Kolflow WP2
continuous extraction, edition, and annotation

ANR kolflow

Organization

Task number 2	**Extraction**			
Responsible	Orpailleur			
Activities	extracting knowledge units from texts using data mining			
Participants	Score	Edelweiss	Orpailleur	Silex
Person-months per participant	30.2	2	31.4	1

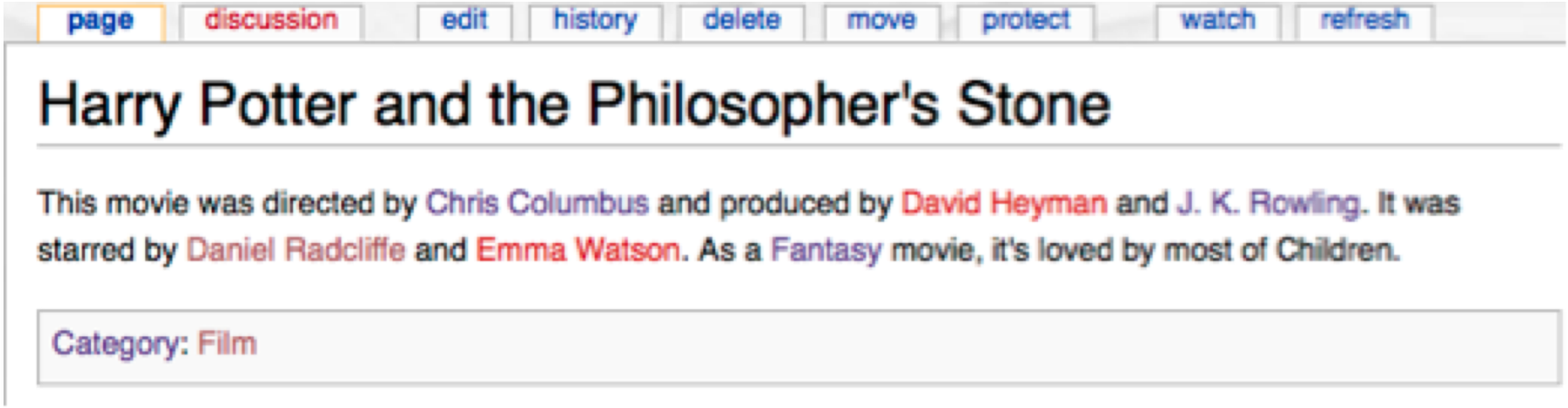
- T2.1 Collecting data (0-6)
- T2.2 Formal methods for knowledge Extraction (6-12)
- T2.3 Continuous extraction of knowledge (6-36)
- T2.4 Semantic Annotation (6-36)

- T2.1 Collecting data (0-6)
- T2.2 Formal methods for knowledge Extraction(6-12)
- T2.3 Continuous extraction of knowledge (6-36)
- T2.4 Semantic Annotation (6-36)

- D2.1: Building a corpus for experimenting continuous knowledge extraction (joined with D6.1) (6)
- D2.2: Integrating a knowledge extraction system from texts in a semantic wiki (12)
- D2.3: Specification of a continuous knowledge extraction system (18)
- D2.4: Dynamic semantic annotation in a semantic wiki: definitions and specifications (36)
- D2.5: Continuous extraction, edition and annotation in a semantic wiki (36)

Formal Concept Analysis for knowledge building (T2.2)

Formal Concept Analysis for knowledge building




```
[[Category:Film]]This movie was directed by [[directedBy::Chris Columbus]]
and produced by [[producedBy::David Heyman]] and [[producedBy::J. K.
Rowling]]. It was starred by [[Starring::Daniel Radcliffe]]
and [[Starring::Emma Watson]]. As a [[hasGenre::Fantasy]] movie, it's loved
by most of Children.
```

