

# The Semantics of History

<http://www2.let.vu.nl/oz/ctl/semhis/index.html>

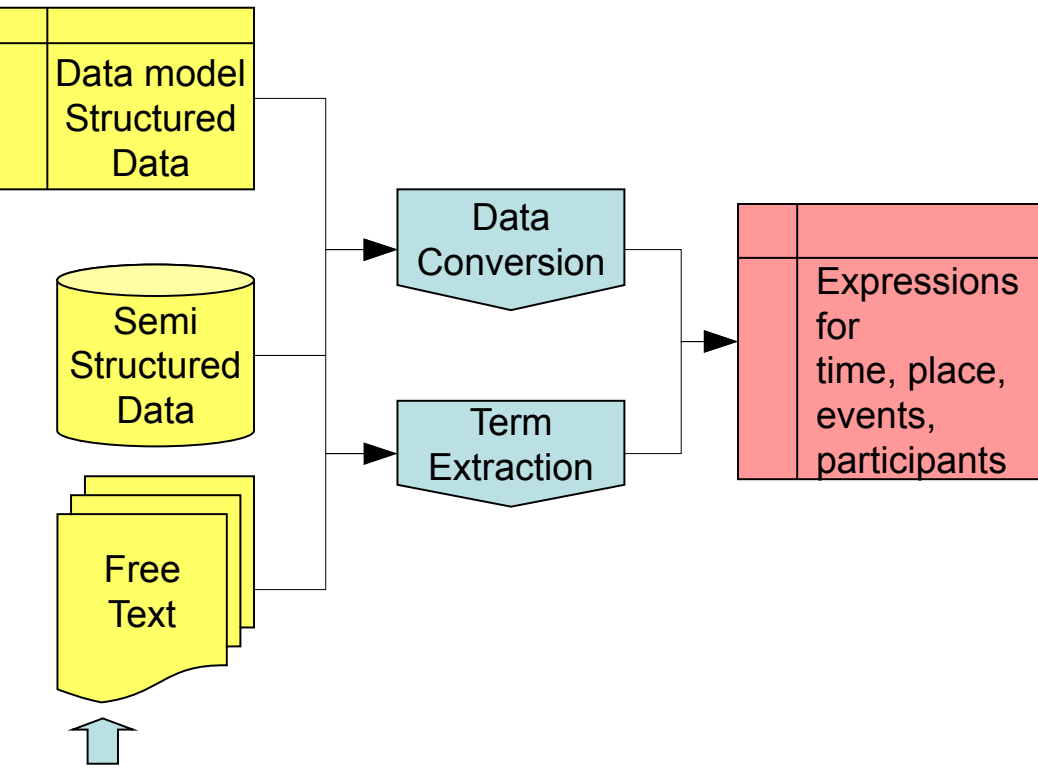
PHDs: Agata Cybulska and Roxane Segers  
Supervision: Piek Vossen and Guus Schreiber

CATCH meeting "Finding, Linking and Organizing  
Sources for Media Research", Rotterdam

October, 22nd, 2010

# The Semantics of History

- CAMeRA project at the VU University involving 1 PhD from the Faculty of Arts (Agata Cybulska) and 1 PhD from the Faculty of Exact Science (Roxane Segers)
- Goal: an ontology and lexicon for a mining events that can be related to the historical multimedia archive of the Rijksmuseum.
- Applied to an innovative information system for accessing the historical archive developed in the CATCH project Agora.

