



NLP-based Indexing

Dr. Elizabeth D. Liddy

**Center for Natural Language Processing
School of Information Studies
Syracuse University**



A Continuum from Human to Statistical Indexing

- **Manual**
 - Term selection from content
 - Controlled vocabularies
- **Mixed Initiative**
 - Machine-aided / Human-assisted
 - Machine Learning
- **Automatic**
 - Statistical indexing
 - Natural Language Processing indexing



Central Problem of IR

- **How to represent documents for retrieval (Blair, 1990)**
- **The quality of the representation of documents determines:**
 - the ‘richness’ of the indexing
 - the ‘quality’ of access to relevant information
 - the ‘value-add’ analytics the system can accomplish for users



NLP-based Indexing

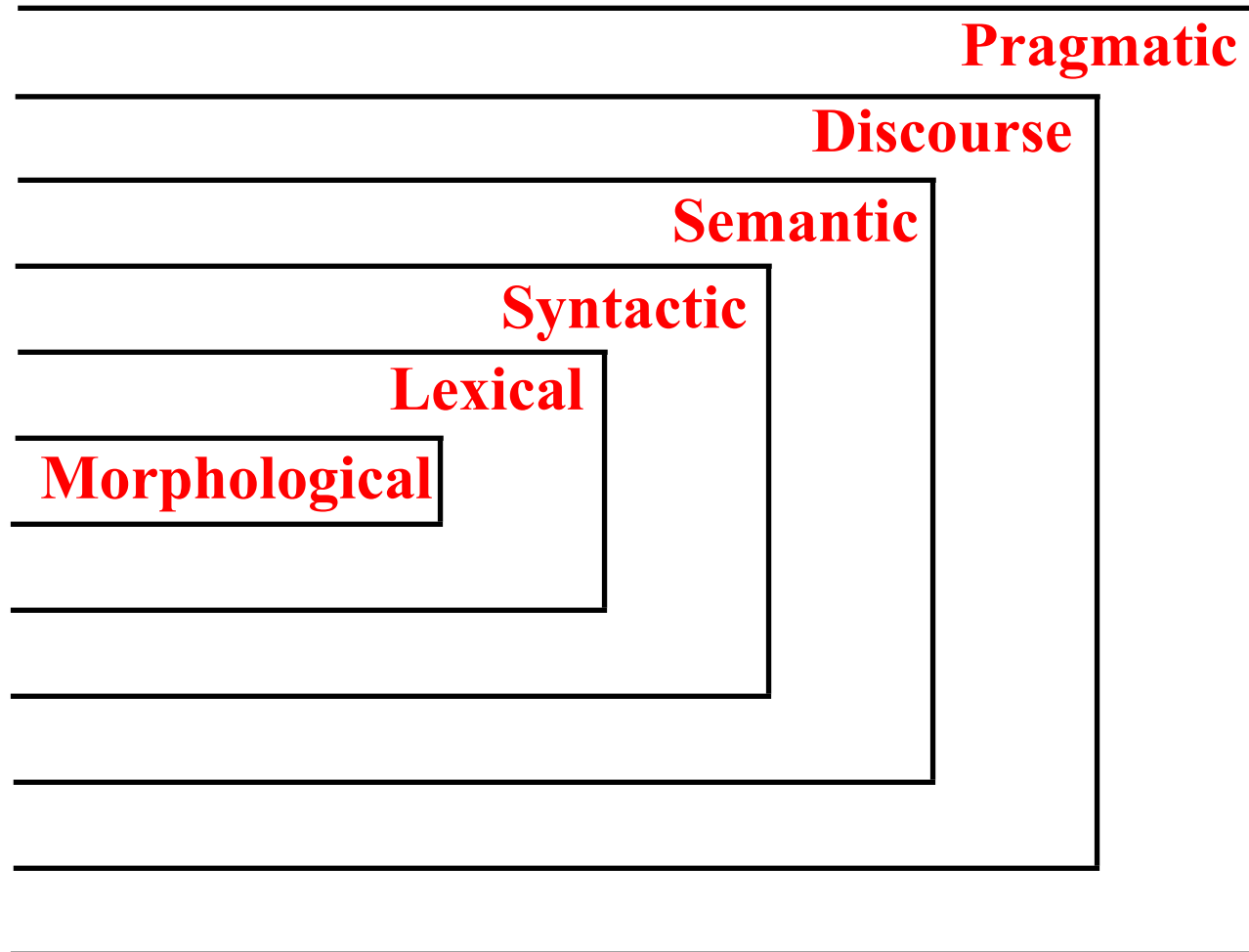
- the computational process of **identifying**, **selecting**, and **extracting** useful information from massive volumes of textual data:
 - for potential review by indexers
 - OR -
 - automatic production of a searchable representation of content
 - using Natural Language Processing