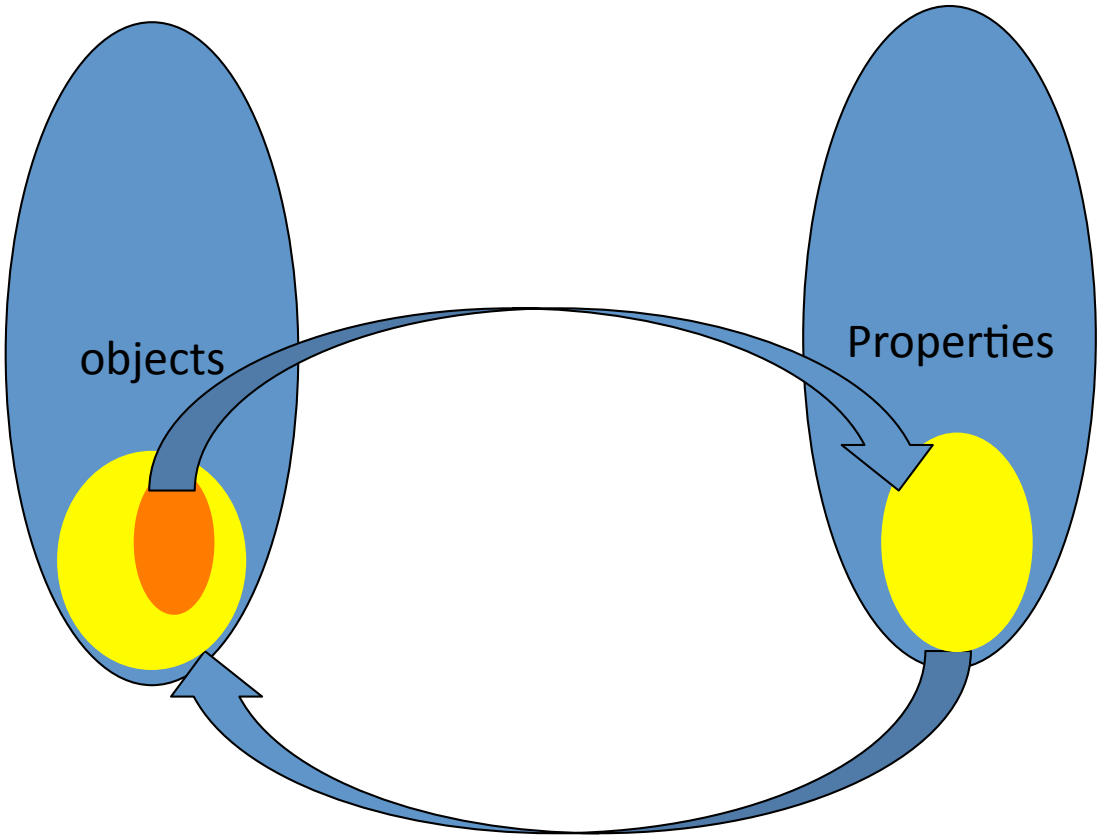
Object, properties, and relations

	French	English	Germany	Romance	ComedyDrama	hasAwards	hasRunningTime	hasYear	American
Jeux d'enfants	x			x			x	x	
Good Bye, Lenin			x		x	x	x	x	
Catch me if you can		x			x		x	x	x
And now my love	x			x		x	x	x	
America's Sweethearts		x		x	x		x	x	x
Kleinruppin Forever				x			x	x	

