Unsupervised ontology acquisition applications

Kalliopi Zervanou
k.zervanou@let.ru.nl

Centre for Language Studies
Radboud University Nijmegen
PO Box 9103, 6500 HD Nijmegen,
The Netherlands

Ontologies & knowledge resources

- Formal representations of knowledge
- Lexico-semantic resources important components in many NLP applications
- Resource acquisition: laborious & complicated

Automatic ontology acquisition

- Reduced time & effort of manual development
- Improved consistency
- *Techniques:* IE, data mining/machine learning
- *Data sources:*
  - unstructured/free text
  - structured documents
  - legacy/ existing domain specific resources
  - dictionaries / existing general purpose resources

Ontology layers

$\forall x, y \ (\text{sufferFrom}(x, y) \rightarrow \text{ill}(x))$

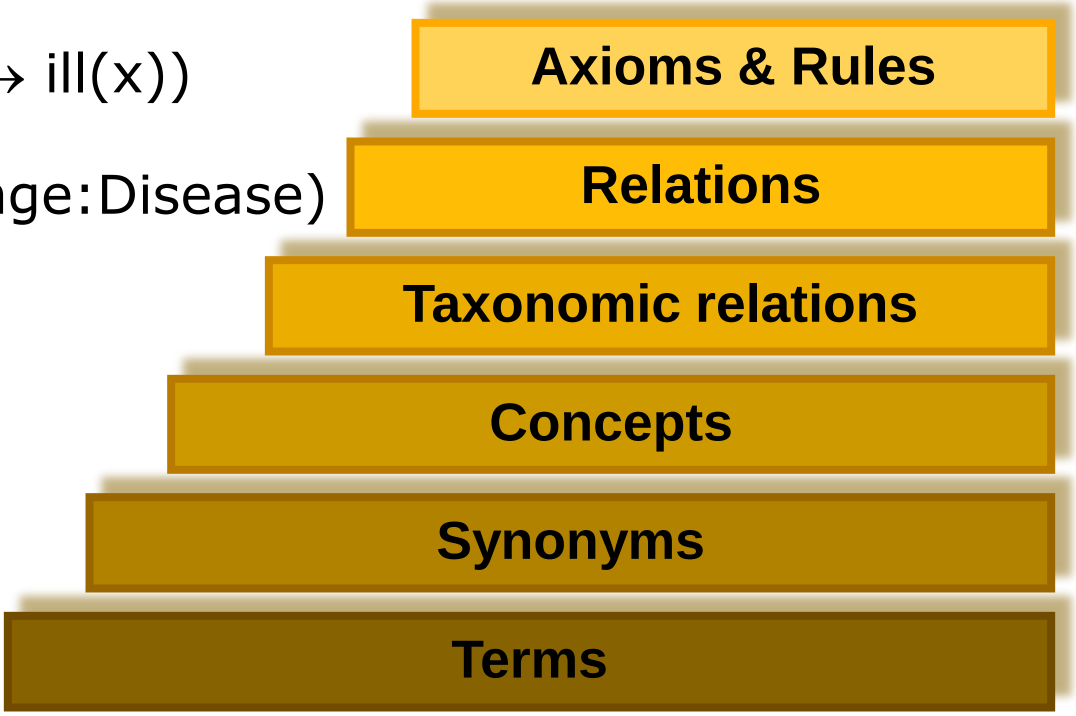
cure (domain:Doctor, range:Disease)

is_a (Doctor, Person)