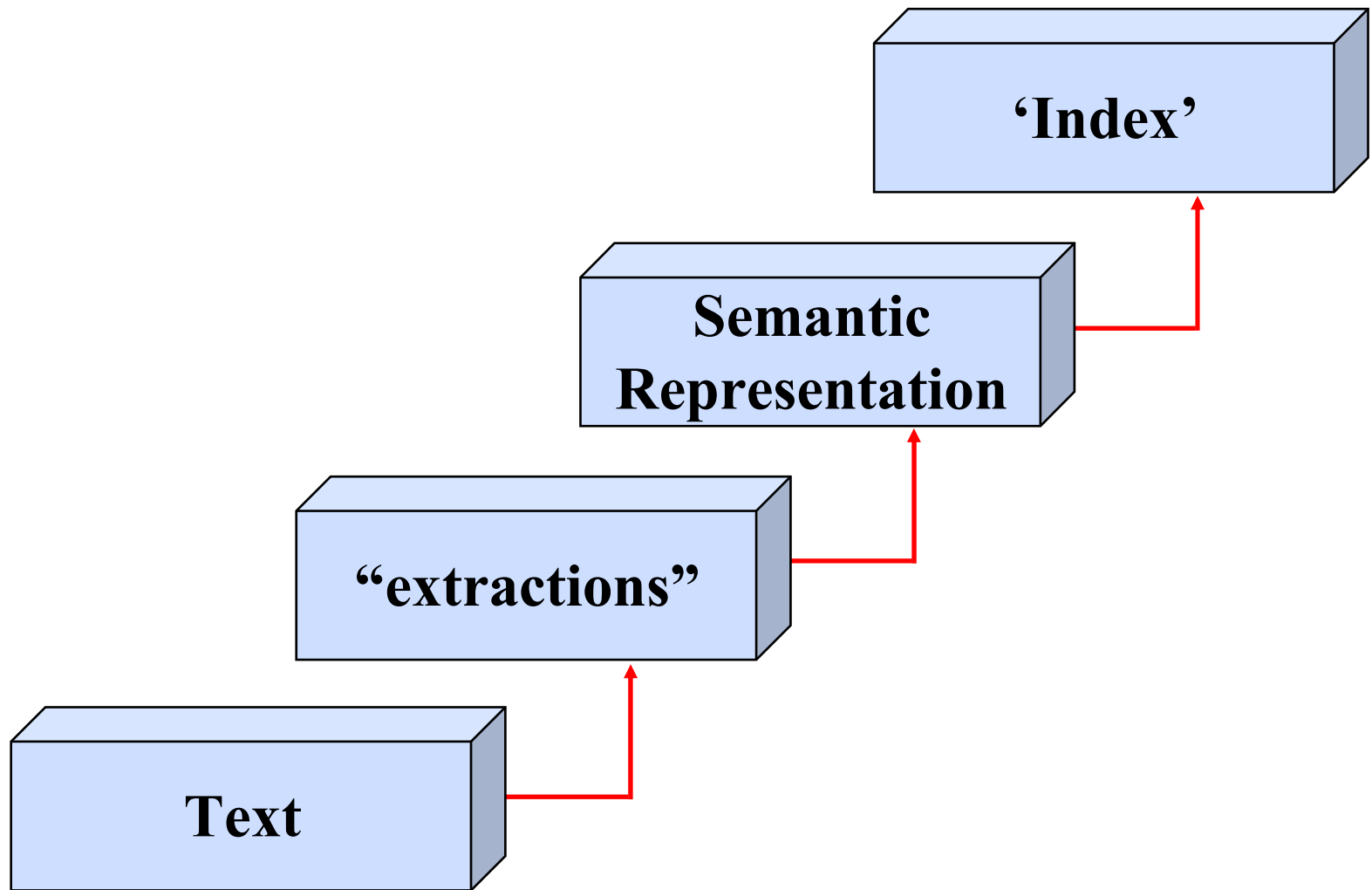
Natural Language Processing

- **a range of computational techniques**
- **for analyzing and representing naturally occurring texts**
- **at one or more levels of linguistic analysis**
- **for the purpose of achieving human-like language processing**
- **for a range of tasks or applications**

Levels of Language Understanding



NLP for Indexing



Tagging & Extraction Process (cont'd)

Westco Bancorp has been merged into MAF Bancorp, and First Federal Savings and Loan Association of Westchester, a wholly-owned subsidiary of Westco Bancorp

Tagging & Extraction Process (cont'd)

Westco Bancorp has been merged into MAF Bancorp, and First Federal Savings and Loan Association of Westchester, a wholly-owned subsidiary of Westco Bancorp

Westco|NP Bancorp|NP has|VBZ been|VBN merged|VBN into|IN MAF|NP Bancorp|NP ,|, and|CC First|NP Federal|NP Savings|NP and|CC Loan|NP Association|NP of|IN Westchester|NP ,|, a|DT wholly-owned|VBN subsidiary|NN of|IN Westco|NP Bancorp|NP.....