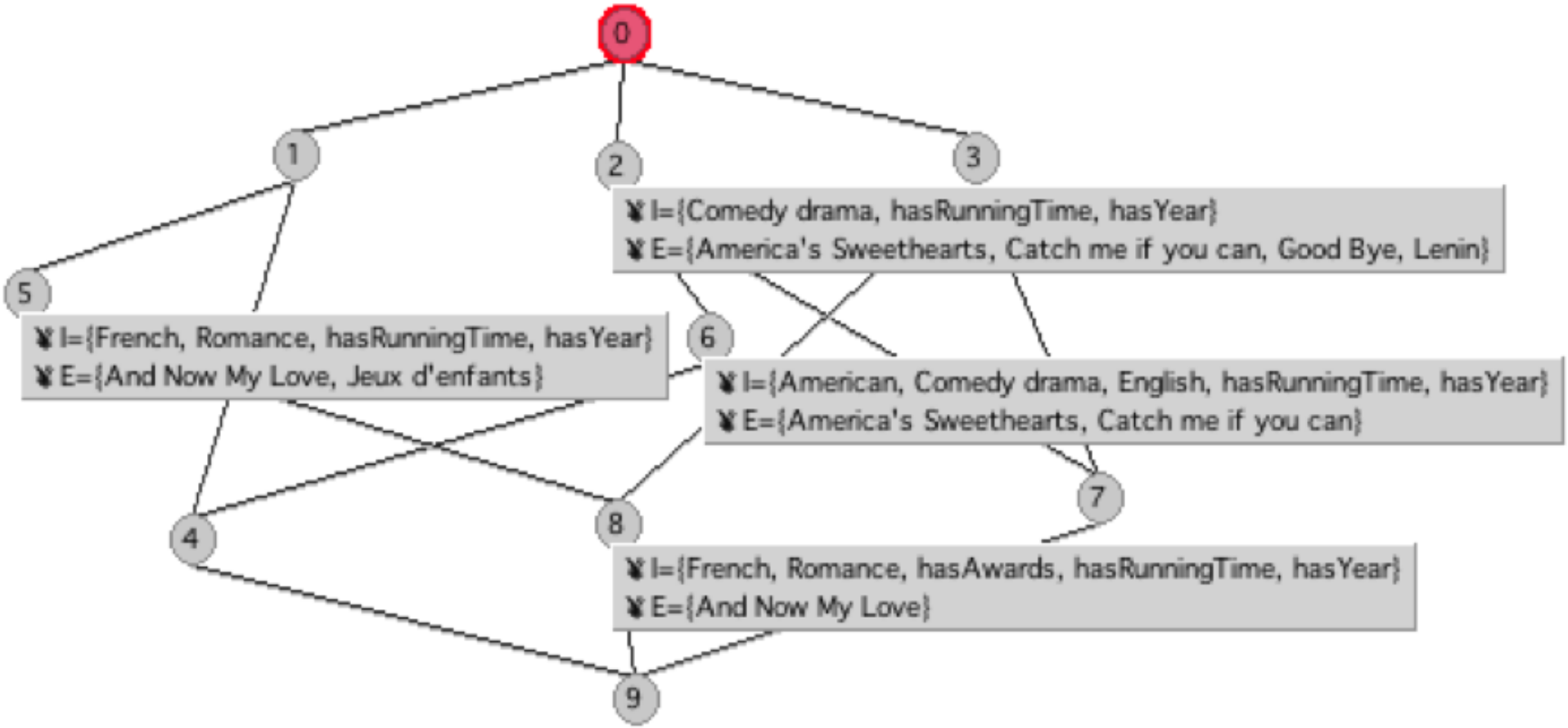
	hasAwards	Female	Male	Age20	Age30
Guillaume Canet	x		x		x
Daniel Brühl	x		x		x
Leonardo DiCaprio	x		x	x	
Marthe Keller		x			
Tolias Schenke		x			
Julia Roberts	x	x			
Catherine Zeta Jones		x			
Anna Maria Muhe		x		x	