Disease:=<I, E, L>

{disease, illness}

disease, illness, hospital



Axioms & Rules

Relations

Taxonomic relations

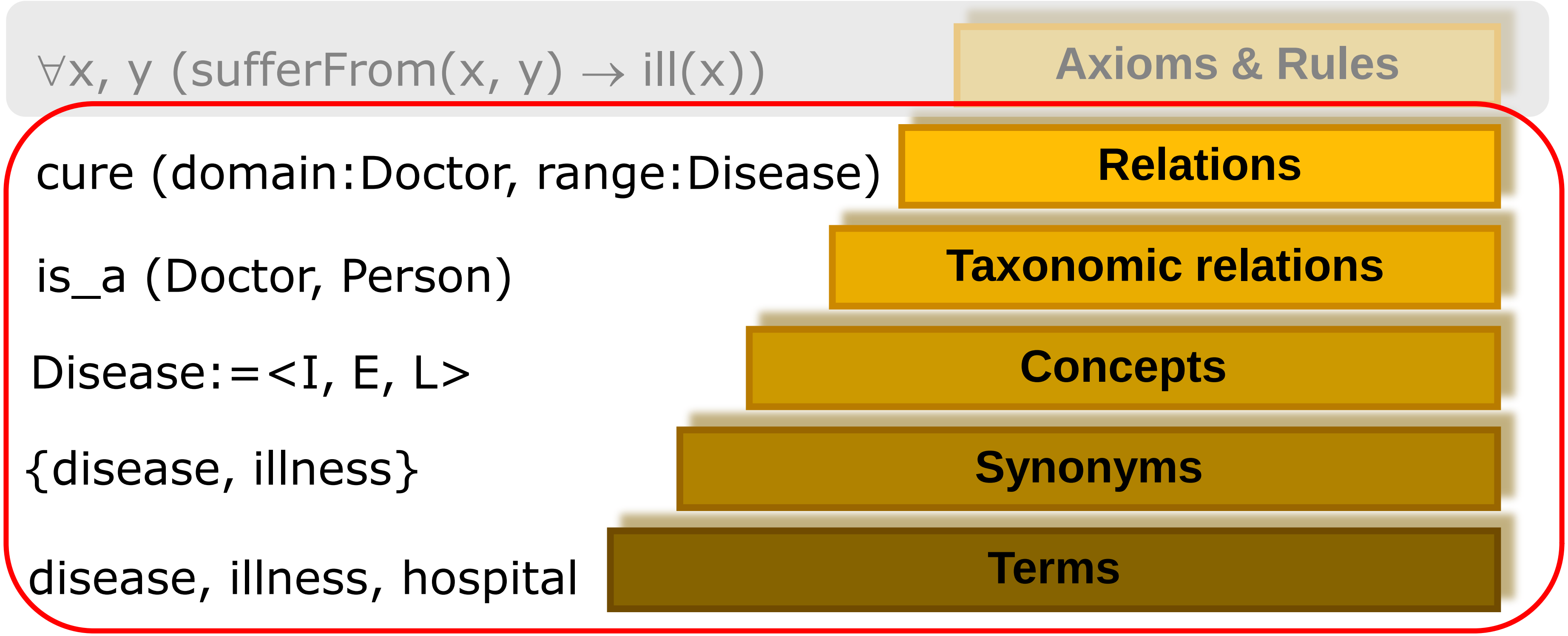
Concepts

Synonyms

Terms

The Ontology Learning Layer Cake ***[Buitelaar et al., 2003]***

Ontology layers



$\forall x, y\ (sufferFrom(x, y) \rightarrow ill(x))$ Axioms & Rules

cure (domain:Doctor, range:Disease) Relations

is_a (Doctor, Person) Taxonomic relations

Disease:=<I, E, L> Concepts

{disease, illness} Synonyms

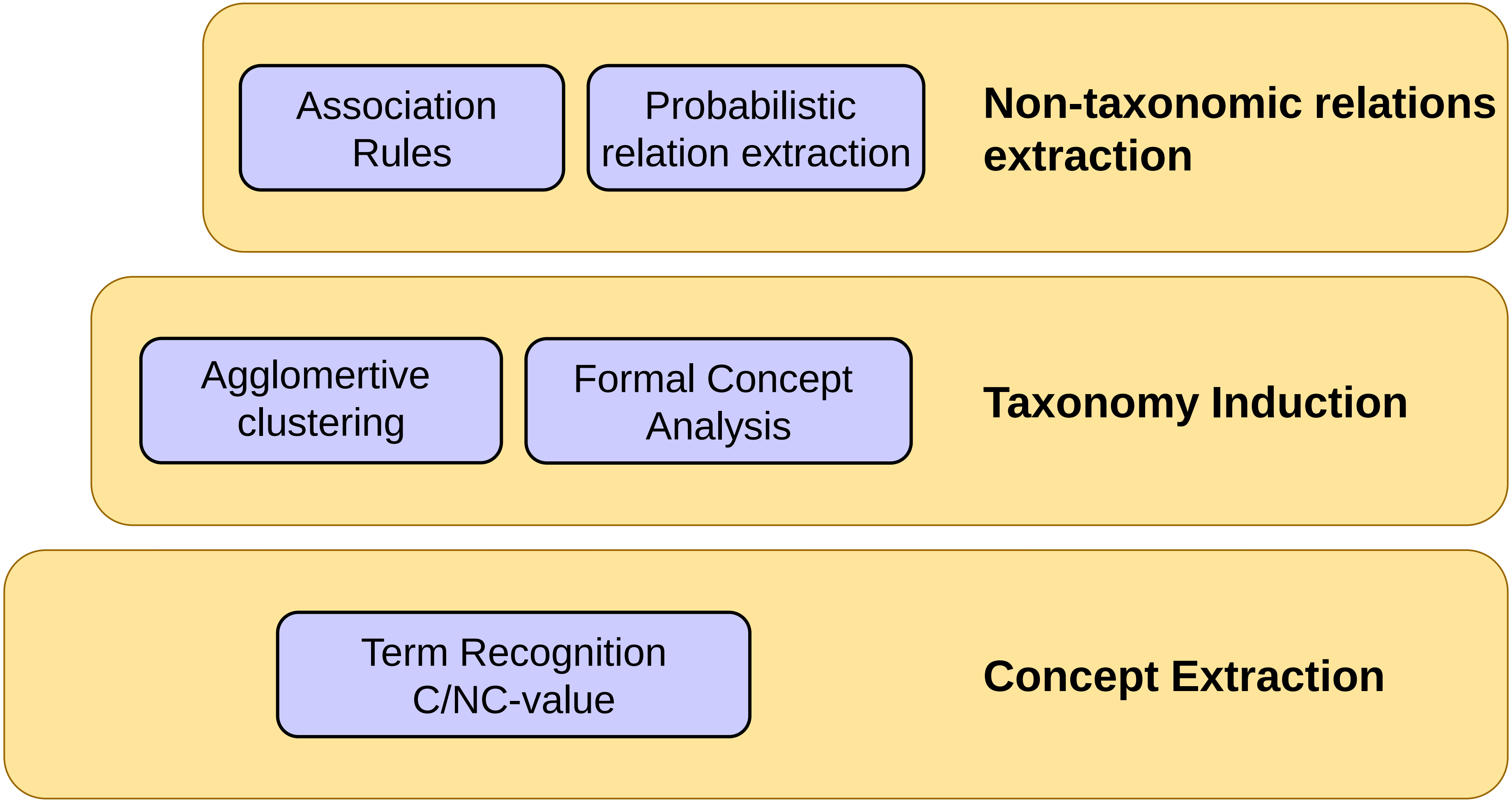
disease, illness, hospital Terms

The Ontology Learning Layer Cake ***[Buitelaar et al., 2003]***

Unsupervised ontology acquisition

- **OntoGain method [Drymonas et al., 2010]**
  - domain independent, multi-word term concepts
  - OWL output
  - implemented for the medical & computer science domains
- **EAD taxonomy method [Zervanou et al., 2011]**
  - EAD metadata (free/semi-structured text)
  - multi-word term concepts
  - social history domain

