot)



## Content and Structure Queries

- user controls the notion of “document”
- queries embedding multiple keyword subqueries



## Conclusions

- Semantic annotation is costly,
- but has possibly many unforeseen beneficial exploitations.
- E.g., DBLP, DBpedia.
- The surplus value of semantic annotations is proportional to the value of these “unexpected” applications.