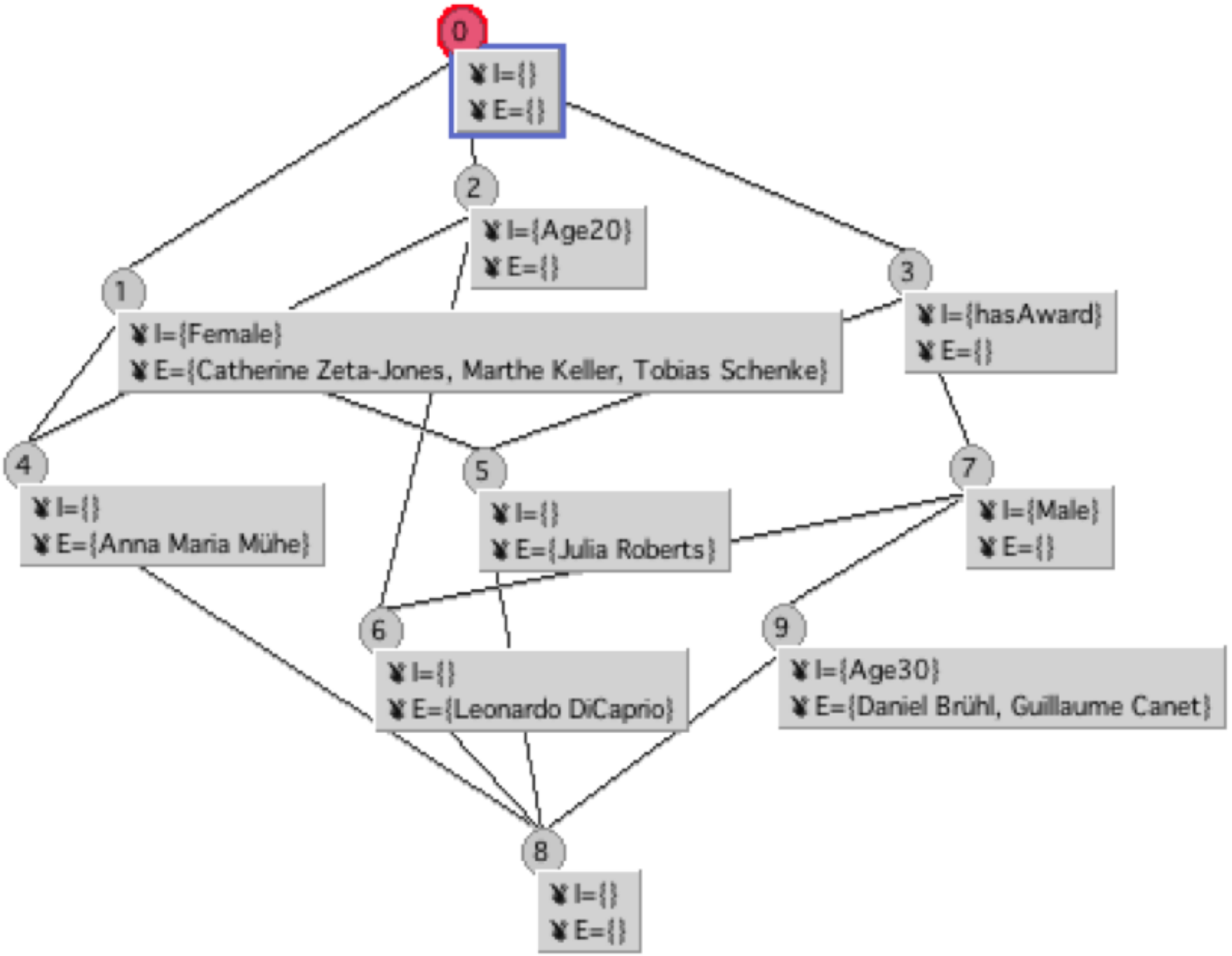
Formal concepts

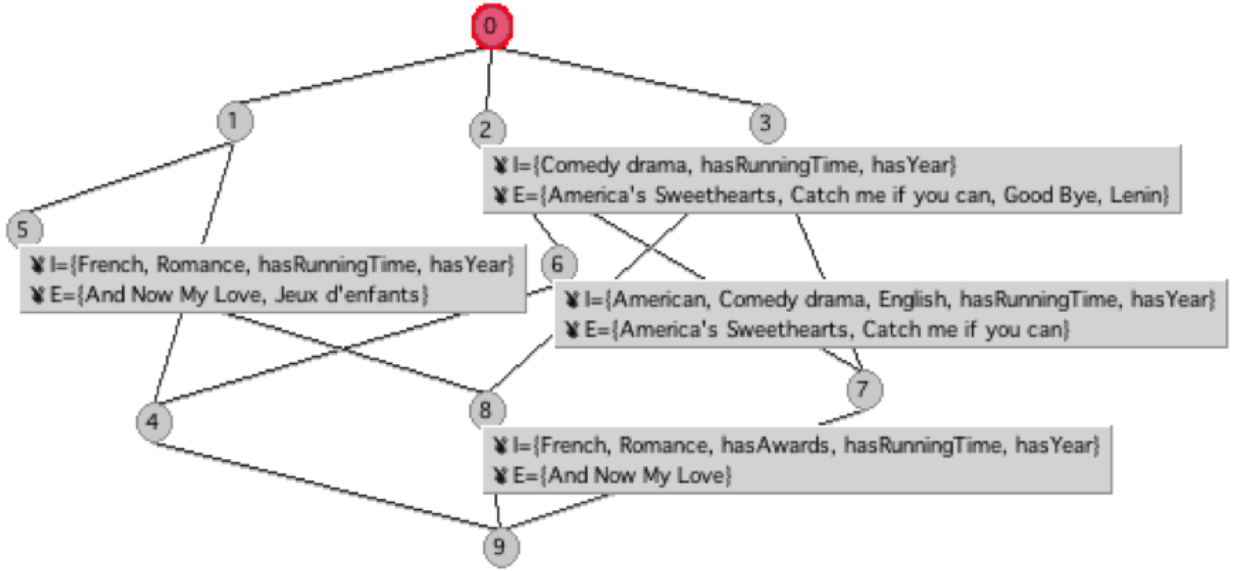
formal concept;
(A,B)
(extent , intent)