OntoGain method steps



Association Rules

Probabilistic relation extraction

Non-taxonomic relations extraction

Agglomertive clustering

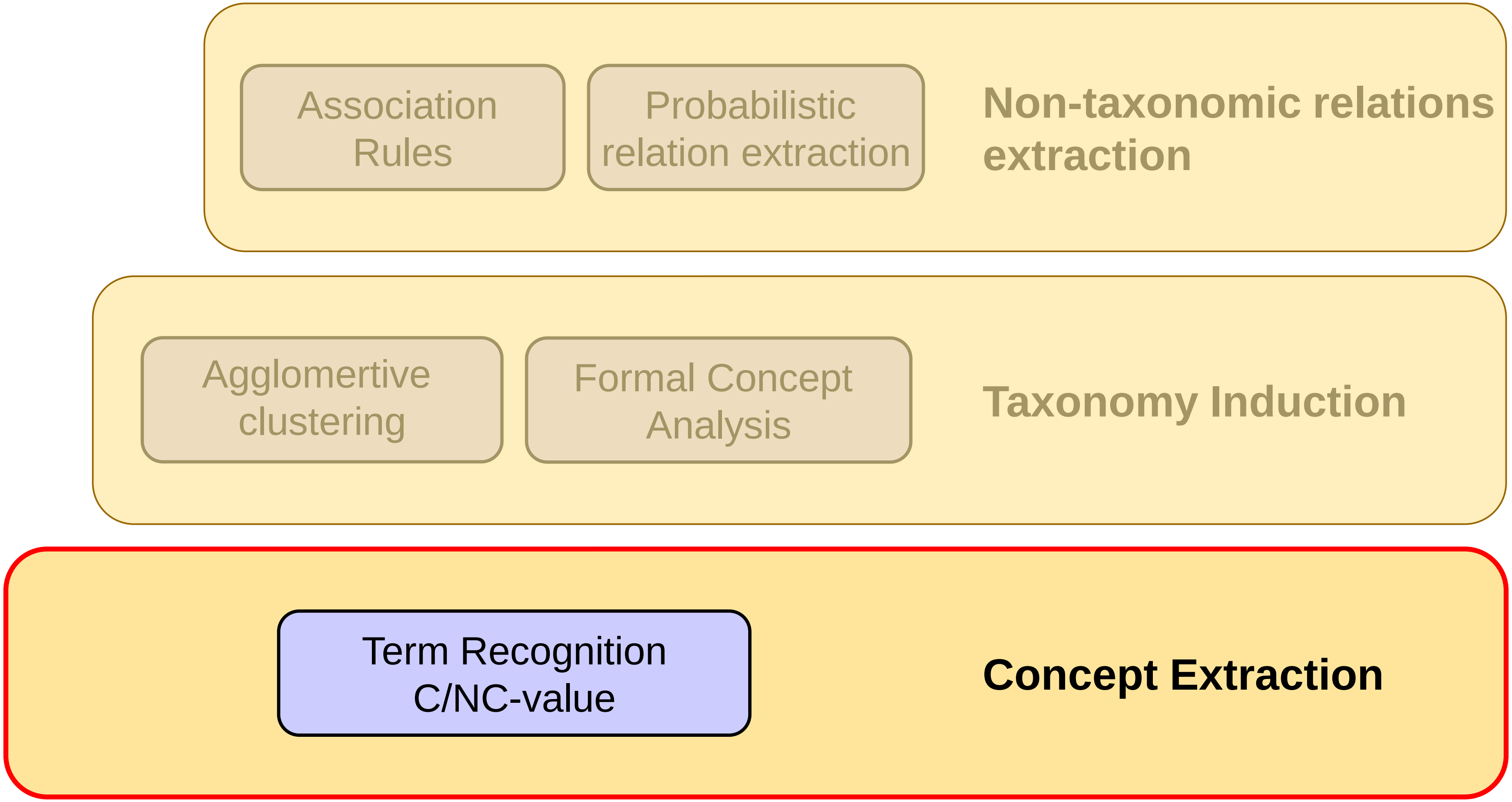
Formal Concept Analysis

Taxonomy Induction

Term Recognition C/NC-value

Concept Extraction

OntoGain method steps



Association Rules

Probabilistic relation extraction

Non-taxonomic relations extraction

Agglomertive clustering

Formal Concept Analysis

Taxonomy Induction

Term Recognition C/NC-value

Concept Extraction

Term extraction: C/NC value

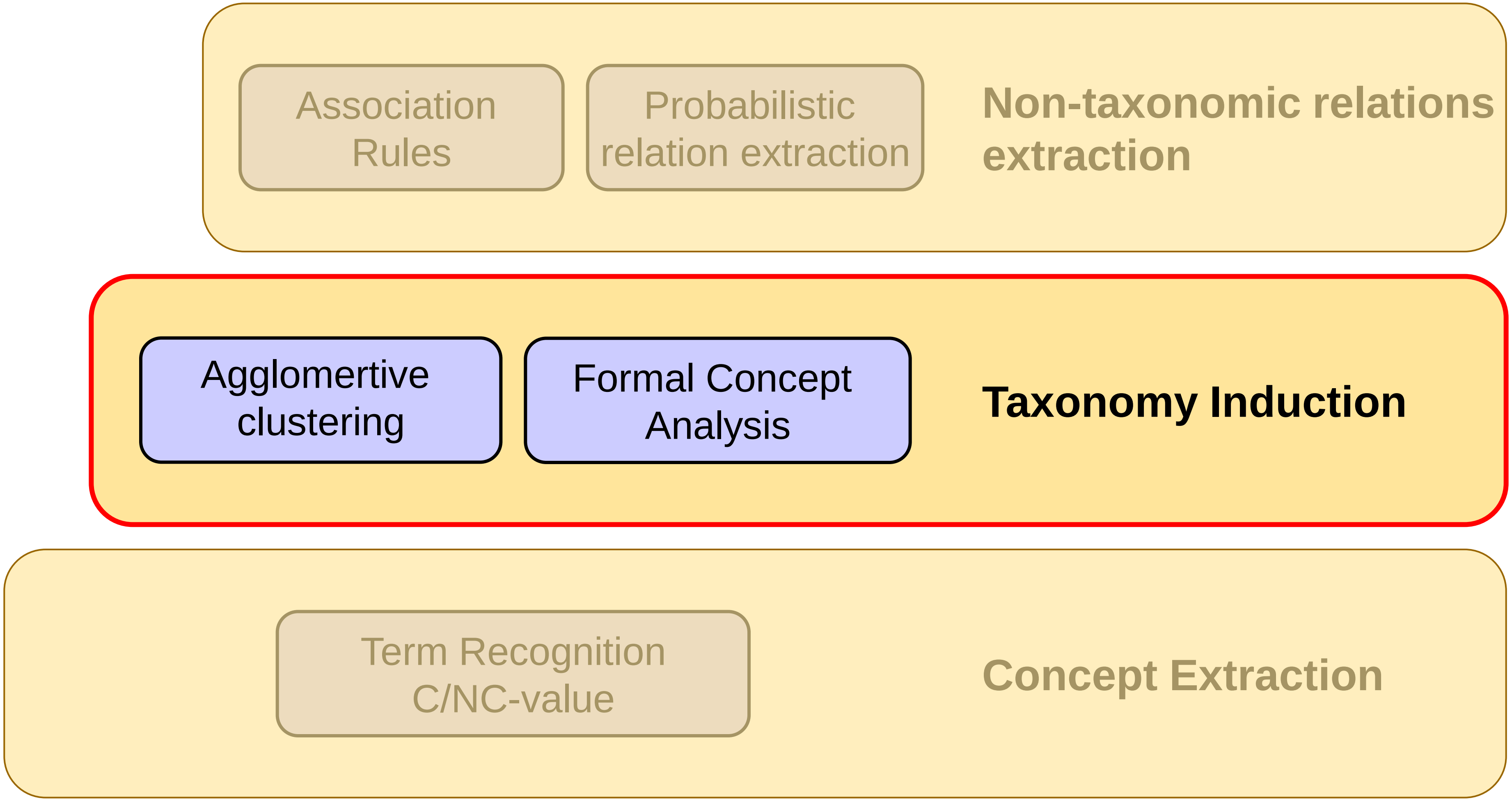
[Frantzi et.al. , 2000]

- Domain independent
- Recognition of multi-word terms
- Hybrid linguistic/statistical method
- Based on the hypothesis that:
  - longer term phrases consist of nested terms
  - term phrases tend to appear in specific context

C/NC value sample results

output term	***C/NC value***
web page	1740.11
information retrieval	1274.14
search engine	1103.99
machine learning	727.70
computer science	723.82
experimental result	655.125
text mining	645.57
natural language processing	582.83
world wide web	557.33
large number	530.67
artificial intelligence	515.73
relevant document	468.22

OntoGain method steps



Association Rules

Probabilistic relation extraction

Non-taxonomic relations extraction

Agglomertive clustering

Formal Concept Analysis

Taxonomy Induction

Term Recognition C/NC-value

Concept Extraction

Taxonomy Induction

- hierarchical structure of concepts
- **`is_a`** (hypernymic/hyponymic) relationships
- Two methods used in OntoGain:
  - Agglomerative clustering
  - Formal Concept Analysis (FCA)

Agglomerative clustering

- Proceeds bottom-up: at each step, the most similar clusters are merged
  - initially each term is considered a cluster
  - merge most similar clusters
  - similarity based on terms sharing common constituents (e.g. heads, modifiers etc.)
  - group average similarity for term clusters is computed

Formal Concept Analysis (FCA)

[Ganter et.al. , 1999]

- based on the idea that the objects (terms) are associated with their attributes (verbs)
- cluster objects based on common attributes
- formal concepts are connected with the sub-concept relationship:

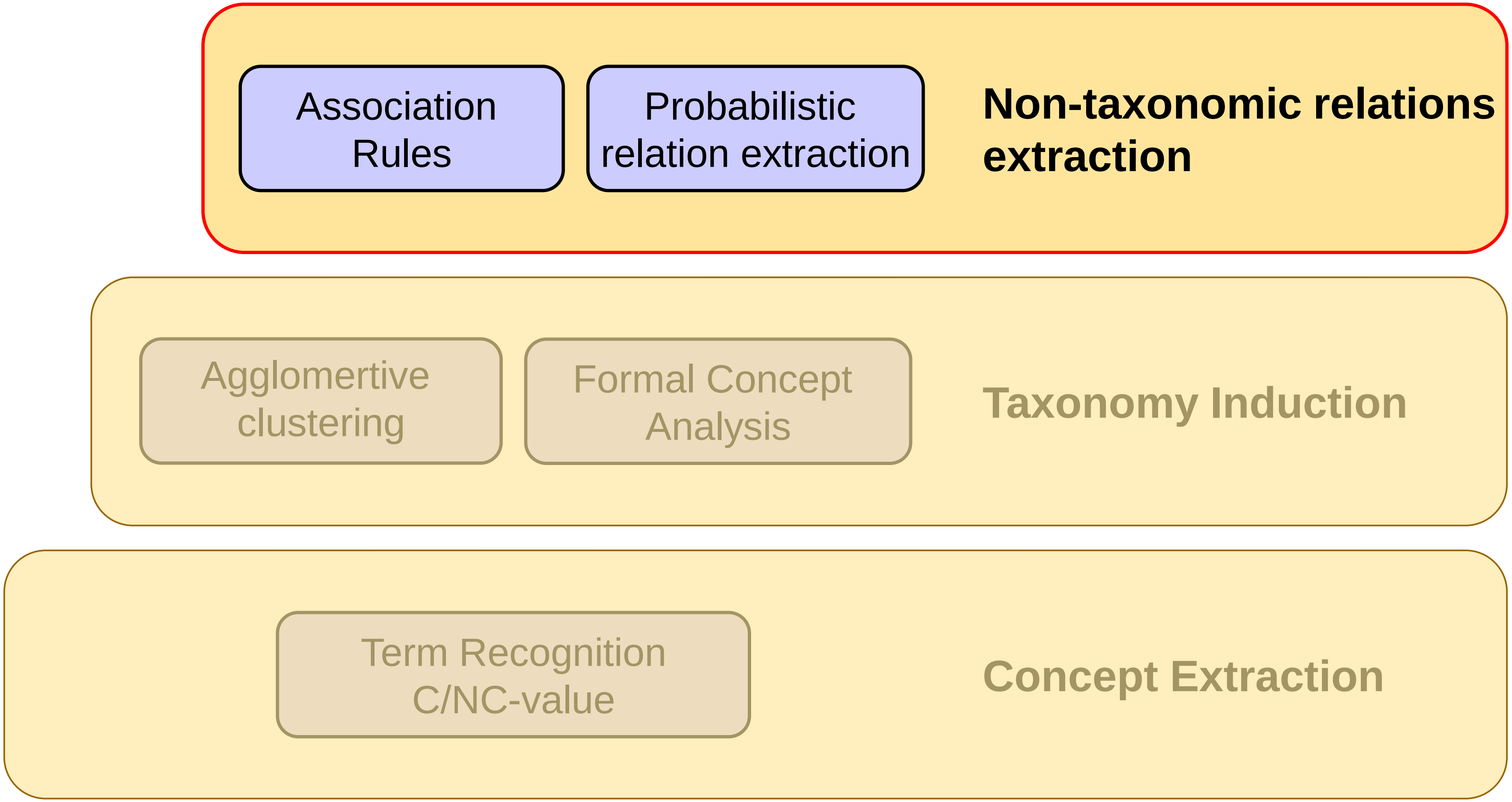
$$(O_1, A_1) \leq (O_2, A_2) \Leftrightarrow O_1 \subseteq O_2 (A_1 \subseteq A_2)$$

FCA Example:

association matrix objects vs. attributes

	submit	test	describe	print	compute	search
html form	*			*		*
hierarchical clustering					*	*
text retrieval						*
root node		*	*		*	*
single cluster			*		*	*
web page				*		*

OntoGain method steps



Association Rules

Probabilistic relation extraction

Non-taxonomic relations extraction

Agglomertive clustering

Formal Concept Analysis

Taxonomy Induction

Term Recognition C/NC-value

Concept Extraction

Non-Taxonomic Relations

- Concept attributes & relations to other concepts
- Typically expressed by a verb relating pair of concepts
- Two approaches:
  - Associations rules
  - Probabilistic approach

Association Rules

[Aggrawal et.al., 1993]

- introduced to predict the purchase behavior of customers
- identify terms connected with some relation *[subject-verb-object]*
- enhance with general terms from the taxonomy
- eliminate redundant relations

Association Rules: example relations

Domain	Range	Label
chiasmal syndrome	pituitary disproportion	cause by
medial collateral ligament	surgical treatment	need
blood transfusion	antibiotic prophylaxis	result
lipid peroxidation	cardiopulmonary bypass	lead to
prostate specific antigen	prostatectomy	follow
chronic fatigue syndrome	cardiac function	yield
right ventricular infraction	radionuclide ventriculography	analyze by

Probabilistic approach

[Cimiano et.al. 2006]

- Collect verbal relations from the corpus
- Find the most general relation using frequency of occurrence:

```
Suffer_from(man, head_ache)
Suffer_from(woman, stomach_ache)
Suffer_from(patient,ache)
```

- Select relationships satisfying a conditional probability measure

Evaluation

- Two domains: medical & computer science
- Evaluation of ontology constituent parts:
  - Terms, Taxonomic & Non-taxonomic relations
- Judgement provided by domain experts:
  - **Precision:** for top 200 terms indicate correct terms & respective relations
  - **Recall:** for 500 lines of each corpus, hand-crafted ontos compared to OntoGain results

Results

Processing Layer	Method	P OMed	R OMed	P CS	R CS
Concept Extraction	C/NC value	**89.7%**	**91.4%**	**86.7%**	**89.6%**
Taxonomic Relations	FCA	**47.1%**	**41.6%**	**44.2%**	**48.6%**
	Agglomerative Clustering	**71.2%**	**67.3%**	**71.3%**	**62.7%**
Non-Taxonomic Relations	Association Rules	**71.8%**	**67.7%**	**72.8%**	**61.7%**
	Probabilistic	**62.7%**	**55.9%**	**61.6%**	**49.4%**

The IISH EAD dataset

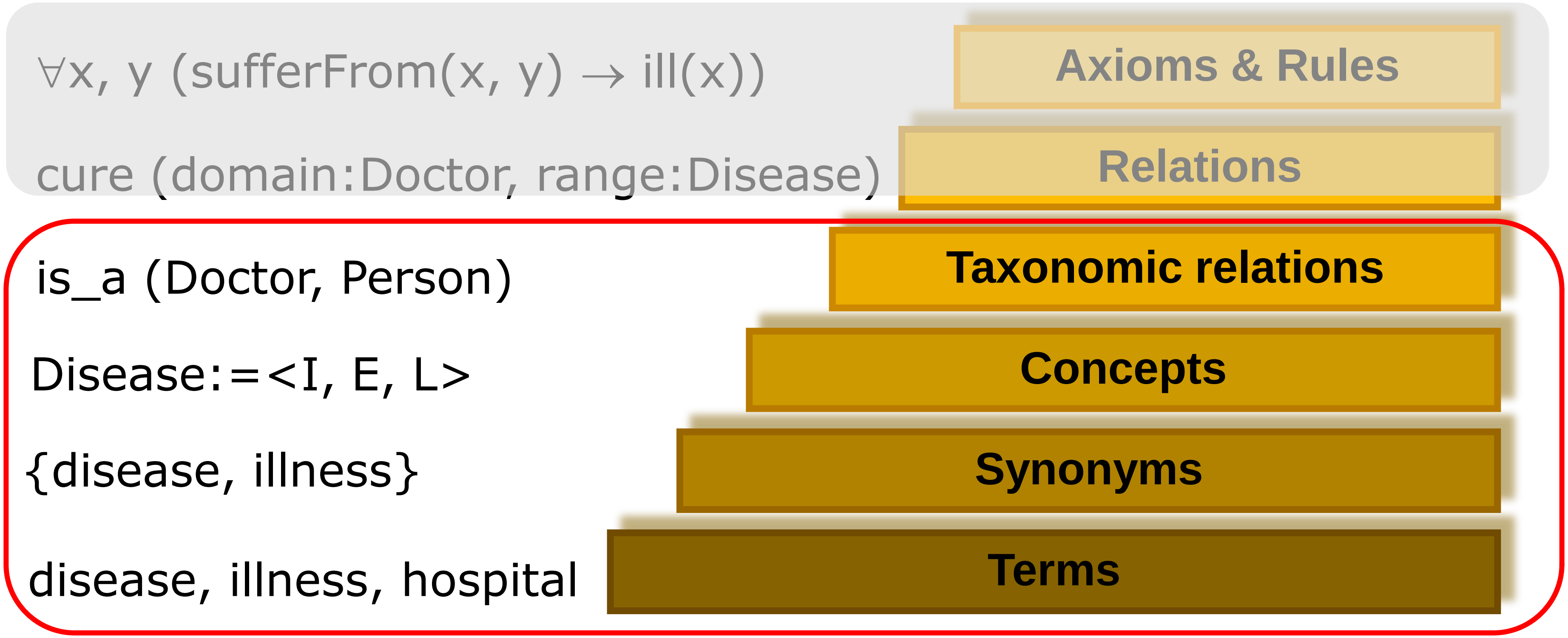
- Semi-structured metadata text (XML)
- EAD: XML standard for encoding archival descriptions