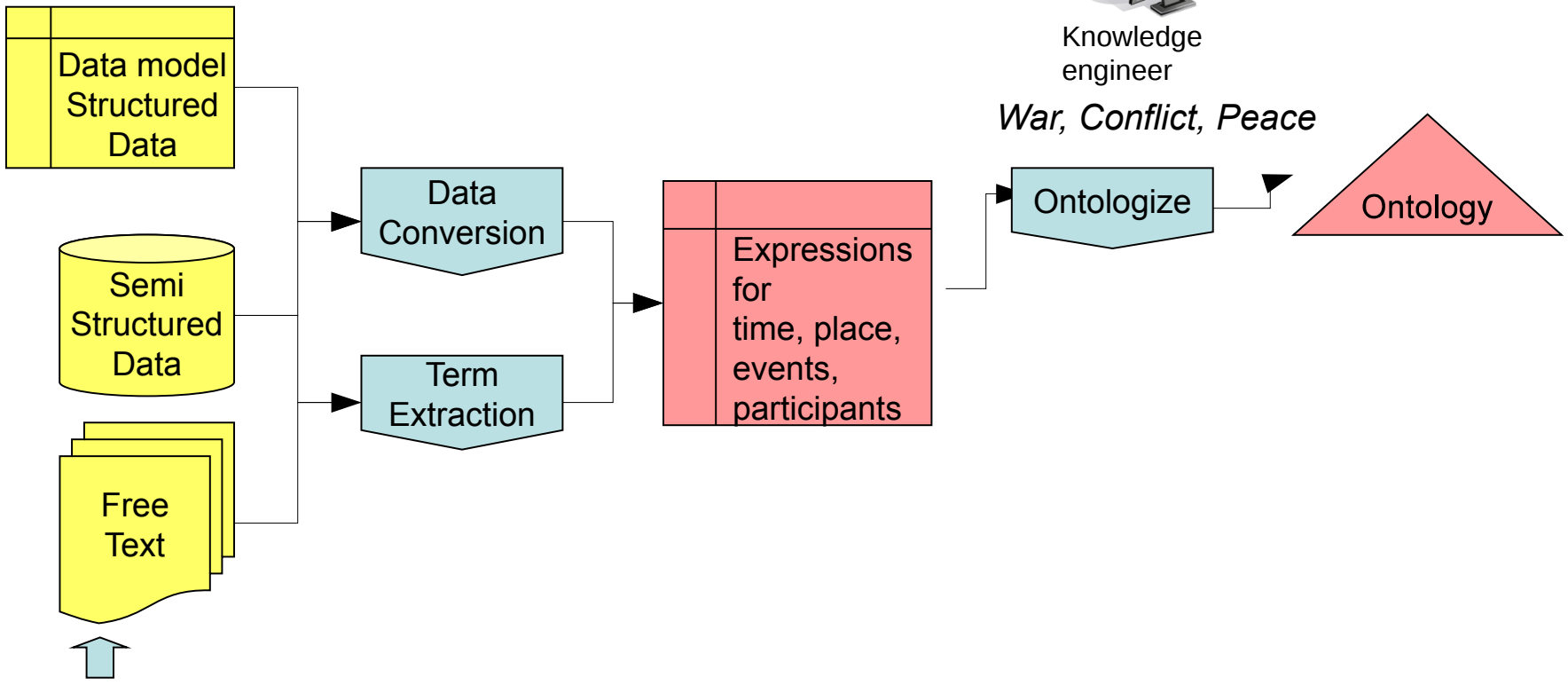


Rijksmuseum  
Wikipedia

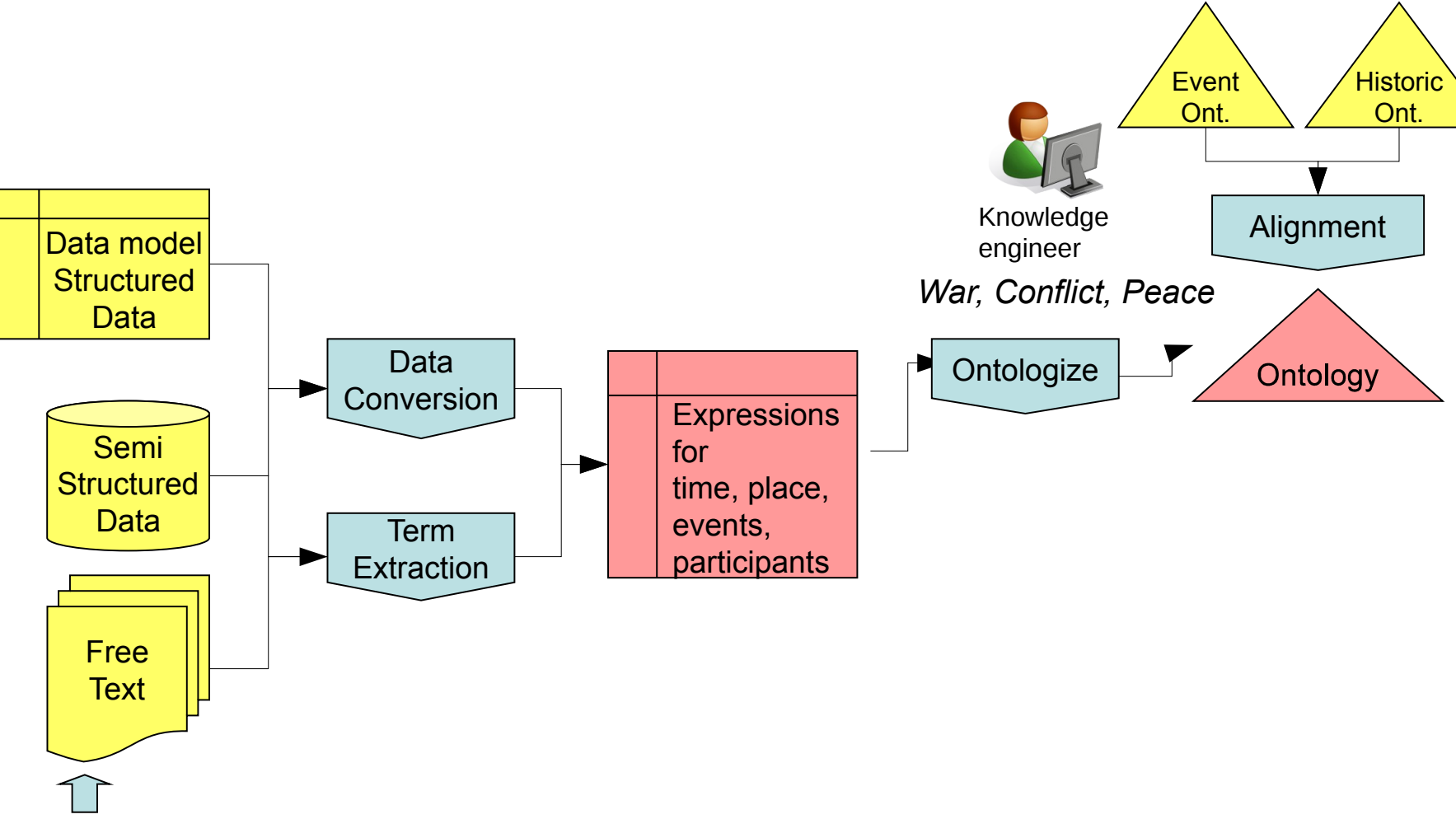


Knowledge engineer

*War, Conflict, Peace*



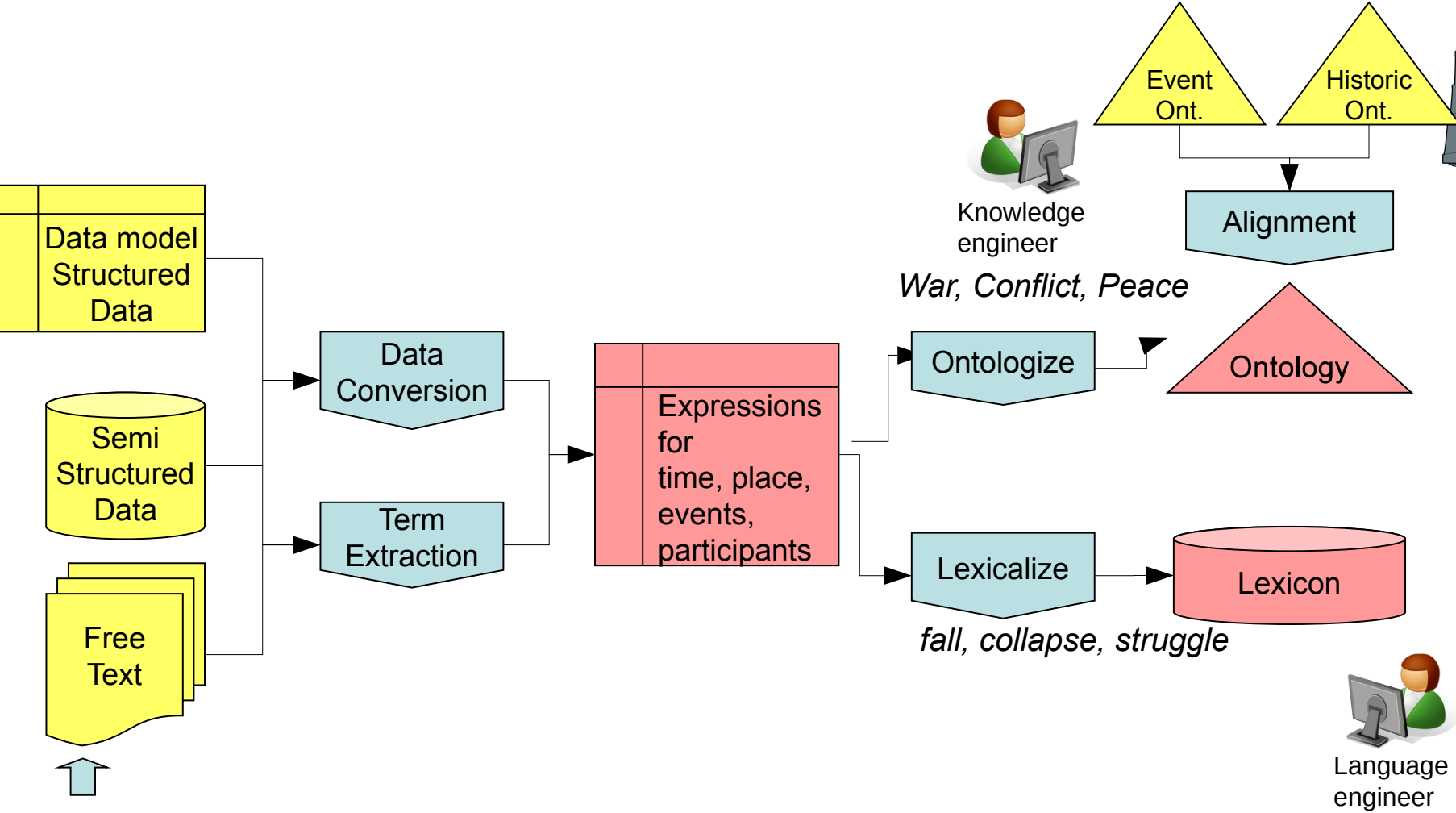