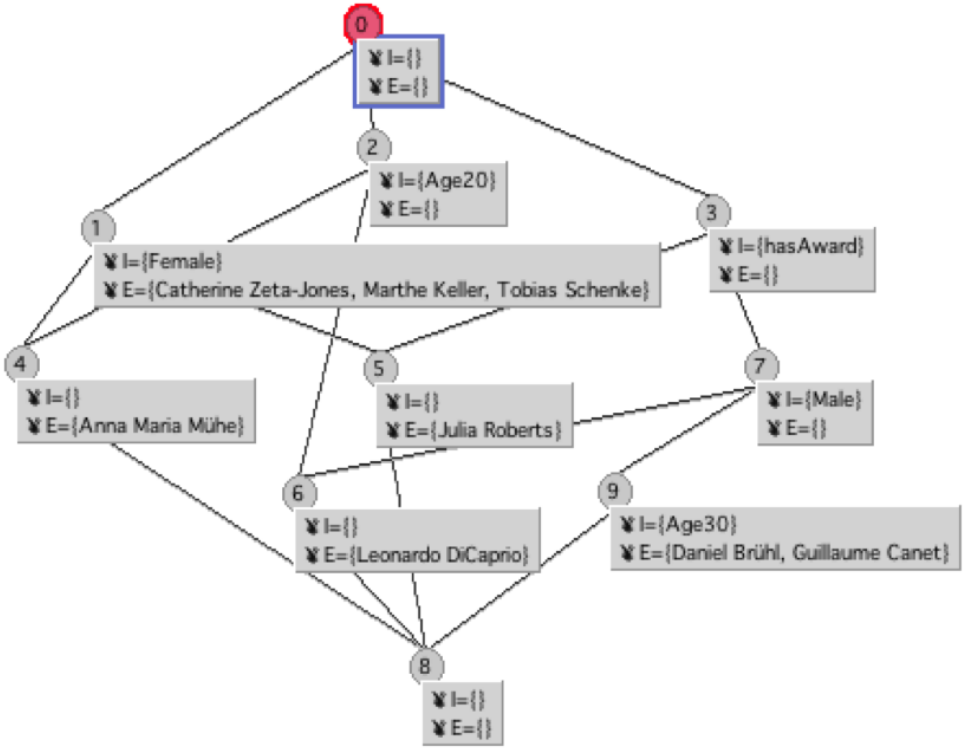
...and the lattice



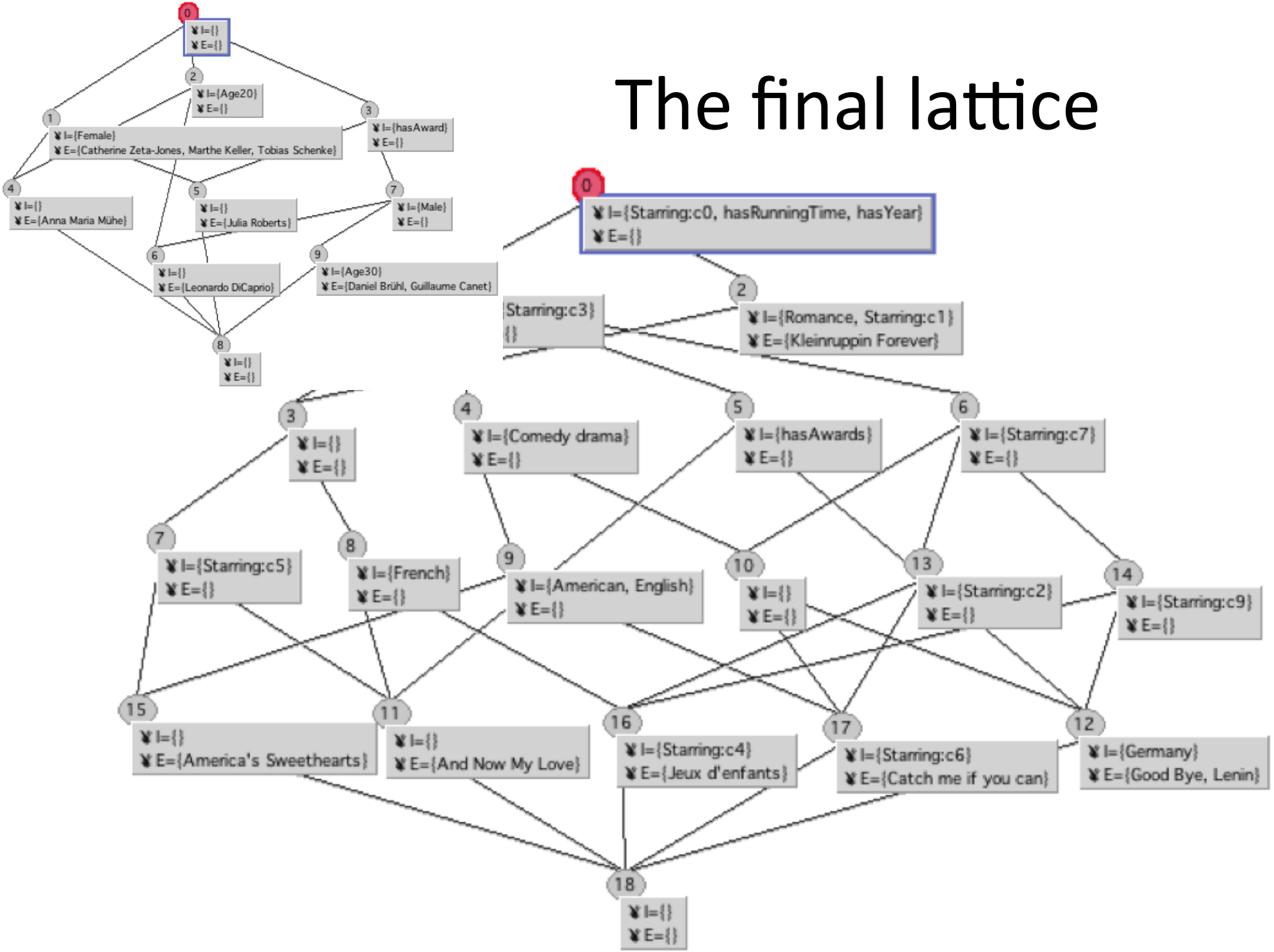

0
I={}
E={}
2
I={Age20}
E={}
1
I={Female}
E={Catherine Zeta-Jones, Marthe Keller, Tobias Schenke}
3
I={hasAward}
E={}
4
I={}
E={Anna Maria Mühe}
5
I={}
E={Julia Roberts}
7
I={Male}
E={}
6
I={}
E={Leonardo DiCaprio}
9
I={Age30}
E={Daniel Brühl, Guillaume Canet}
8
I={}
E={}



