

Incremental development of a  
shared urban ontology  
The Urbamet Experience

Jacques Teller  
Université de Liège

Jacques Guyot, Gilles Falquet,  
Université de Genève

Urbamet

Thesauruses and other knowledge resources

An analysis of Urbamet with an automatic classification tool

Methodology for the evolution of thesauruses

# An Analysis of the Urbamet Thesaurus

URBAMET

Thesaurus for indexing bibliographic notes  
produced by the French Centre for Urban  
Documentation.

In 1969:

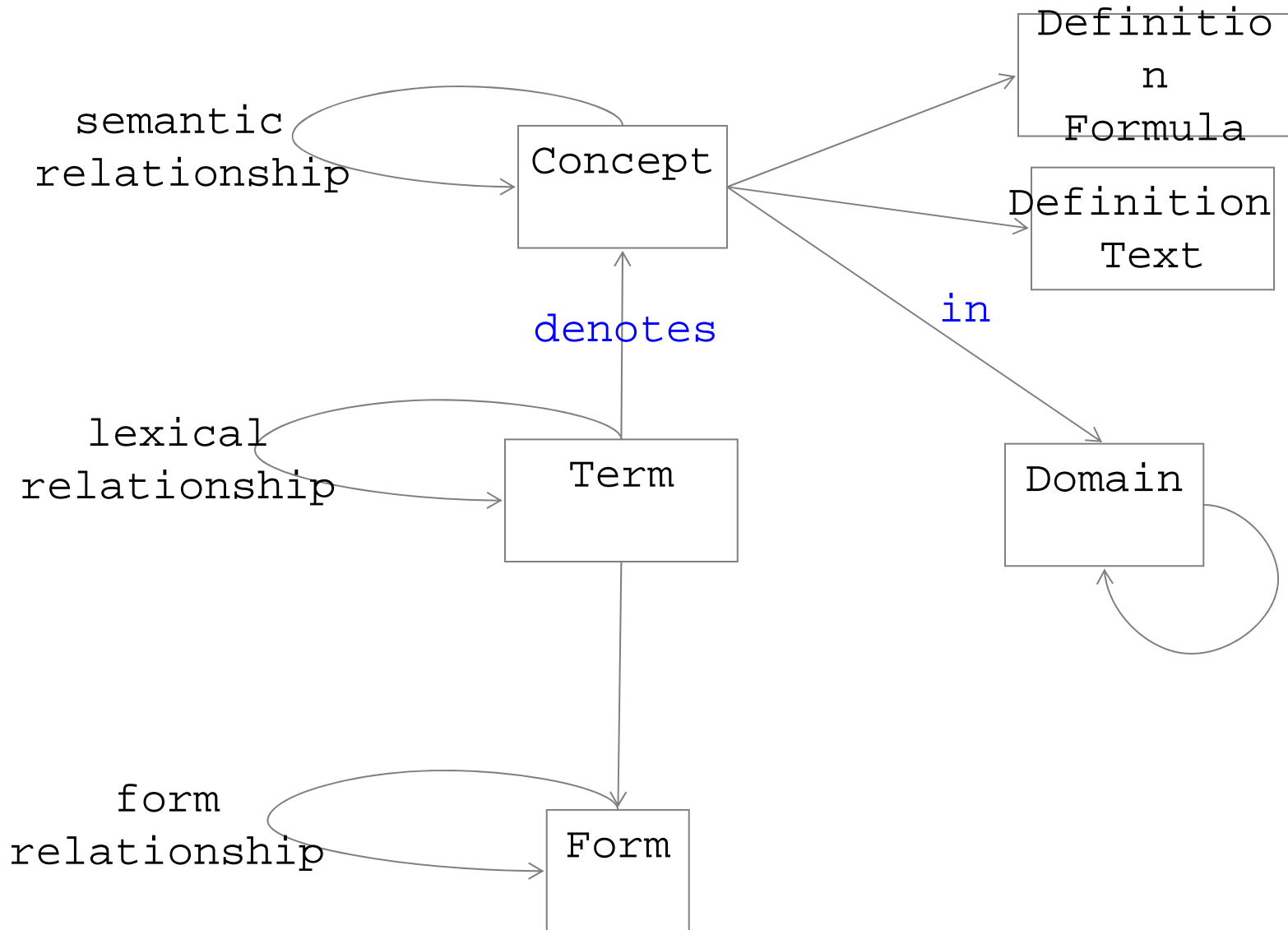
2300 terms

Currently:

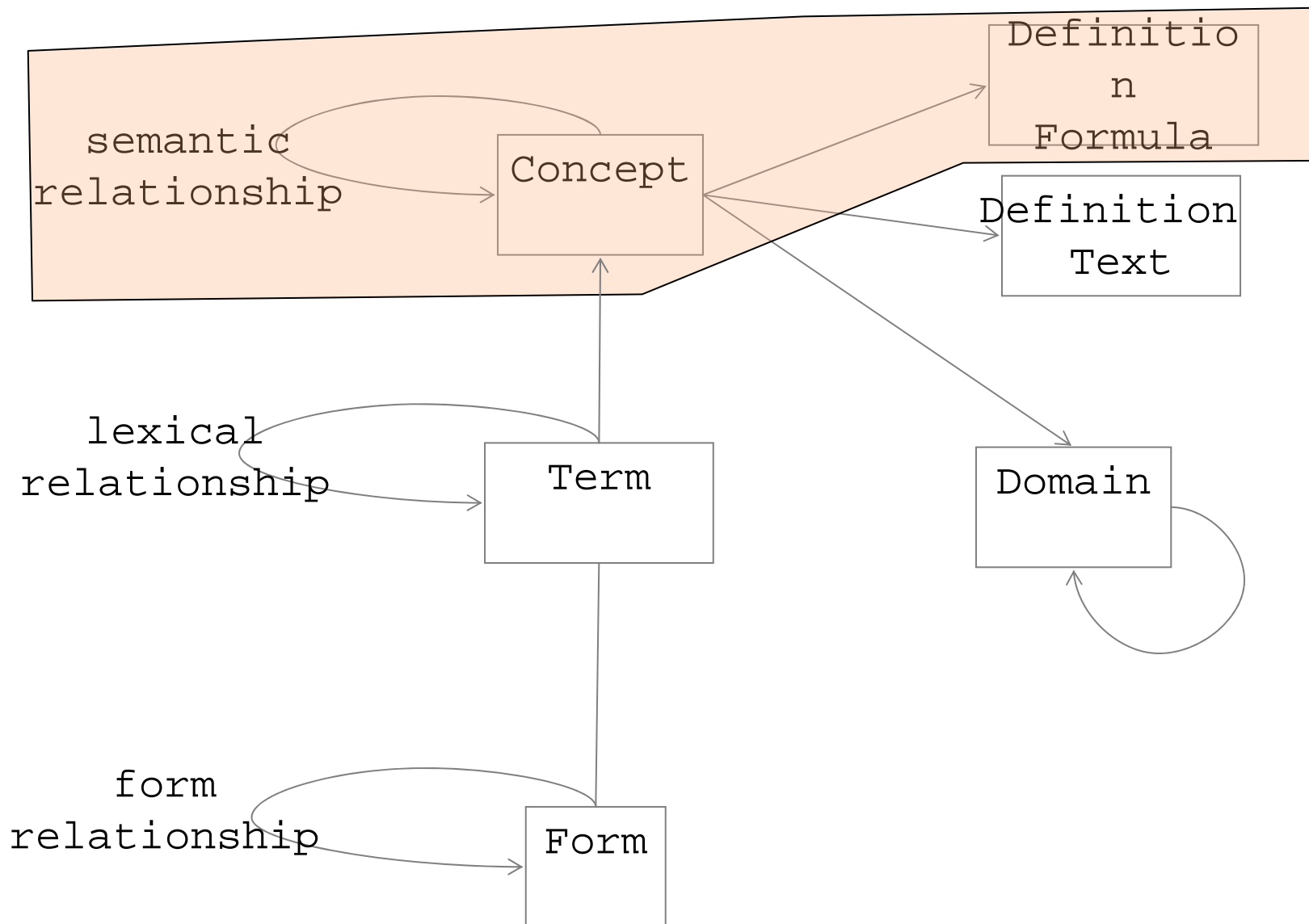
4200 terms

230'000 documents

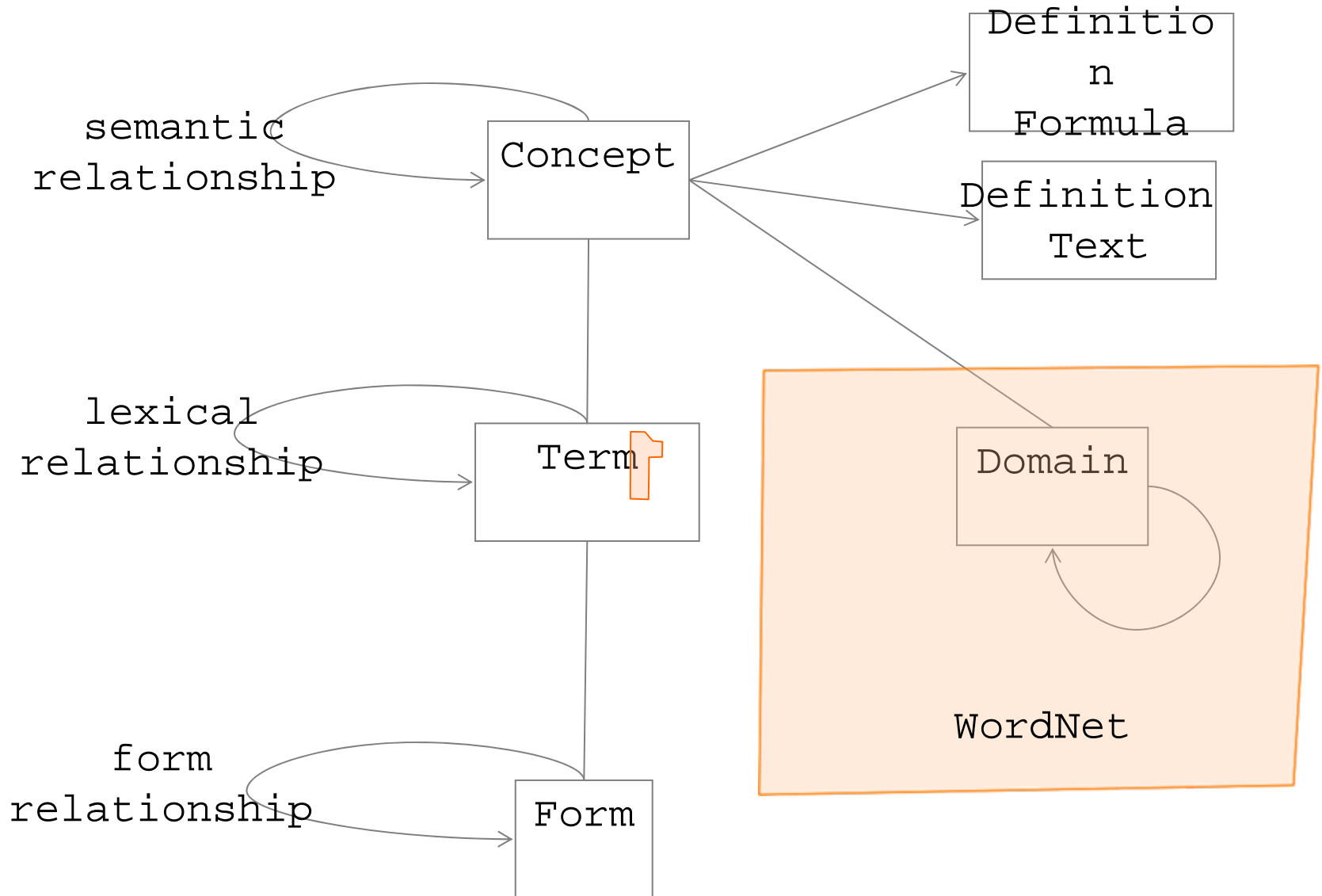
# Ontologies and Linguistic Resources



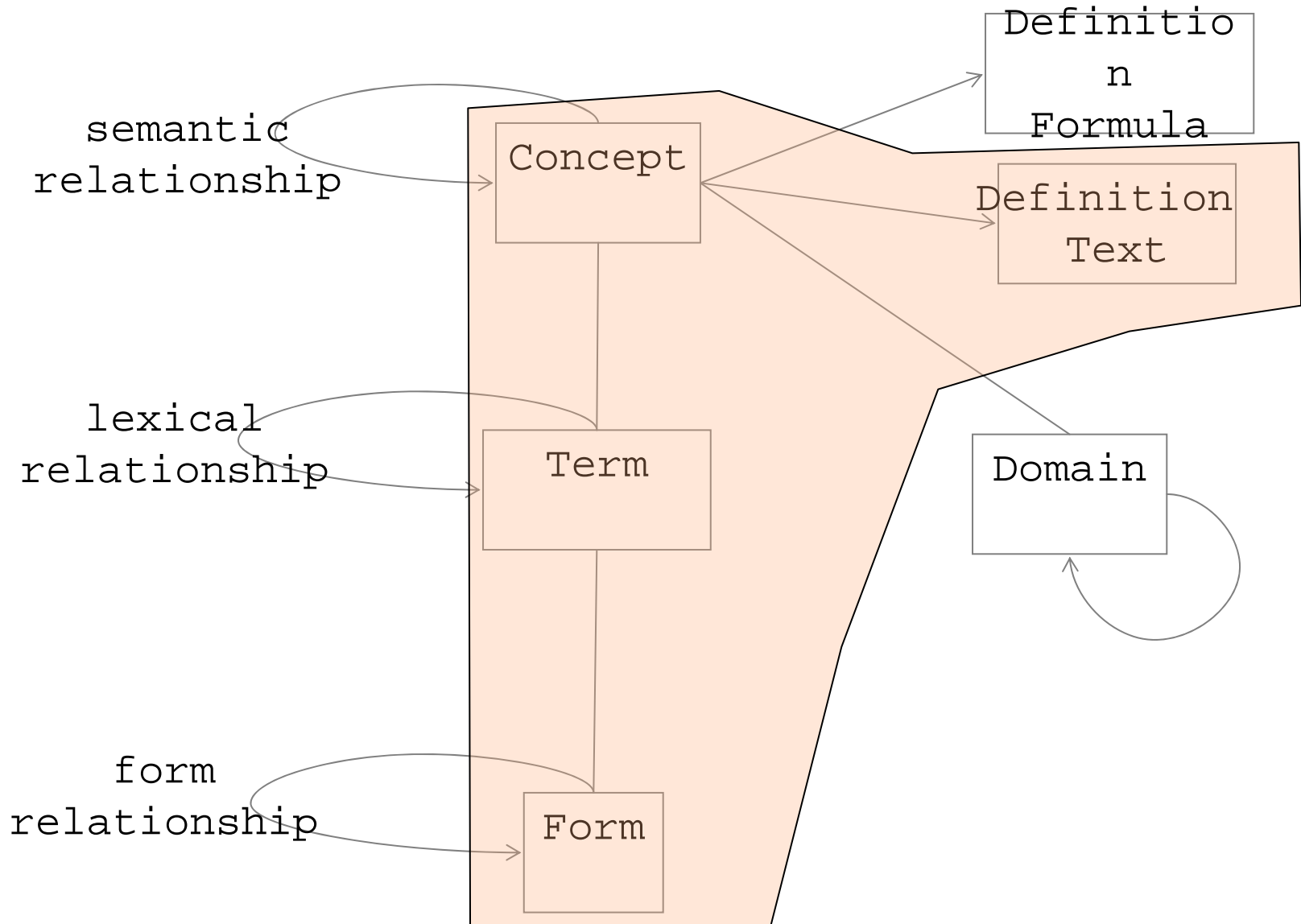
# Purely Formal Ontology



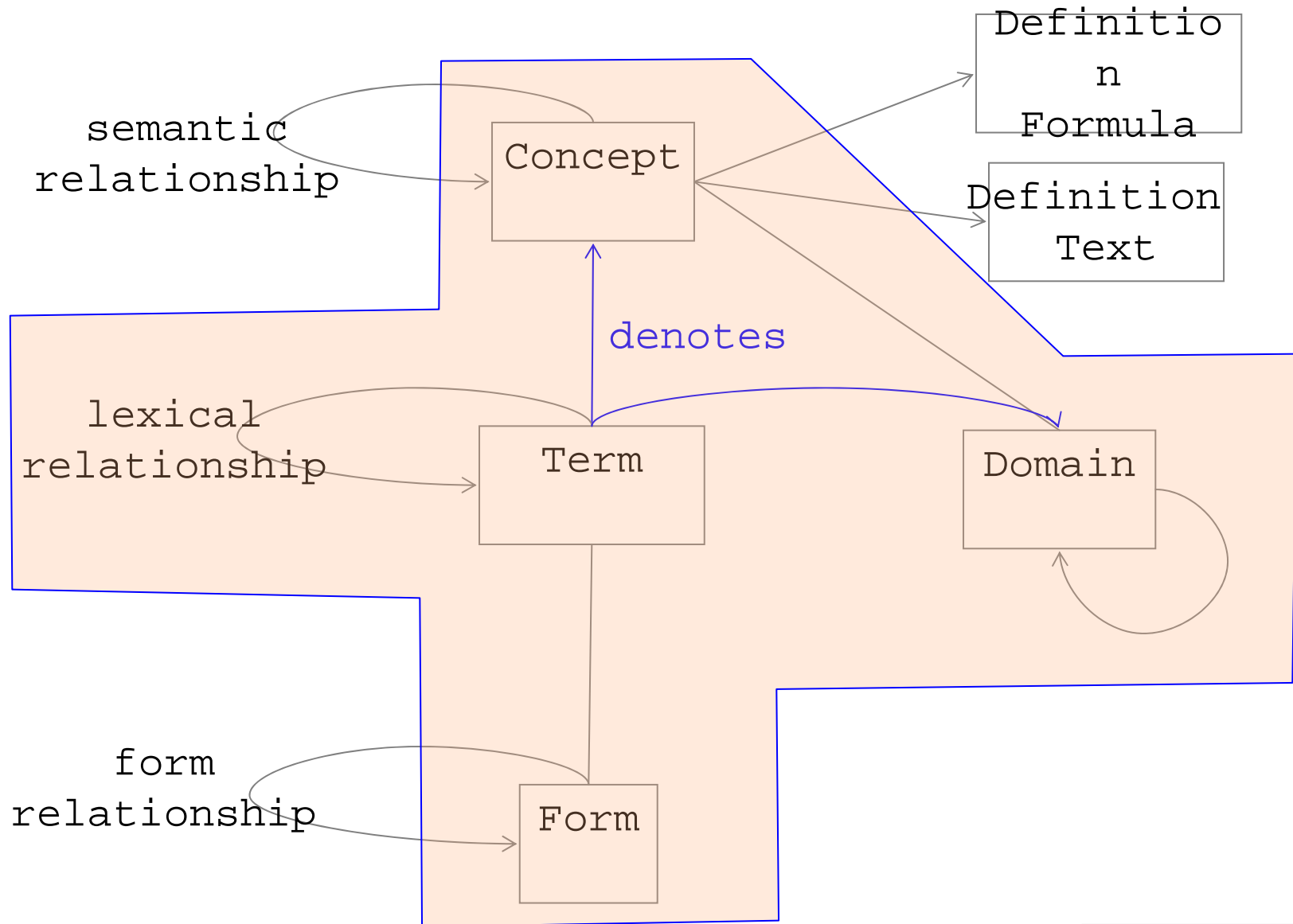
# Lexical Ontology



# Glossary



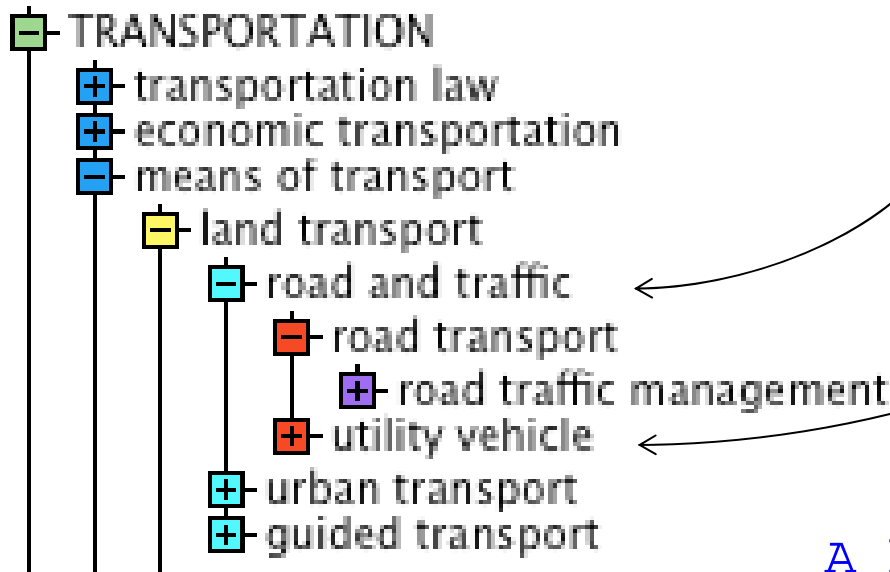