Tagging & Extraction Process (cont'd)

Westco Bancorp has been merged into MAF Bancorp, and First Federal Savings and Loan Association of Westchester, a wholly-owned subsidiary of Westco Bancorp

Westco|NP Bancorp|NP has|VBZ been|VBN merged|VBN into|IN MAF|NP Bancorp|NP ,|, and|CC First|NP Federal|NP Savings|NP and|CC Loan|NP Association|NP of|IN Westchester|NP ,|, a|DT wholly-owned|VBN subsidiary|NN of|IN Westco|NP Bancorp|NP

<NP> Westco_Bancorp </NP> has|VBZ been|VBN merged|VBN into|IN <NP> MAF_Bancorp </NP> ,|, and|CC <NP> First_Federal_Savings_and_Loan_Association_of_Westchester </NP> ,|, a|DT <NP> wholly-owned_subsidary </NP> of|IN <NP> Westco_Bancorp </NP>

Tagging & Extraction Process (cont'd)

Westco Bancorp <type=company > has|VBZ been|VBN merged|VBN into|IN MAF Bancorp <type=company> ,|, and|CC First Federal Savings and Loan Association of Westchester <type=company> ,|, a|DT <NP> wholly-owned subsidiary </NP> of|IN Westco Bancorp <type=company>

Tagging & Extraction Process (cont'd)

Westco Bancorp <type=company > has|VBZ been|VBN merged|VBN into|IN MAF Bancorp <type=company> ,|, and|CC First_Federal_Savings_and_Loan_Association_of_Westchester <type=company> ,|, a|DT <NP> wholly-owned_subsidary </NP> of|IN Westco_Bancorp <type=company>

generic relation: isa

entity1: First Federal Savings & Loan Assoc. of Westchester (company)

entity2: subsidiary

generic relation: characteristic

entity1: subsidiary

entity2: wholly-owned

generic relation: possess

entity1: Westco Bancorp Inc (company)

entity2: subsidiary

Event Extraction Process

<Westco Bancorp <type=company> <EV;type=combine; id=1-1-1> has|VBZ
been|VBN merged|VBN into|IN </EV> <MAF Bancorp Inc <type=company>