	Guillaume	Daniel Brühl	Leonardo DiCaprio	Marthe Keller	Tolias Schenke	Julia Roberts	Catherine Zeta Jones	Anna Maria Muhe
Jeux d'enfants	x							x
Good Bye, Lenin		x						
Catch me if you can			x					
And now my love				x		x		
America's Sweethearts						x	x	
Kleinruppin Forever				x	x			



The final lattice



Reading results

- A new structure given by the lattice
- Equivalence between classes (C9)
- Suggesting new categories (C5)
- Category subsumption (C6 subsumed by C2)
- Defining categories by properties and relations

Open problems

- What the objects, properties, relations are?
- Multi-valued attributes
- Formal concept = concept?
  - What about noise and silence?
- Naming concepts
- Adapting the lattice to expert needs

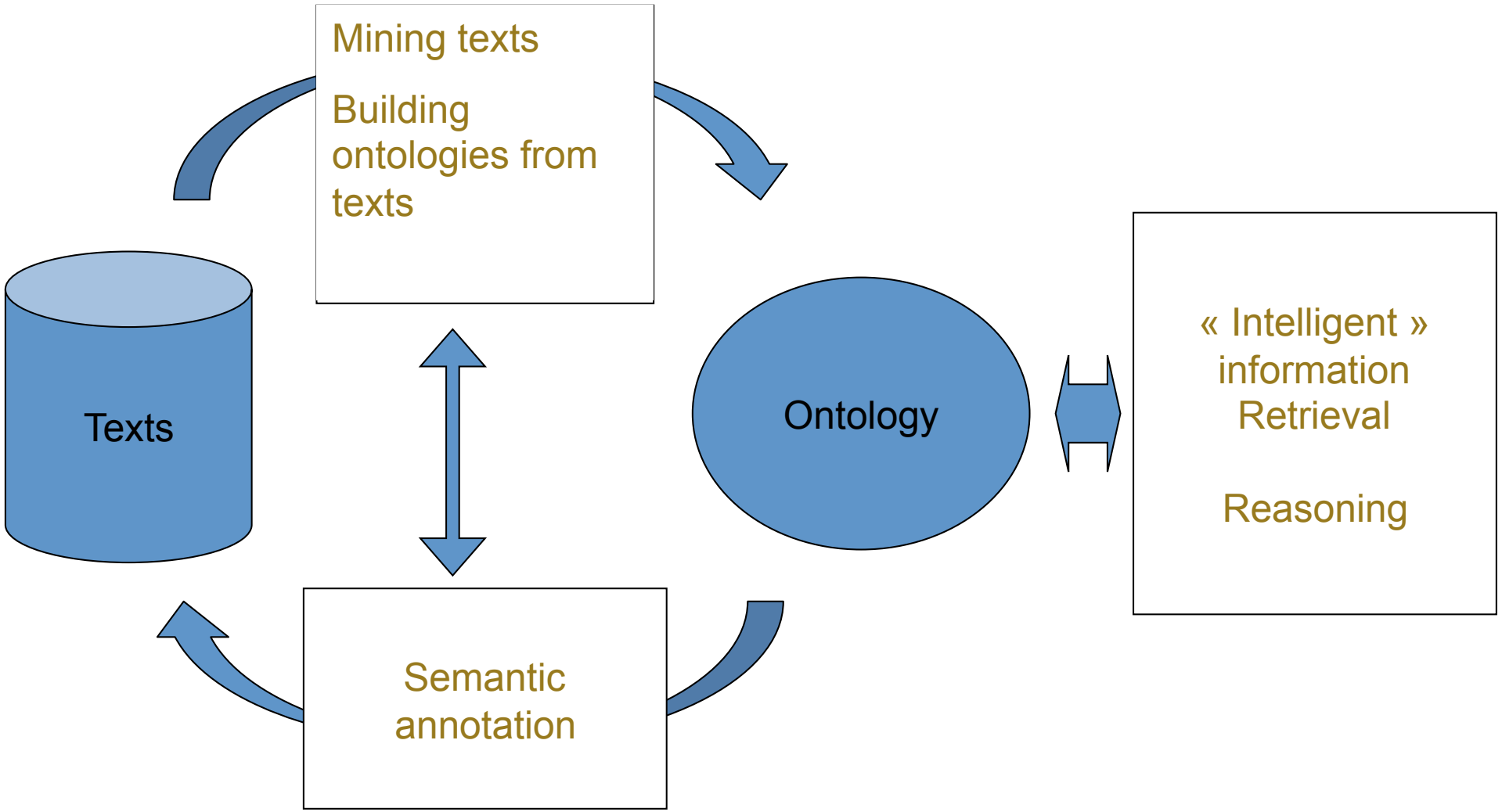
Towards a continuous process: Adapting to expert needs