- ***Challenges:***
  - Variety of languages used
  - Varying type and amount of information
  - Style: enumerations, lists, incomplete sentences

Research on Metadata

- Developing standards:
  - collection specific (e.g. EAD, MARC21)
  - cross-collection (e.g. Dublin Core)
- Provide mappings:
  - across schemas
  - ontologies (ad hoc or standard CDOC-CRM)
- Discard metadata for IR [Koolen et al., 2007]
- Exploit metadata for IR [Zhang&Kamps, 2009]

Ontology layers



∀x, y (sufferFrom(x, y) → ill(x))

Axioms & Rules

cure (domain:Doctor, range:Disease)

Relations

is_a (Doctor, Person)

Taxonomic relations

Disease:=<I, E, L>

Concepts

{disease, illness}

Synonyms

disease, illness, hospital

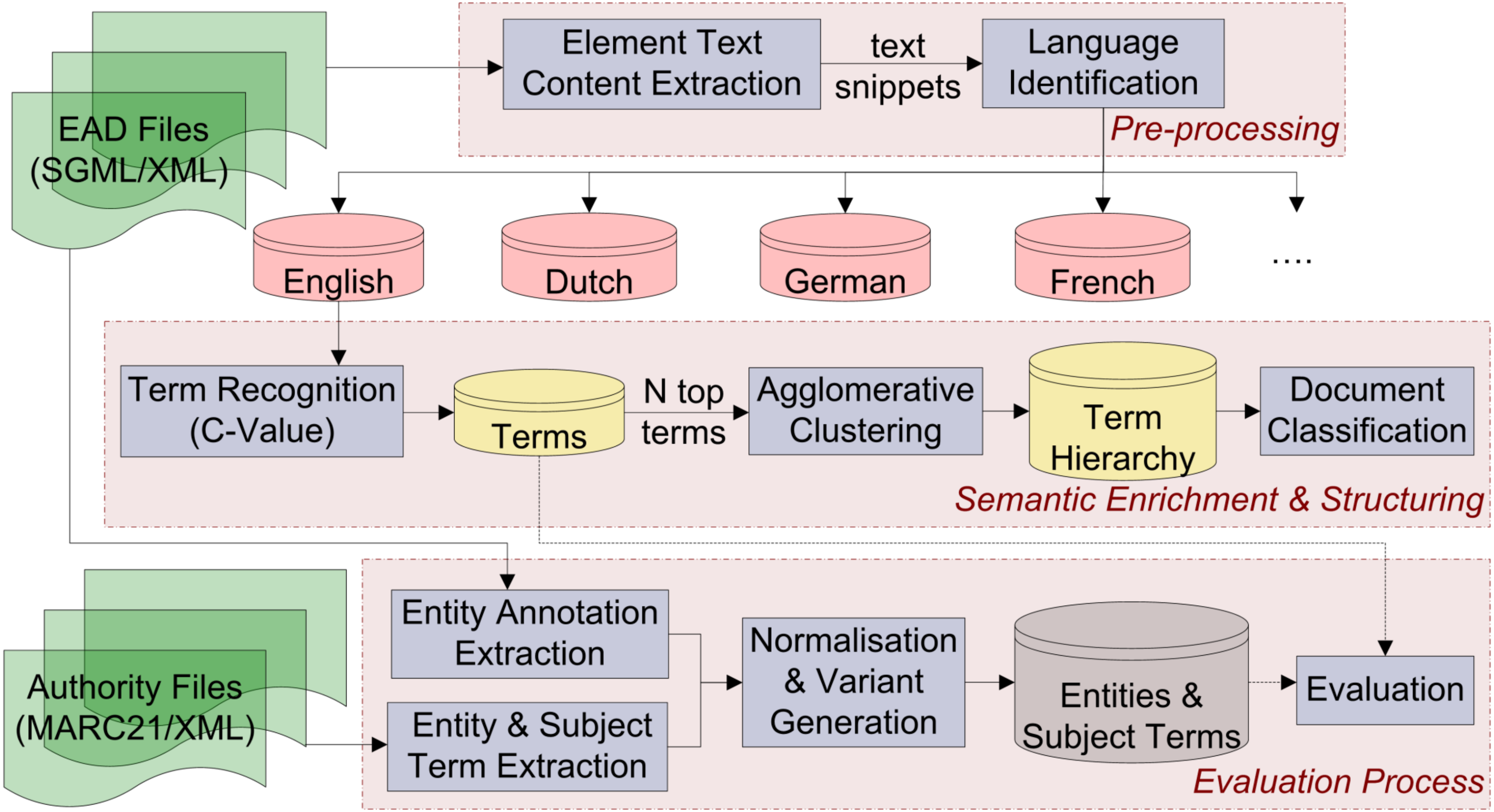
Terms

The Ontology Learning Layer Cake ***[Buitelaar et al., 2003]***

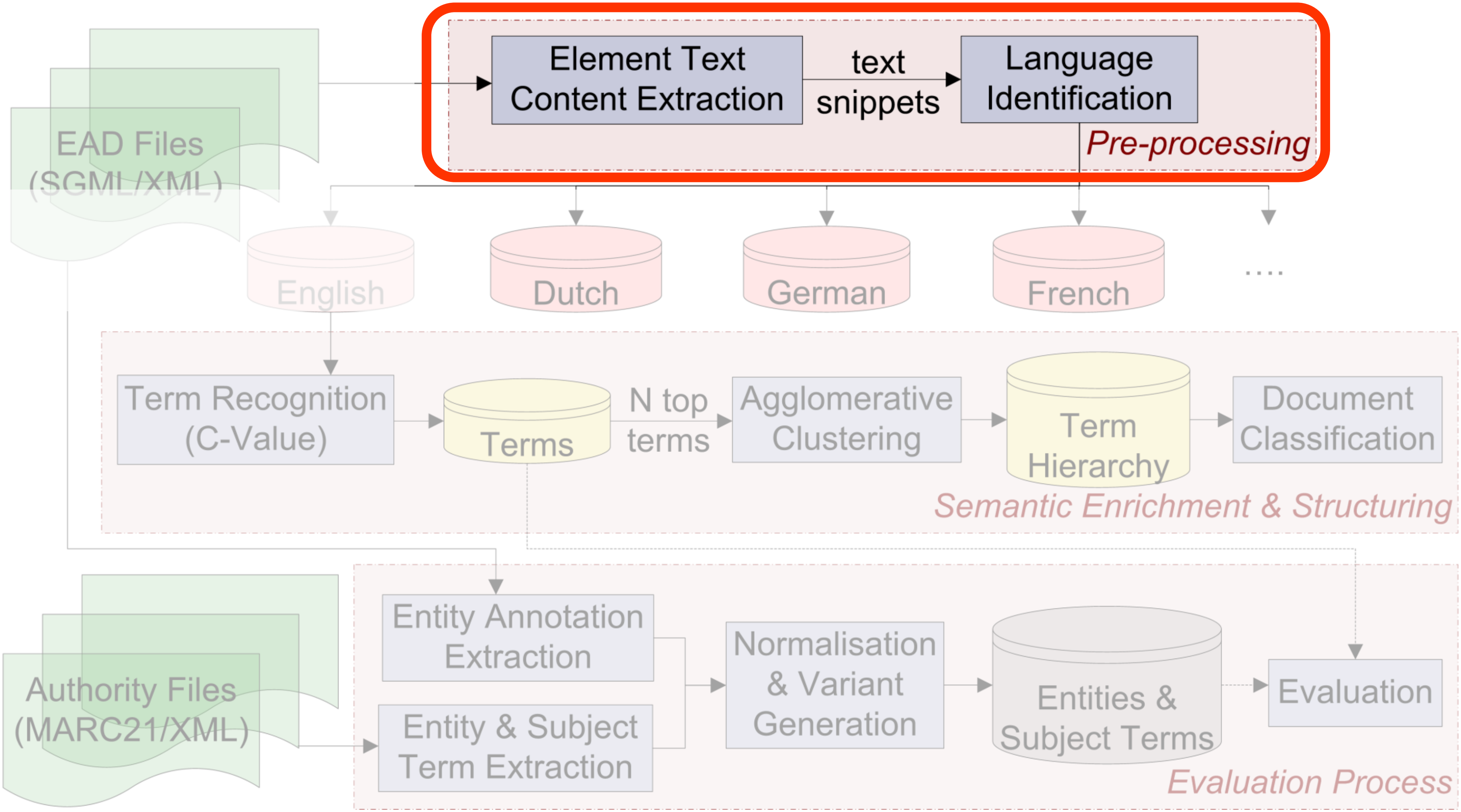
Improved search and retrieval

- Cluster metadata documents based on content
- Support content-based/semantic search
- Support exploratory research
- Link across collections, metadata formats & institutions
- Create unified metadata knowledge resources
  - Ontology & respective knowledge base