\$1<KOS='groups'> \$2<VB=COMBINE> \$3<KOS='groups'>

→

event: merge

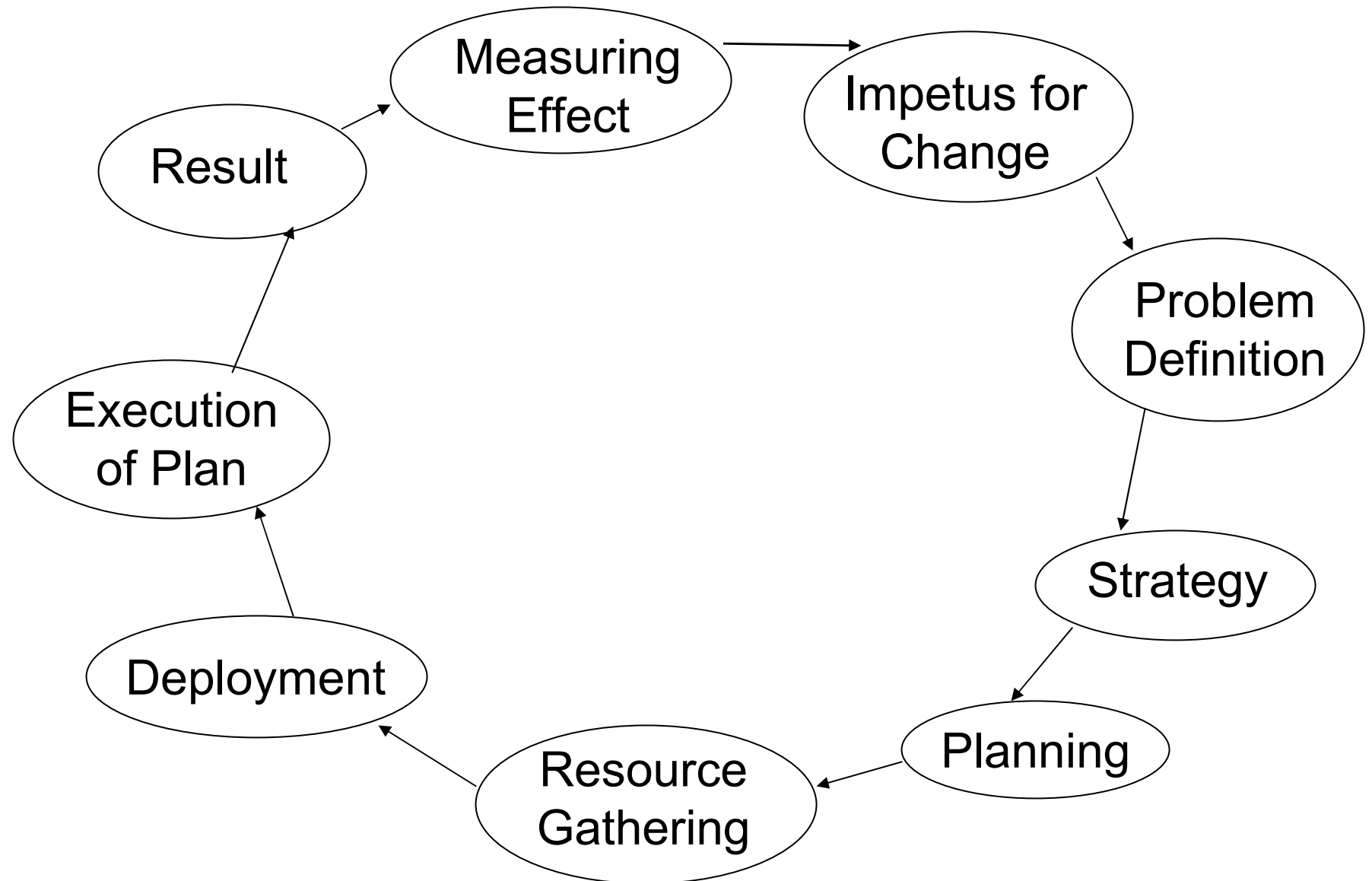
entity1: \$1

entity2: ?

