

# Automatic Topic Map Generation from Free Text using Linguistic Templates

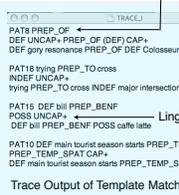
Ann Houston      Grammarsmith

## MOTIVATION

Free text contains vast amounts of useful information  
Information extraction from free text is not easy  
Complexity of natural language  
Complexity of NLP technologies  
Mapping is from 'messy' NL to well-defined TAOs



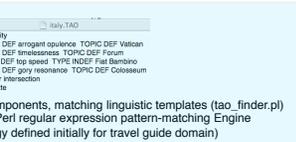
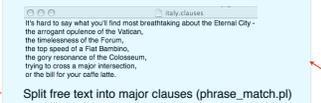
## Template ID and unmatched portion of phrase



Tracing captures portions of phrases that are not consumed by the template matching. These are important for debugging current templates and also determining what new templates should be added. Could add 'Remaining Phrase' as Topic Type and inspect these via Omnigator as well.

## HYPOTHESIS

Can a set of linguistic templates extract TAOs automatically to form a useful topic map base from free text?  
One pass pattern matching, no lexicon, no parse trees



Rapid iterative cycle is possible through use of advanced visualization tool for Topic Maps, Ontopia's Omnigator

## APPROACH



Generate xtm code and explore output in Omnigator



**Topic Map Overview**

- Ontology
- Master Index
- Index of Individuals
- Index of Themes
- Unnamed Topics

**Topic Types (4)**

- Descriptor
- General Descriptor
- Generic Entity
- Point of Interest

**Association Role Types (2)**

- General Characteristic
- Generic Entity

**Occurrences Types (1)**

- Descriptor

## Point of Interest

**Untyped Names (1)**

- Point of Interest

**Scoped Names (2)**

- Features (Sight Description)
- Features (Specific Agentive Association)

## Subject Identifiers (1)

file:/Users/annhouston/Desktop/tomcat/webapps/omnigator/W

## Topics of this Type (64)

- Alghieri
- Astoria
- August
- Boboli
- Cave
- Christmas
- Close
- December
- Early
- Easter
- Europe
- Excell
- Excellent
- Florence
- Gaddi
- Giambologna
- Giotto
- Italian
- Italy
- July
- ...

	Recall		Precision	
	POI	GE	POI	GE
Italy	64/84	59/77	55/64	40/59
	0.76	0.77	0.86	0.68
Libya	116/162	104/150	101/115	68/104
	0.72	0.69	0.88	0.65
Yellowstone	17/29	39/55	16/17	28/39
	0.59	0.71	0.94	0.72
Chichen Itza	17/24	14/15	15/17	9/14
	0.71	0.93	0.88	0.64

Preliminary Results on Point of Interest (POI) and Generic Entity (GE) extraction from 4 free text sources

## Unidentified Entity

**Unidentified Entity**

- Unknown: 630 BC

## Conclusions

A simple set of linearly-applied 'shallow orthographic/functional' templates can extract TAOs. Noun phrases contain a lot of the basic information, templates for NPs easier to formulate than for VPs. Linguistic templates are language specific, and to some extent, domain specific. Approach is fast, iterative and utilizes robust existing visualization tool - Ontopia's Omnigator. Approach is a fast bootstrap to create topic maps from minimal ontology, human editing can then build upon it. Tracing tools are very important in refining the developing set of linguistic templates.

## Issues for Further Research

- Ordering of Templates - implications for which templates will match phrases in which sequences
- Coreference - recognize when 'the city' = 'Florence' (a difficult problem to solve automatically)
- Conjunction Scope - distinguish 'expensive shops and restaurants' versus 'expensive shops and interesting restaurants'
- Verbs - rich source of associations - probably need some lexicon to extract these, not just templates

**Unidentified Entity**

- Unknown: 630 BC

**Subject Identifiers (1)**

- file:/Users/annhouston/Desktop/oke-professional-3.3.0/apache-tomcat/webapps/omnigator/WEB-INF/topicmaps/Guide\_Libya.xtm#unidentified\_

**Topics of this Type (1)**

- Unknown: 630 BC

**Generic Entity**

- Has (General Characteristic)

**Unidentified Entity**

- Unidentified Entity

**Subject Identifiers (1)**

- file:/Users/annhouston/Desktop/oke-professional-3.3.0/apache-tomcat/webapps/omnigator/WEB-INF/topicmaps/Guide\_Italy.xtm#descriptor

**Topics of this Type (9)**

- built
- overwhelming
- situated
- the arrogant opulence
- the gory resonance
- the heart
- the timelessness (the Forum)
- thought (Boboli)
- usually mild

**Occurrences of this Type (9)**

- built (Astoria)
- overwhelming (Florence)
- situated (Medici)
- the arrogant opulence (the Vatican)
- the gory resonance (the Colosseum)
- the heart (the Renaissance)
- the timelessness (the Forum)
- thought (Boboli)
- usually mild (Winters)

**Internal occurrences identified by matched templates (Descriptors)**

## Need new template to match numeric dates

## Subject Identifiers (1)

file:/Users/annhouston/Desktop/oke-professional-3.3.0/apache-tomcat/webapps/omnigator/WEB-INF/topicmaps/Guide\_Italy.xtm#generic\_entity

## Topics of this Type (59)

- altering the original architecture PREP\_TEMP\_SPAT a anarchical suggestive way Together PREP\_INSTR Pre Palace.
- autumn
- citrus fruit trees
- concerts
- cranmy
- culture
- economic horizon
- external terraces extending towards the river
- fir
- literature
- luty domes
- mountains
- pleasant
- sunny days
- the citys main museums
- then head PREP\_SPAT again PREP\_TEMP\_SPAT 1800 PREP\_TO take advantage of the cooler evening
- wooden planks
- BE NUM of the citys best known images
- BE indeed picturesque
- COMP means very light traffic

Template Mismatches of Generic Entities