Rijksmuseum  
Wikipedia



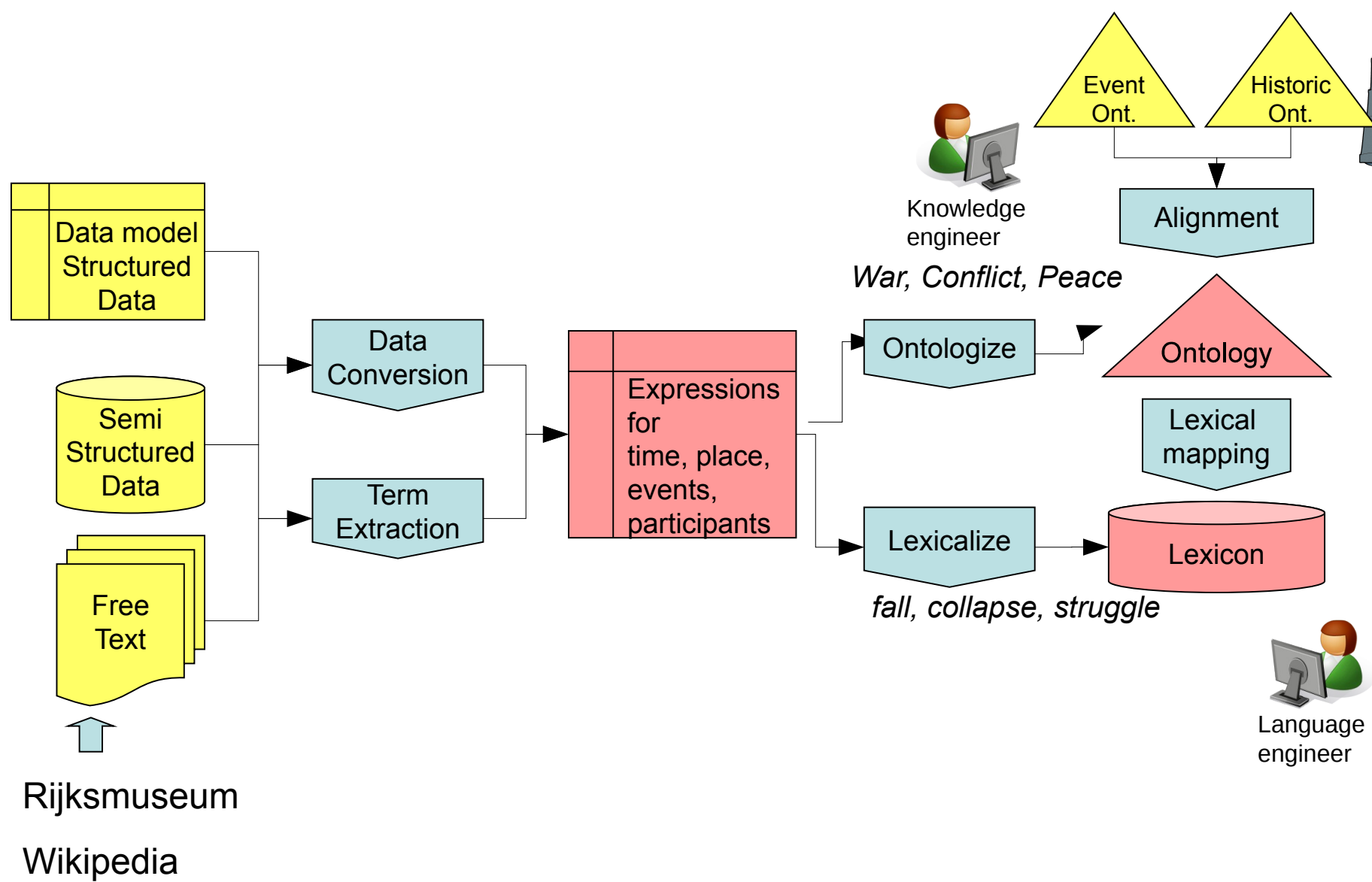
Rijksmuseum  
Wikipedia

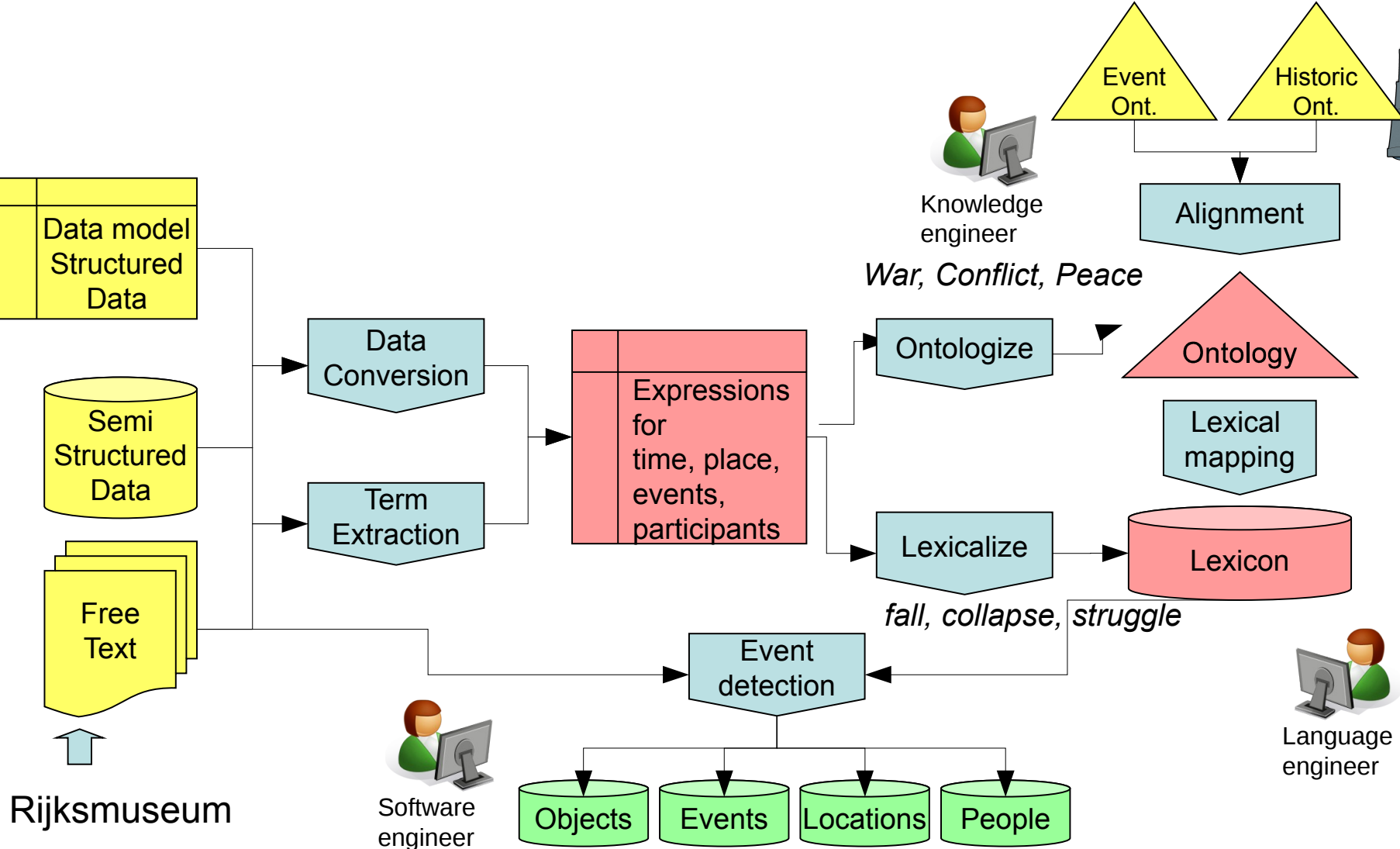
Agora

5



Rijksmuseum  
Wikipedia