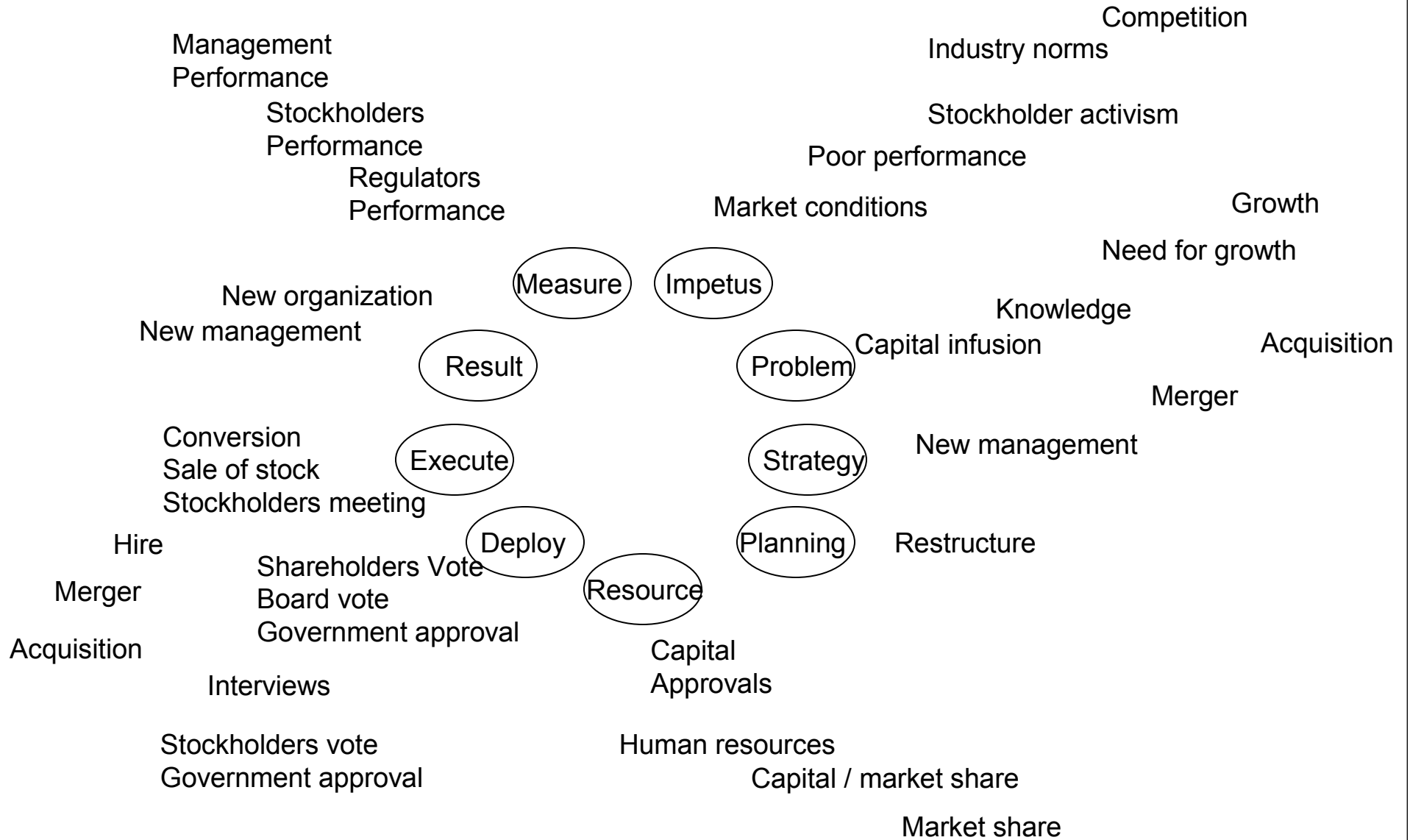
target: \$3

point-in-time: ?