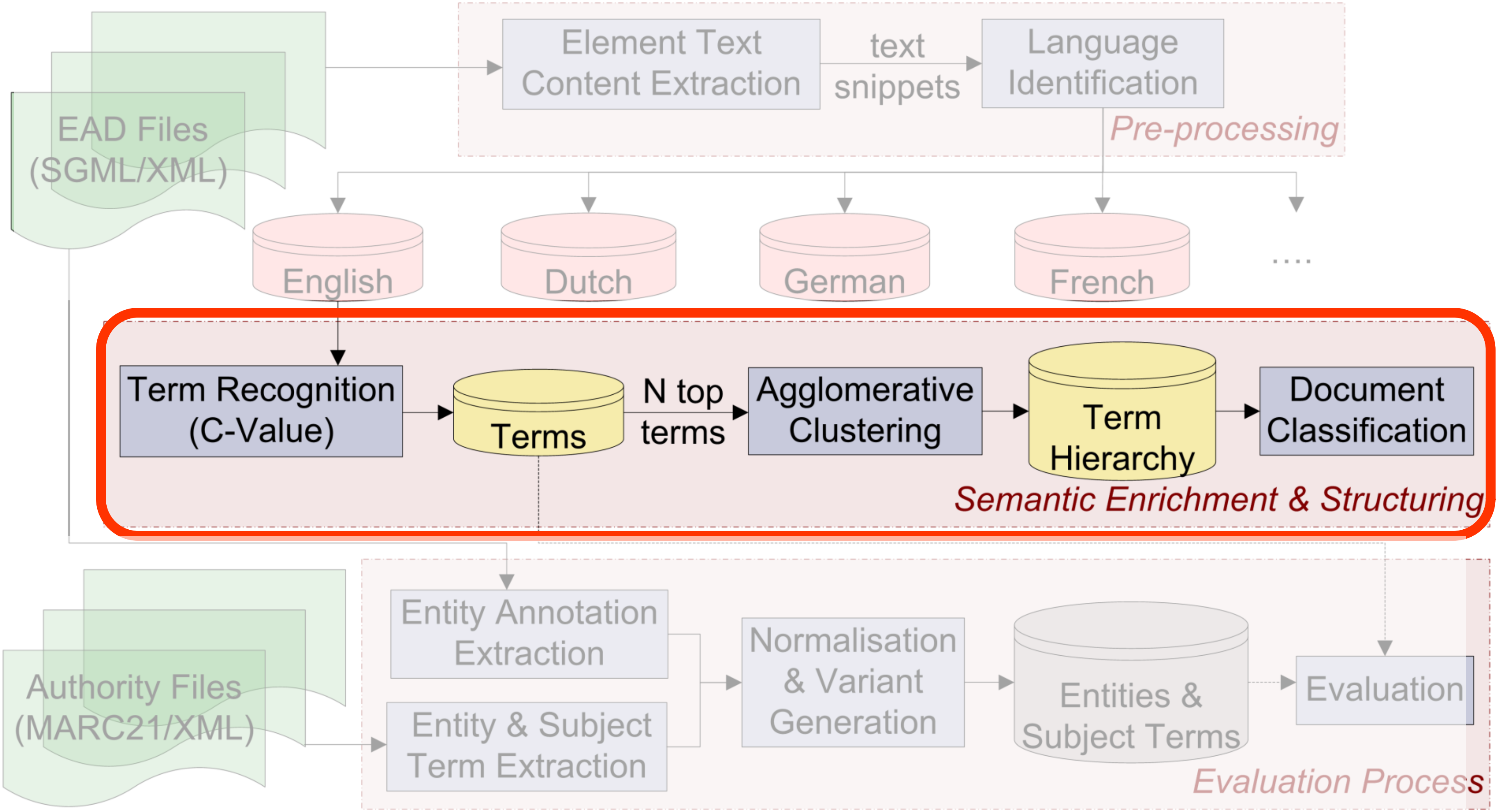
EAD Taxonomy method



EAD Taxonomy method



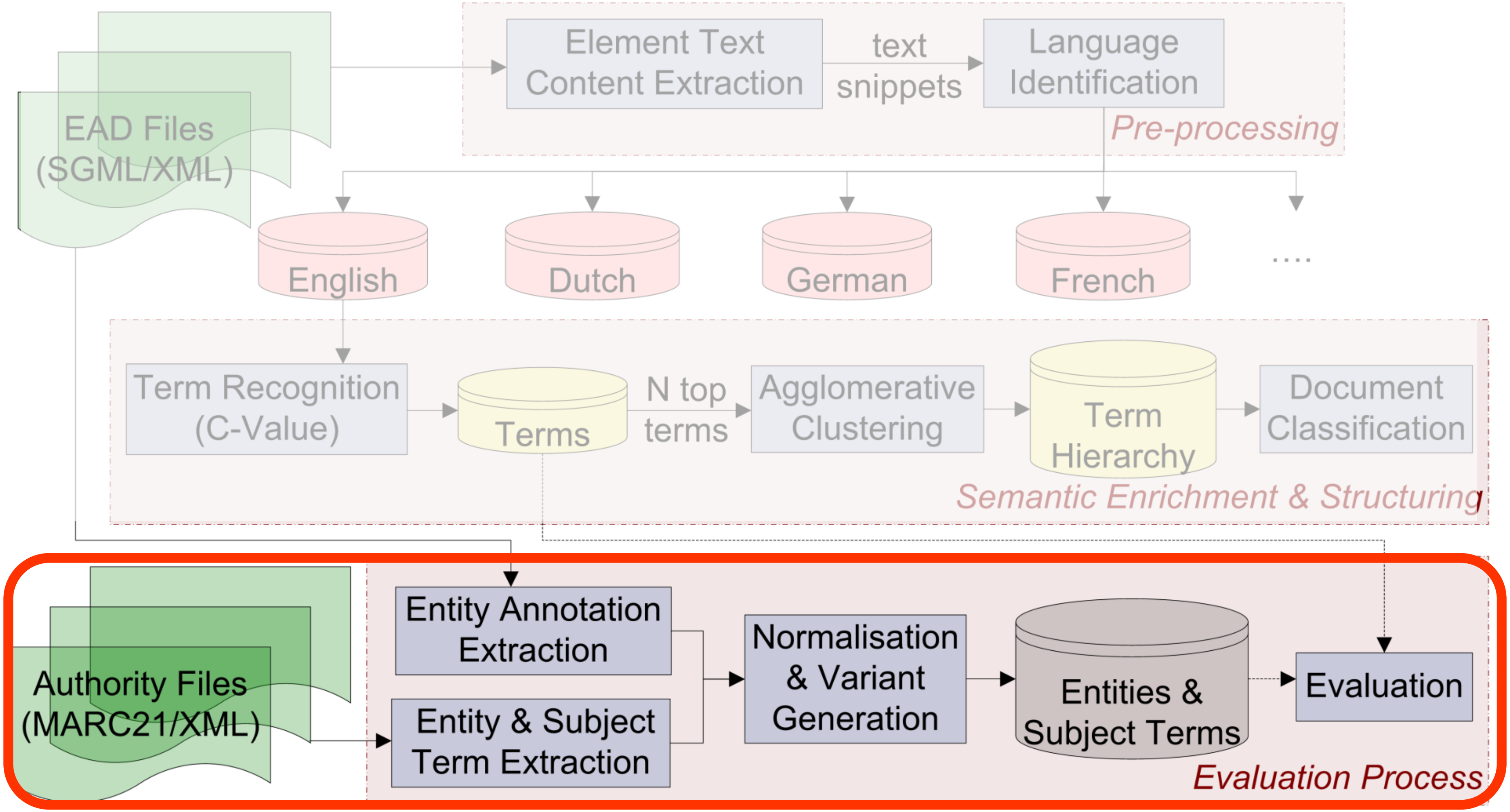
EAD Taxonomy method



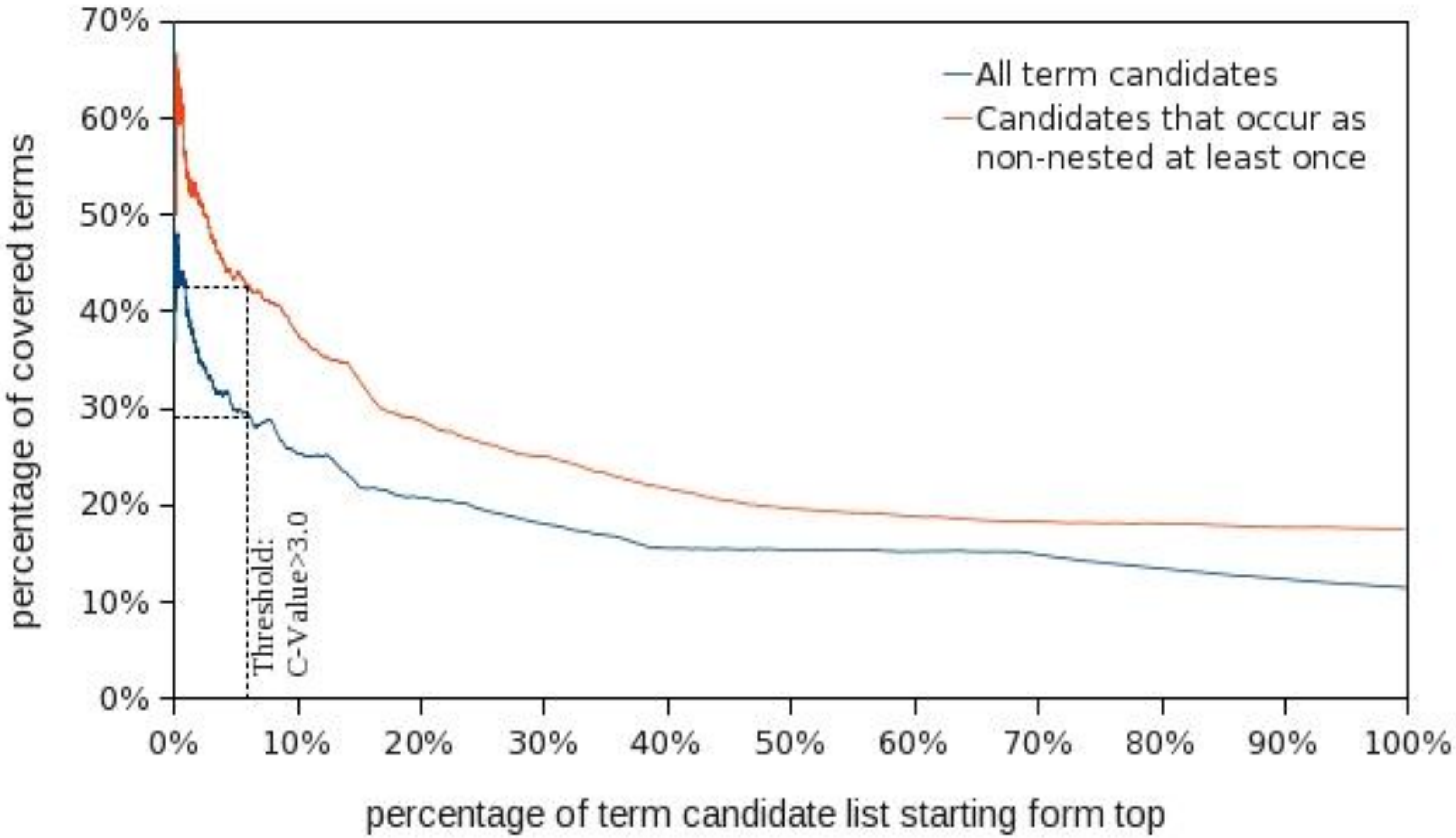
Enrichment & structuring

- *Topic detection:*
  - automatic term recognition (C value)
- *Agglomerative clustering:*
  - complete, single & average linkage criteria
  - document co-occurence & lexical similarity measures