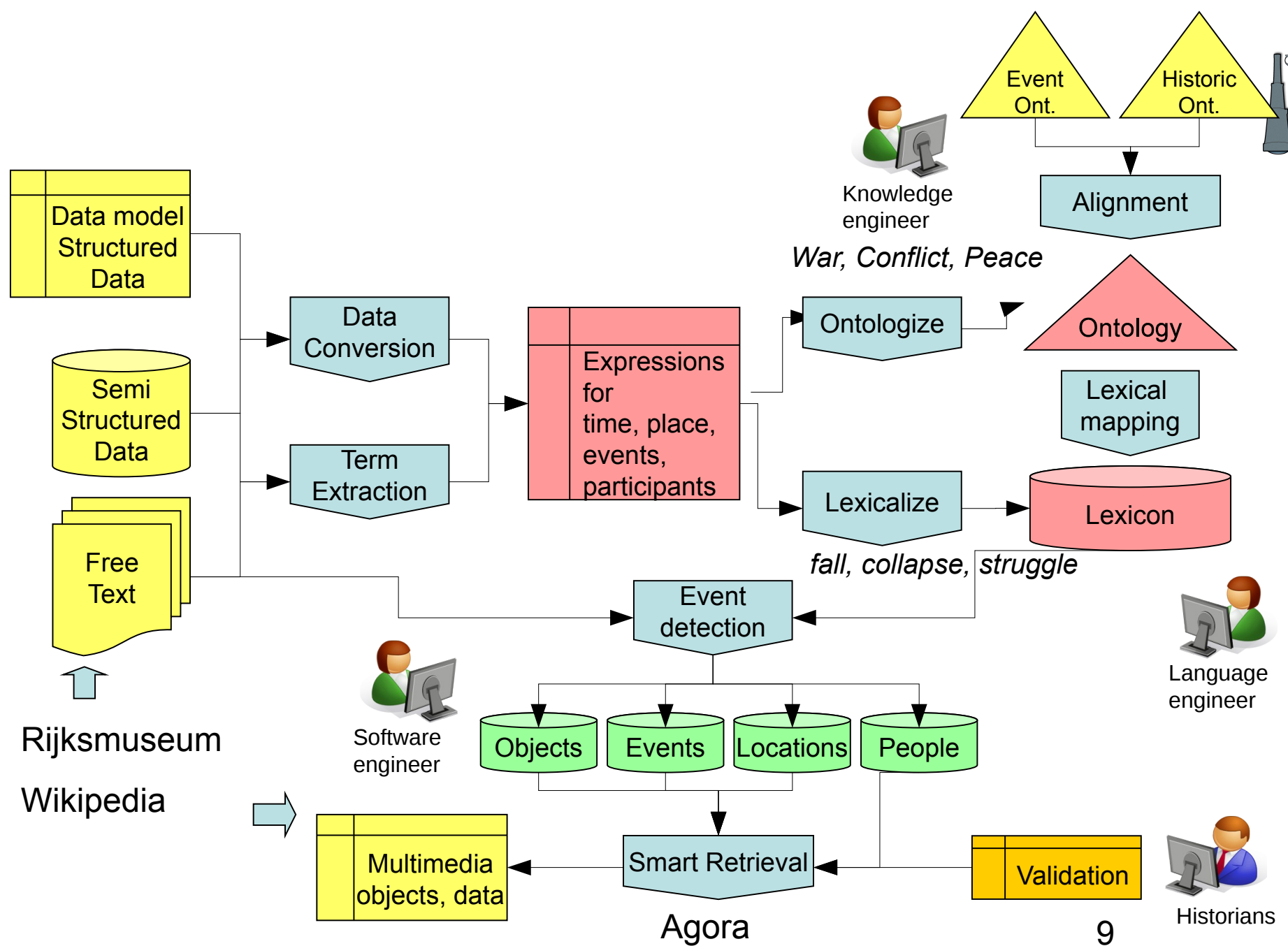
Rijksmuseum  
Wikipedia

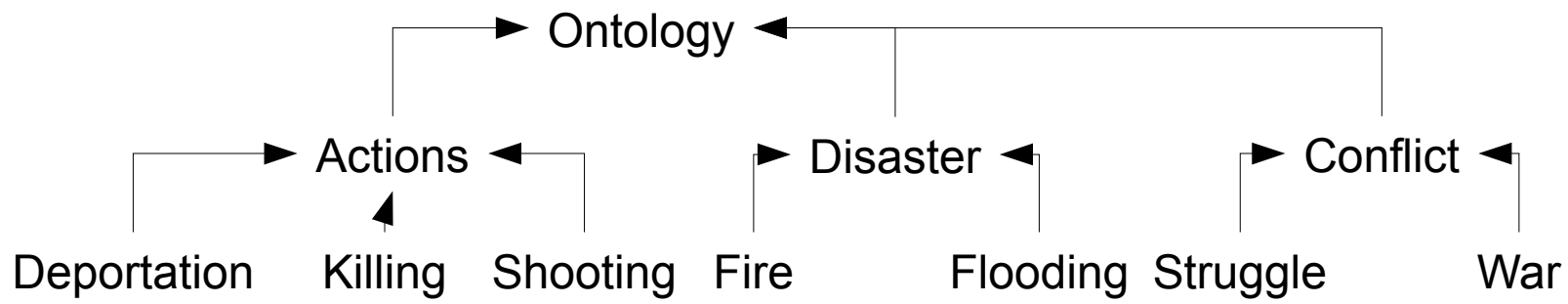
Software engineer

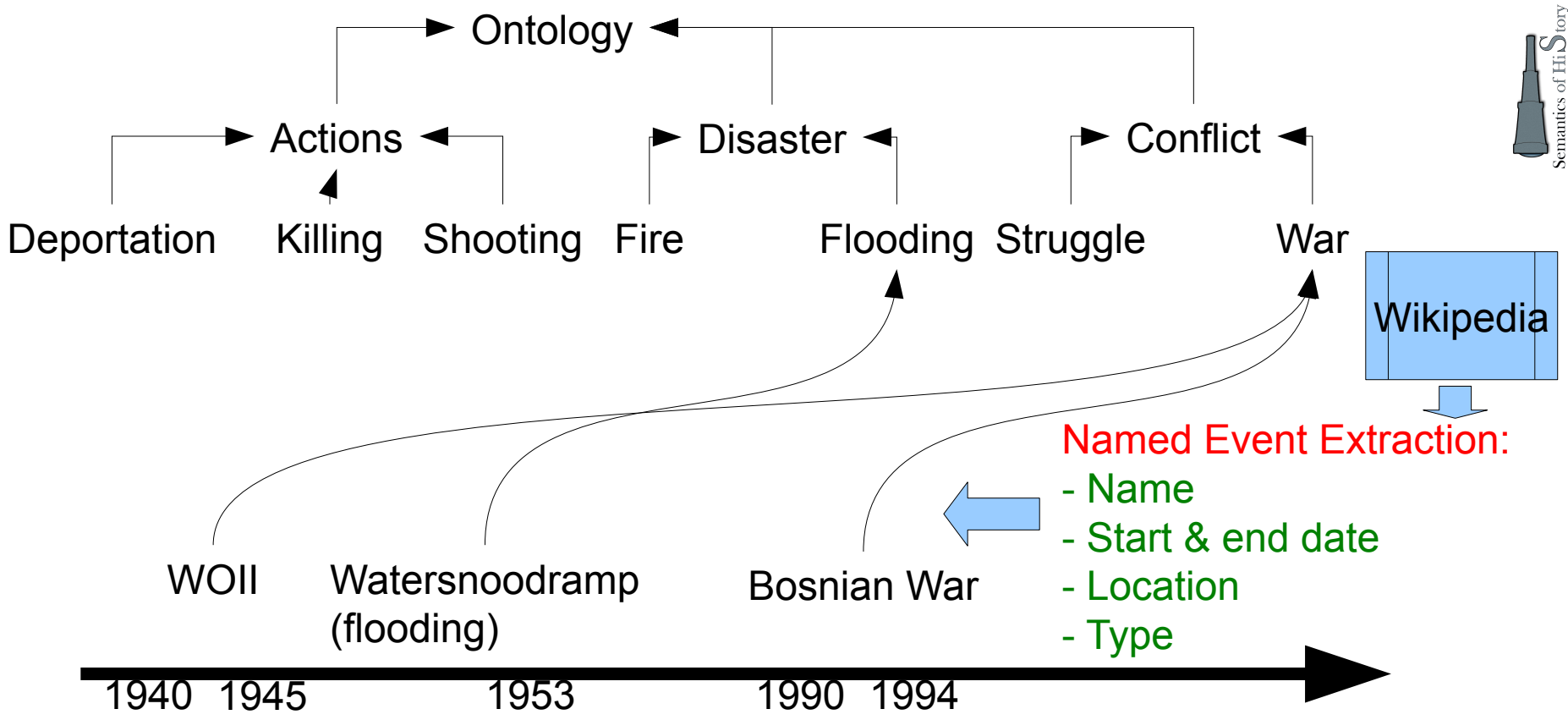
Language engineer