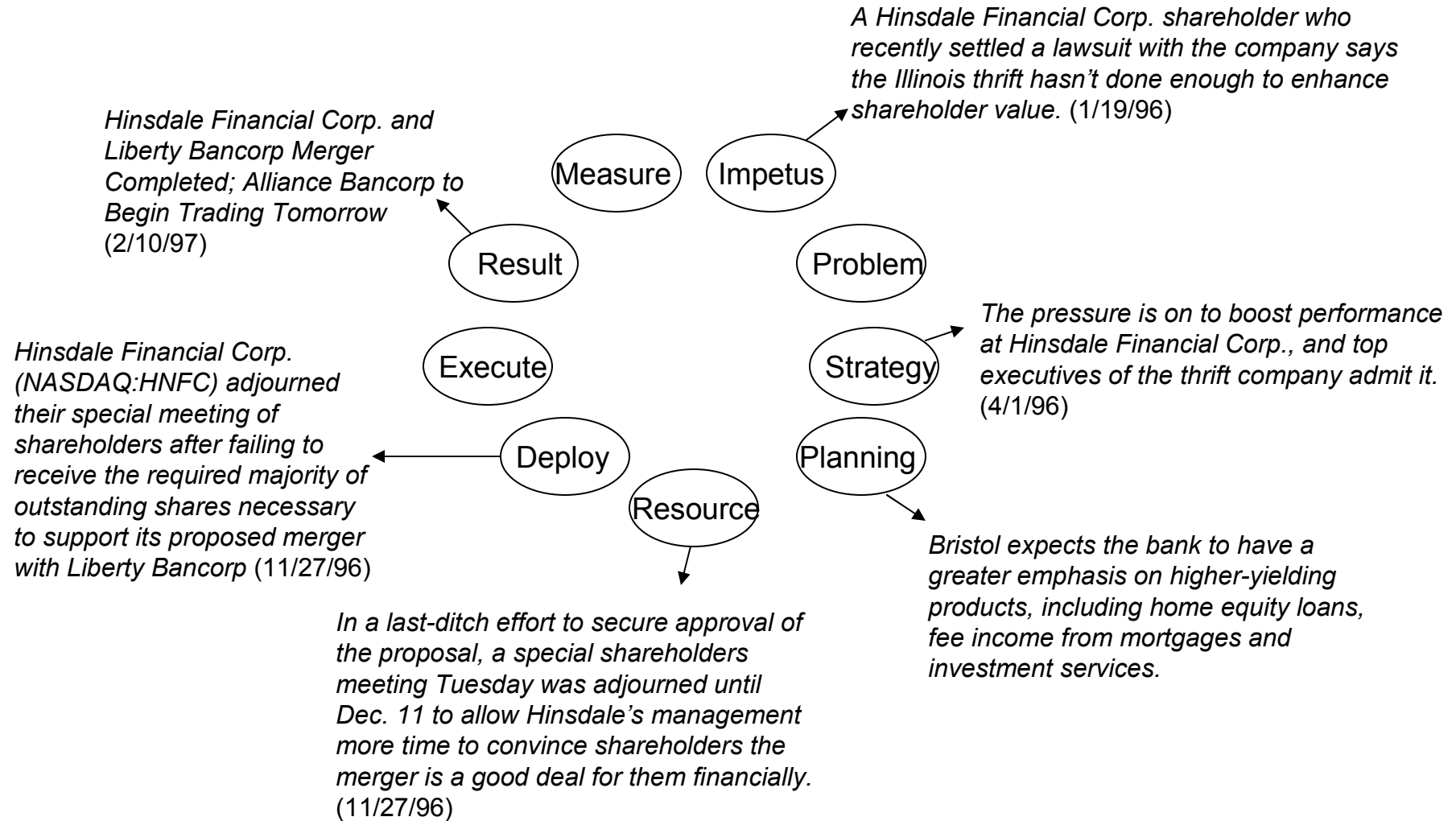
General Model



Banking Application of the Model



Instantiation of Banking Model



Event Extraction Process

<Westco Bancorp <type=company> <EV;type=combine; id=1-1-1> has|VBZ
been|VBN merged|VBN into|IN </EV> <MAF Bancorp Inc <type=company>

\$1<KOS='groups'> \$2<VB=COMBINE> \$3<KOS='groups'>

→

event: merge

entity1: \$1

entity2: ?

target: \$3

point-in-time: ?

\$1: 'Westco Bancorp Inc.'
category='company' -> 'groups'

\$3: 'MAF Bancorp Inc'
category='company' -> 'groups'

event: merge=1-1-1

entity1: Westco Bancorp Inc (company)

entity2: ?

target: MAF Bancorp Inc (company)

point-in-time: ?

Knowledge Organization Structure (KOS)

0. Life & Living Things (other than human)

1. People

1.1 Titles / Positions [*Director*]

1.1.1 Honorifics [*Mr.*]

1.1.2 Roles [*Dutch Auction Agent*]

1.1.3 Military Ranks [*Lt.Col.*]

1.2 Groups

1.2.1 Organizations [*Save the Children*]

1.2.1.1 Government Orgs [*DARPA*]

1.2.1.1 Courts [*Bankruptcy Court*]

1.2.1.2 Lawmaking groups [*Parliament*]

1.2.1.3 Military divisions [*Boys from Syracuse*]

1.2.1.2 Terrorist Groups [*Jihad*]

.....

1.2.1.20 Organizational subdivisions [*Director's Comm.*]

1.2.2.Companies [*ABC Bancorp*]

1.2.2.20 Company subdivisions [*Public Relations Dept.*]

2. Thought, Communication and Communication Channels

2.1 E-mail [*liddy@syr.edu*]

.....