- Using external resources
- Modifying input data
  - objects, properties, relations
  - modifying the "incidence" relation
  - Modifying external resources

But

- How to know what should be changed in input data ?
- What about traceability with texts
- Annotation (using terms or definitions)
- How to trace the change between the previous lattice and the new one.

Semantic Annotation


Mining texts
Building ontologies from texts
Texts
Ontology
« Intelligent » information Retrieval
Reasoning
Semantic annotation

AminoAcid Antibio Bacterie Codon Enzyme Gene NomMutant Plasmid Protein

Fluoroquinolone resistance (FQ-R) in clinical isolates of Enterobacteriaceae species has been reported with increasing frequency in recent years. Two mechanisms of FQ-R have been identified in gram-negative organisms: mutations in DNA gyrase and reduced intracellular drug accumulation. A single point mutation in gyrA has been shown to reduce susceptibility to fluoroquinolones. To determine the extent of gyrA mutations associated with FQ-R in enteric bacteria, one set of oligonucleotide primers was selected from conserved sequences in the flanking regions of the quinolone resistance-determining regions (QRDR) of Escherichia coli and Klebsiella pneumoniae. This set of primers was used to amplify and sequence the QRDRs from 8 Enterobacteriaceae type strains and 60 fluoroquinolone-resistant clinical isolates of Citrobacter freundii, Enterobacter aerogenes, Enterobacter cloacae, E. coli, K. pneumoniae, Klebsiella oxytoca, Providencia stuartii, and Serratia marcescens. Although similarity of the nucleotide sequences of seven species ranged from 80.8 to 93.3%, when compared with that of E. coli, the amino acid sequences of the gyrA QRDR were highly conserved. Conservative amino acid substitutions were detected in the QRDRs of the susceptible type strains of C. freundii, E. aerogenes, K. oxytoca (Ser-83 to Thr), and P. stuartii (Asp-87 to Glu). Strains with ciprofloxacin MICs of >2 microg/ml expressed amino acid substitutions primarily at the Gly-81, Ser-83, or Asp-87 position. Fluoroquinolone MICs varied significantly for strains exhibiting identical gyrA mutations, indicating that alterations outside gyrA contribute to resistance. The type and position of amino acid alterations also differed among these six genera. High-level FQ-R frequently was associated with single gyrA mutations in all species of Enterobacteriaceae in this study except E. coli.

- What are the cues for semantic annotation?
  - Label of new concepts, localisation in texts

Job positions