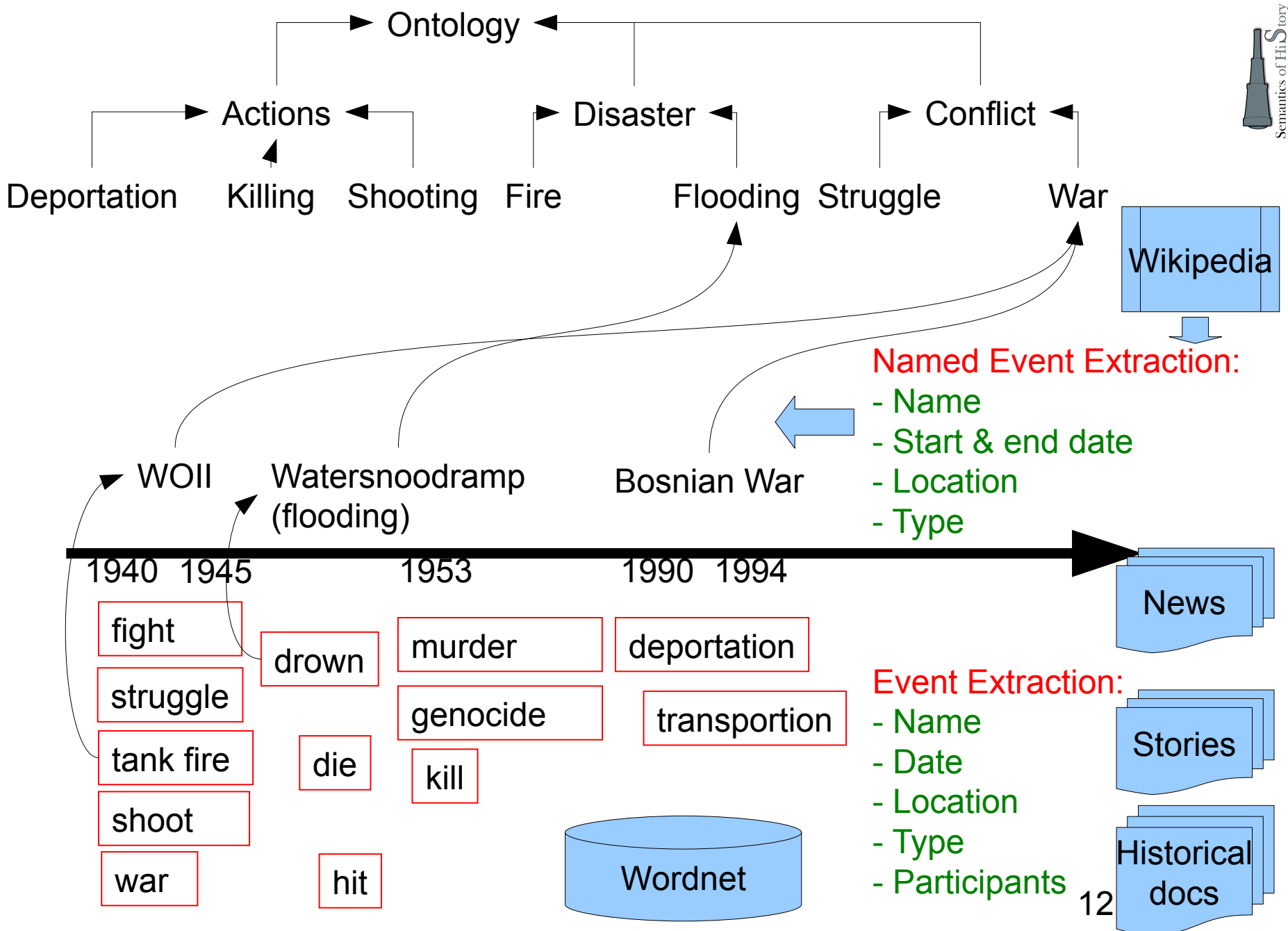
Agora

8









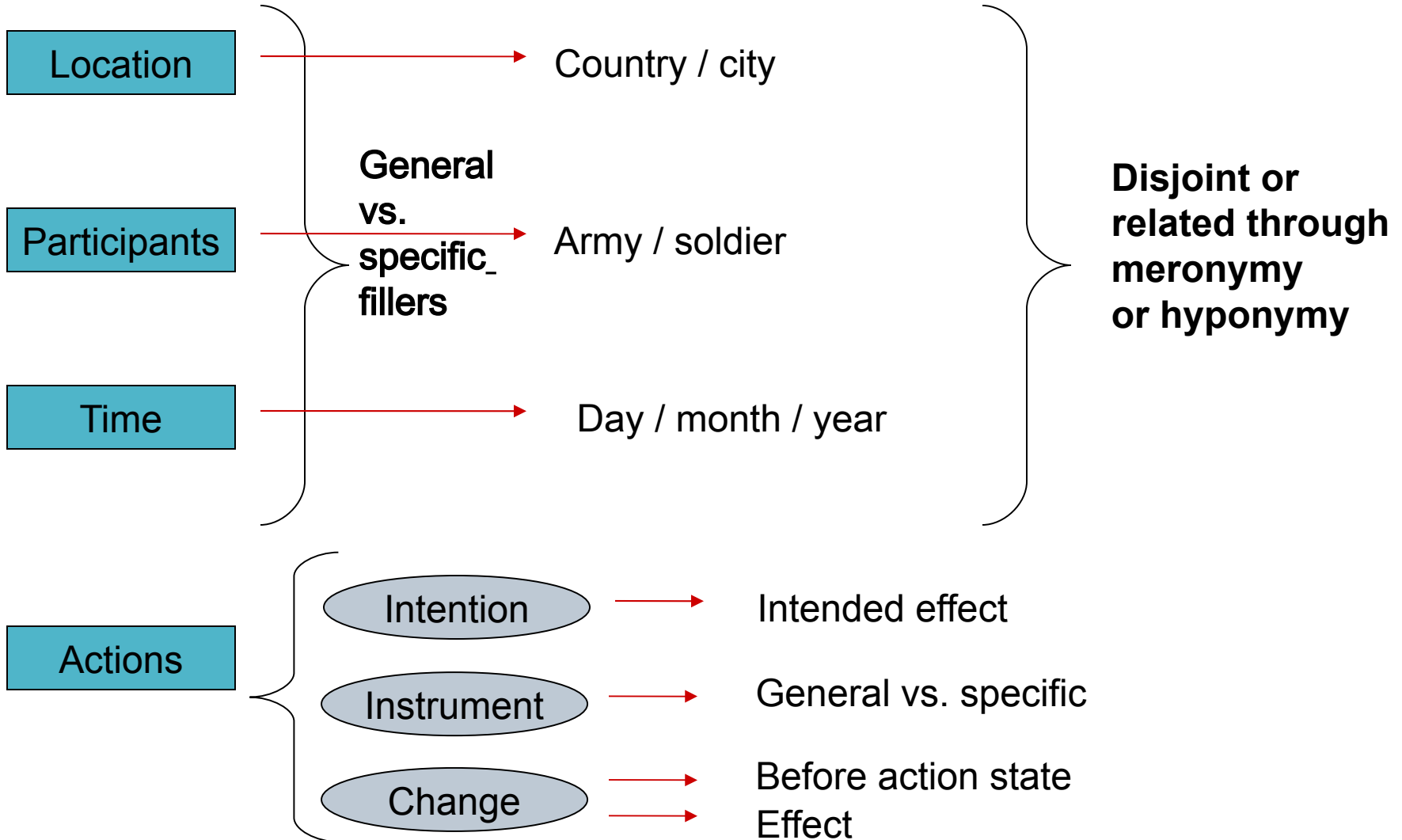
# Changing realities & perspectives

- Represent changing realities:
  - related as (*causal*) changes over time in regions
  - what remains stable and what changed
  - representing different views or perspectives on the same reality, e.g. from a different historical angle or from different geographical or social parties.
- Changes are typed as events in time and place and involving participants but not every event is also a historical event