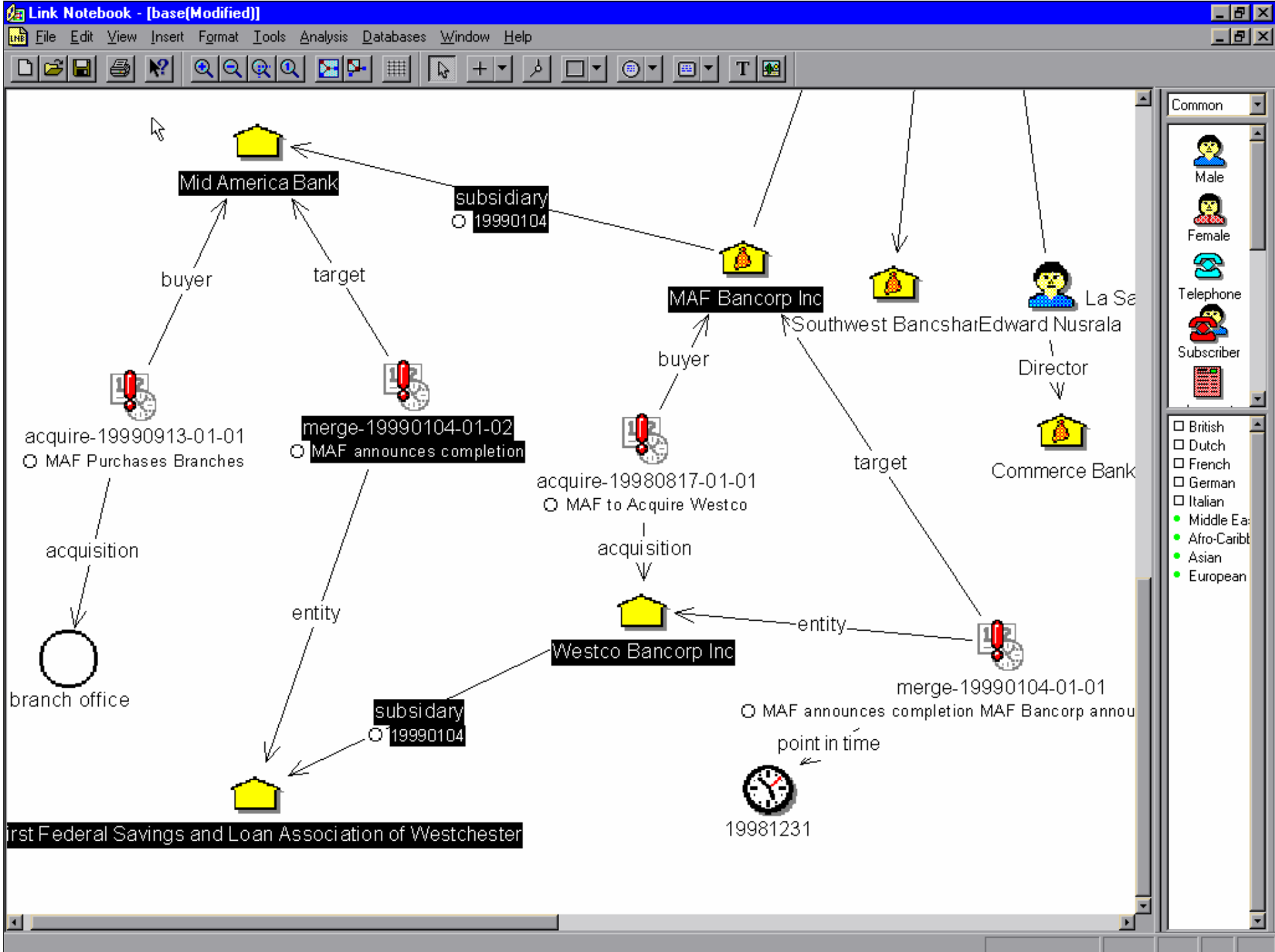
3. Buildings & Structures

4. Substances, Materials, Objects, and Equipment

.....

What NLP can extract & index:

- 1. Domain-independent Entities**
 - *Person, Country, Organization, Company*
- 2. Domain-independent Relations**
 - *Agent, object, location, point-in-time*
- 3. Domain-dependent Entities**
 - *Merger, subsidiary, perpetrator*
- 4. Domain-dependent Relations**
 - *Takeover, complain*
- 5. Model-specific Events**
 - *Acquisitions & Mergers*
 - *Terrorism, Smuggling*
 - *Public Health Interventions*





Knowledge Organization Structures

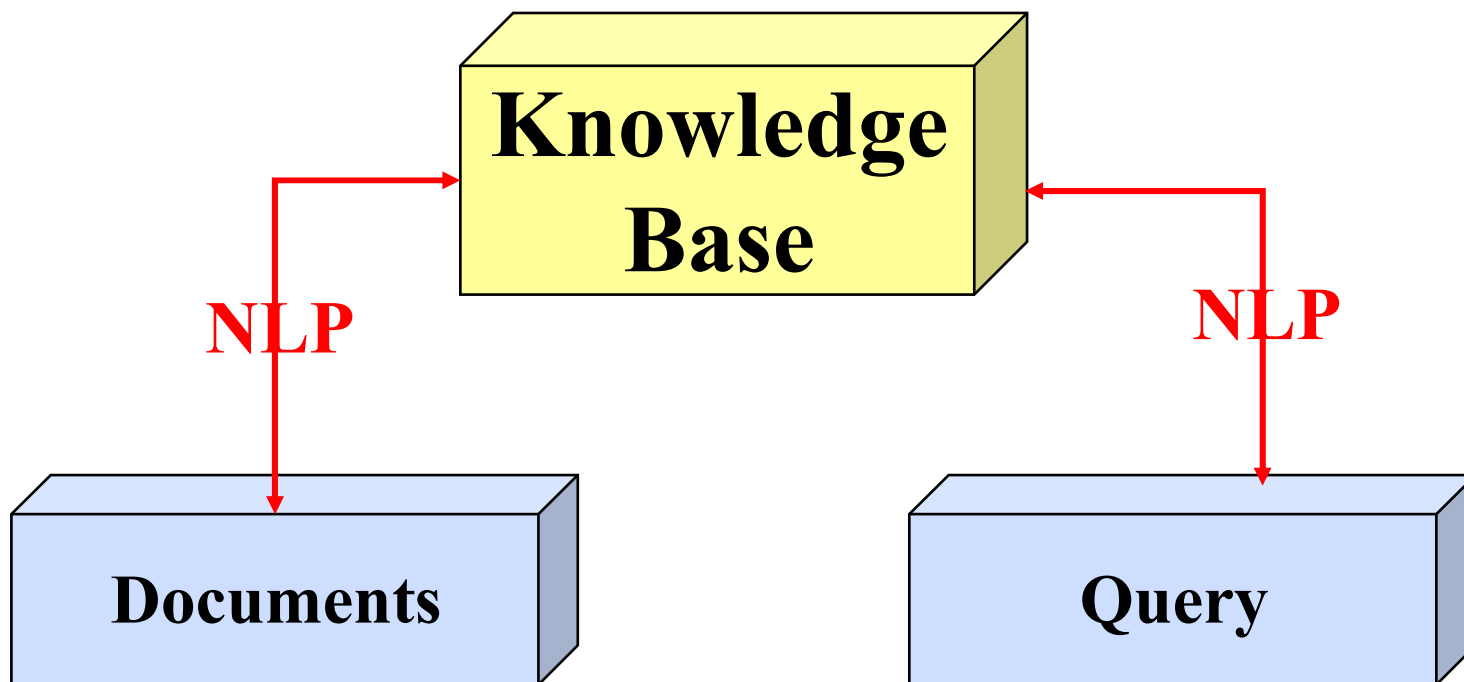
- **Key to knowledge-rich organizational tasks:**
 - **Information retrieval**
 - **Question Answering**
 - **Visualization**
 - **Automatic thesaurus construction**
 - **Information retrieval**
 - **Knowledge Base construction**
 - **Internal documents**
 - **Web Sites**
 - **I**