EAD Taxonomy method



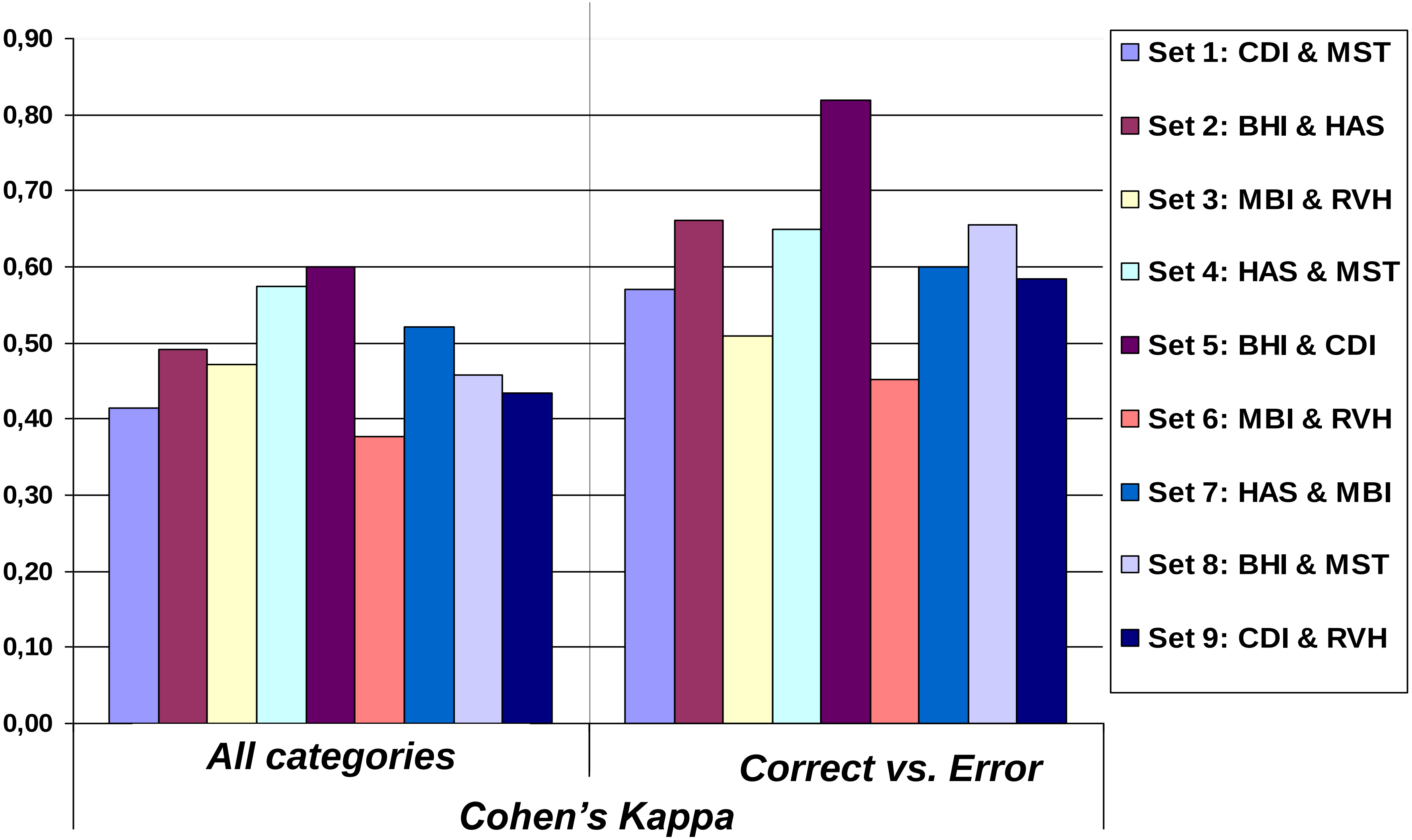
Term results (auto evaluation)



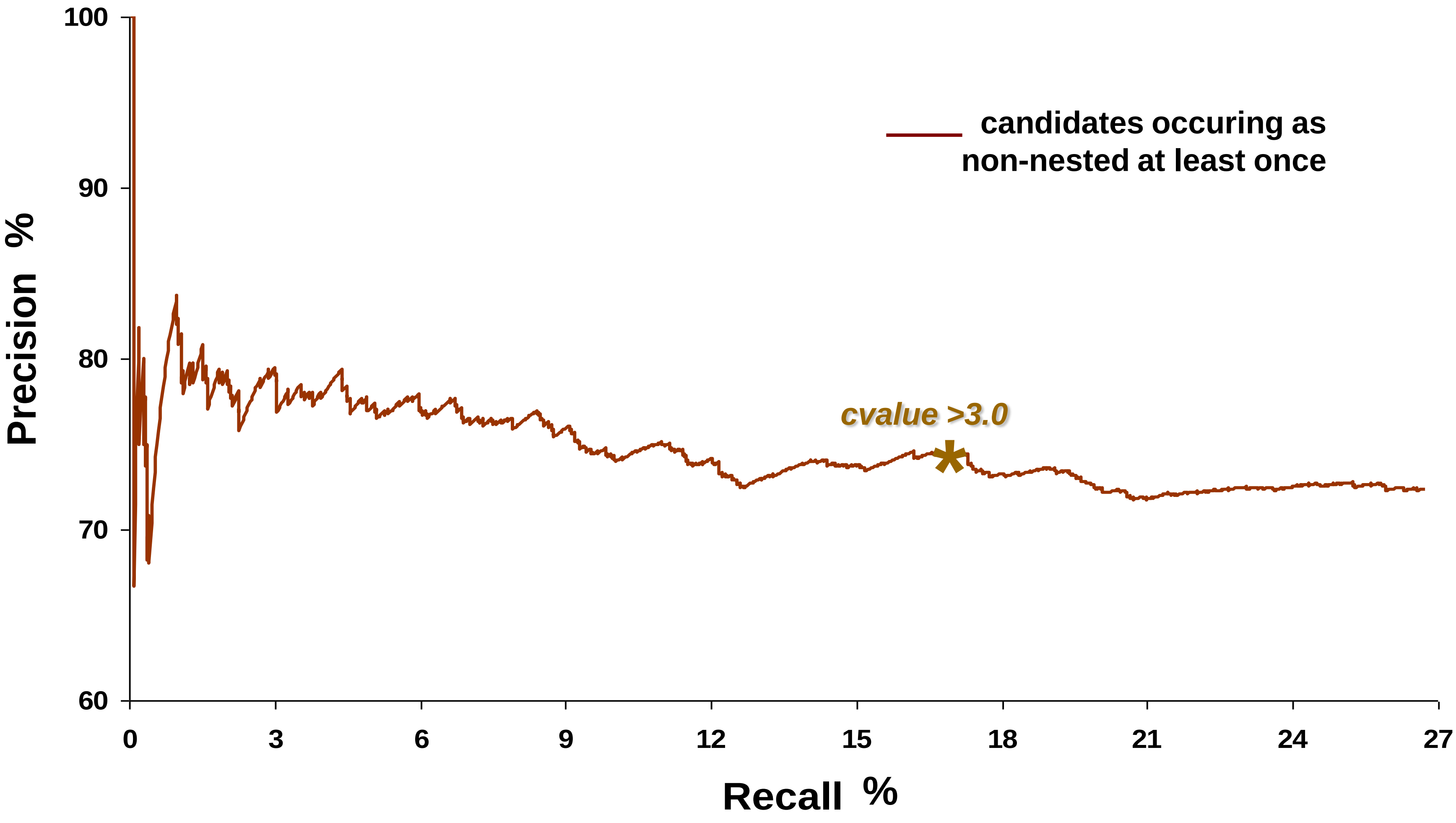
Manual evaluation

- 9 annotation sets, each consisting of 100 non-annotated terms (total: 900 candidates)
- Terms ordered by decreasing cvalue
- CValue threshold > 2.77
- Annotators: curators & archivists at IISH
- 5 categories + “Error”

Interannotator agreement



Precision vs. Recall



Results

- **C-value best performance:** candidates that occur as non-nested at least once
- Good results in term extraction
  - In manual set 1104 correct in 1526 top cand.
  - Found “new” terms/concepts
- **Average linkage criterion & Doc Co-occurence:** seem to provide broader and richer hierarchies
- Former evaluation of clustering required

Questions?

Main references:

- ***Drymonas, E., K. Zervanou and E.G.M. Petrakis (2010). Unsupervised Ontology Acquisition from Plain Texts: the OntoGain System. In: NLDB 2010, Springer: LNCS, vol. 6117, pp. 277-287.***

- ***Zervanou, K., I. Korkontzelos, A. van den Bosch and S. Ananiadou (2011). Enrichment and structuring of archival description metadata. In: LaTeCH-2011, pp. 44-53.***