# Events as key notions

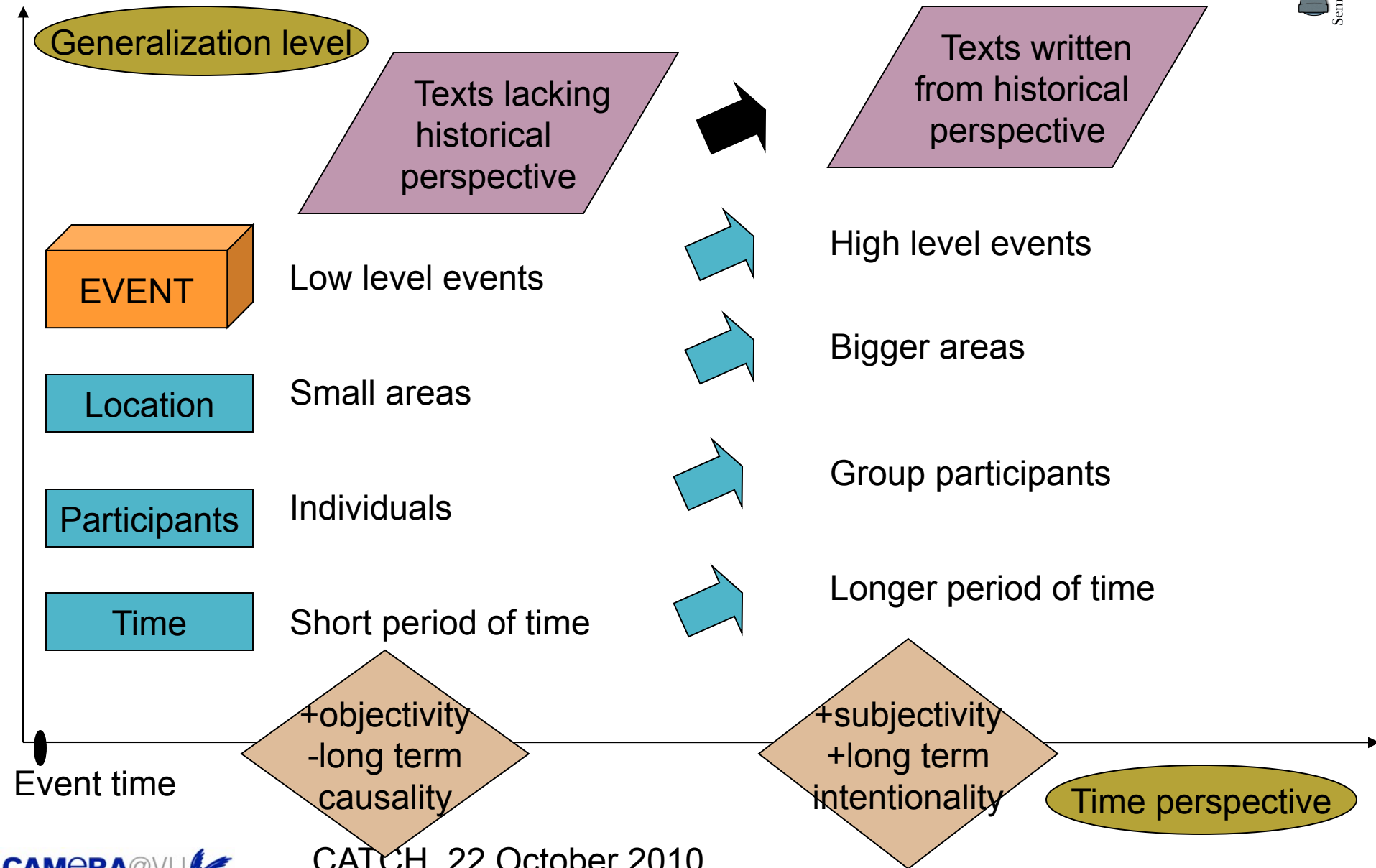
- Historical events:
  - events considered from a distance in time and abstraction of detail.
  - referenced by names (*WOII, Srebrenica Massacre*), nouns (*war*) or nominalizations (*violation of human rights*)
- News events:
  - Reports on (the same) reality but more in the active verbal form: *US soldiers shoot Iraqi citizens.*
  - Closer to the actual event
  - Lacking a historical abstraction and filtering.
- News becomes history over time, and we therefore expect a smooth transition in the use of language to refer to the same events, adding more and more historical perspective.

# Modeling Historical Events in Text



...

# Event Descriptions vs. Time Perspective





**Separation  
of men**



**Deportation  
of Muslim men**



**Srebrenica  
Massacre**



**Bosnian War**

Srebrenica sub -corpus		News <i>Volks- krant</i>	News <i>Het Parool</i>	Texts with hist. persp.
Frequency				
Time	Relative freq.types	0,045	0,029	0,026
	Relative freq.tokens	0,006	0,004	0,004
	Type-token ratio	1,1	1,1	2,1
	Unique types %	88,4 %	91,7 %	48 %
Loc	Relative freq. ypes	0,06	0,067	0,051
	Relative freq.tokens	0,007	0,008	0,004
	Type-token ratio	1,1	1,1	1,3
	Unique types %	89,5 %	92,6 %	78 %
Actor	Relative freq.types	0,249	0,253	0,169
	Relative freq.tokens	0,031	0,031	0,02
	Type-token ratio	1,1	1,1	1,8
	Unique types %	89,4 %	92 %	56,7 %

- Different historical perspectives of writers correspond with genre distinction and correlate with variation in language use.
- Texts, written shortly after an event happened, use more specific and uniquely occurring event descriptions than texts describing the same events but written from a longer time perspective.

# (Pattern-based) event extraction for named events

Three approaches:

I. Search for event *external* patterns, and apply these to find new events in text: *tijdens de X, gedurende de X, herdenking van de X (during the X, memorial of the X)*

II. Search for event *internal* patterns, and apply these to find new events:

1. Beleg (siege) van Antwerpen -> Beleg van Y [loc] -> Beleg van Leuven

2. Palingoproer -> X + oproer (insurrection)-> Bouwvakoproer

III. Search for new types that are used in event names: use a hierarchical organized lexical database (wordnet) to find possible candidates around the already found types. E.g. aanslag (attack) -> bomaanslag (bomb attack).

# Results: top eight

Pattern	Freq	Total over sample	percentage
verslagen in de X (defeated)	242	163/242	67%
herdenking van X (memorial)	193	90/193	46.6%
nam deel aan de X (took part)	892	148/350	42%
tijdens de X (during)	49532	93/350	26.5%
ten tijde van X (during)	4719	72/350	20.5%
naar aanleiding van X (because of)	1442	42/350	12%
einde van de X (end of)	14407	25/350	7.1%
begin van de X (start of)	11489	20/350	5.7%

# Conclusions

- We will develop an ontology that represents a basic classification of historical events
- We will develop a Dutch lexicon that relates every expression to the historical ontology
- Ontology and lexicon are used to detect events in text, including time, place and participants
- We will develop machinery to match events expressions that are connected by time and place:
  - What expressions are equivalent;
  - What expressions are different perspectives
  - What expressions have sub-event relations
- Indexes of events will be used in the Agora project

Thank you for your attention