Methodology

- **Run NLP modules on free text**
- **Extract entities and ISA relations**
- **Based on training examples, system learns structure of knowledge in this domain**
- **Populate KOS using learned rules**
- **Domain-independent extraction can be specialized for domain of interest**

eQuery





eQuery Representation

What Iranian backed terrorist groups exist in the Middle East region?

What|**WP** Iranian|**JJ** backed|**VBD** terrorist|**NN**
groups|**NNS** exist|**VBP** in|**IN** the|**DT** Middle|**NP**
East|**NP** region|**NN** ?|?



eQuery Representation

What Iranian backed terrorist groups exist in the Middle East region?

What|WP Iranian|JJ backed|VBD terrorist|NN groups|NNS exist|VBP in|IN the|DT Middle|NP East|NP region|NN ?|?

What|WP <NP> Iranian </NP> backed|VBD terrorist|NN groups|NNS exist|VBP in|IN the|DT <NP> Middle_East </NP> region ?|?



eQuery Representation

**What|WP <NP cat=nationality id=1> Iranian </NP>
backed|VBD <CN> terrorist|NN groups|NNS
</CN> exist|VBP in|IN the|DT <CN><NP cat = geo
region id=0> Middle_East </NP> region|NN
</CN>?|?**



eQuery Representation

What|WP <NP cat=nationality id=1> Iranian </NP>
backed|VBD <CN> terrorist|NN groups|NNS
</CN> exist|VBP in|IN the|DT <CN><NP cat = geo
region id=0> Middle_East </NP> region|NN
</CN>?|?

eQuery Representation

What|WP <NP cat=nationality id=1> Iranian </NP>
backed|VBD <CN> terrorist|NN groups|NNS
</CN> exist|VBP in|IN the|DT <CN><NP cat = geo
region id=0> Middle_East </NP> region|NN
</CN>?|?

eQuery Representation

What|WP <NP cat=nationality id=1> Iranian </NP>
backed|VBD <CN> terrorist|NN groups|NNS
</CN> exist|VBP in|IN the|DT <CN><NP cat = geo
region id=0> Middle_East </NP> region|NN
</CN>?|?

What|WP <NP cat=nationality id=1> Iranian </NP>
<EV cat=SUPPORT id=0> backed|VBD </EV>
<CN> terrorist|NN groups|NNS </CN> exist|VBP
in|IN the|DT <CN><NP cat = geo region id=0>
Middle_East </NP> region|NN </CN>?|?



eQuery Representation

(
SUPPORTER (*support**, *Iran*)
AND
SUPPORTED (*support**, ?*x*)
AND
ISA (?*x*, *terrorist_group#*)
AND
LOCATION (?*x*, *Middle_East#*)
)



eQuery identifies:

- appropriate stemming of terms
- useful multi-word concepts
- expansion of terms with synonymous words/phrases
- *required* concept
- implicit logical requirements
- relations amongst concepts



In Summary:

- **There exist a range of approaches for representing documents and queries**
 - NLP provides meaningful representation
- **Each needs to be evaluated in terms of their ability to accomplish the goals of your application**
- **The Web has opened a whole new world of possible technologies which require traditional indexing skills**