# Thesaurus



# Domains and Concepts in Thesauruses

Thesauruses are mostly used to classify documents

Terms in thesauruses denote (sub-)domains and/or concepts



A hierarchy of sub-domains

Not a hierarchy of concepts

Analyze the Urbamet thesaurus *together with the indexed documents (corpus)*.

Analysis steps

1. Corpus extraction
2. Training set creation
3. Classifier training (machine learning tech.)
4. Confusion matrix analysis
5. TOP50 words analysis

# Corpus extraction

Approx. 10'000 abstracts

70 indexed words / document

Vocabulary: 18'000 words (stems)

Document < 1 > /9878

 [Affichage HTML pour impression](#) [Affichage PDF pour impression](#) [Retour à la liste de résultats](#)

<b>Titre</b>	<b>Agora 2020 : synthèse à mi-parcours.-</b>
<b>Auteur(s)</b>	<a href="#">BAIN, Pascal</a> ; <a href="#">CHAPUY, Pierre</a> ; <a href="#">DROUET, Dominique</a> ; <a href="#">FARHI, François</a> ; <a href="#">MAUJEAN, Sébastien</a> ; <a href="#">MIRENOWICZ, Philippe</a> ; <a href="#">THEYS, Jacques</a>
<b>Organisme(s) auteur(s)</b>	<a href="#">CENTRALE MANAGEMENT INTERNATIONAL. PARIS</a> ; <a href="#">FRANCE. MINISTERE DE L'EQUIPEMENT. DIRECTION DE LA RECHERCHE ET DE L'ANIMATION SCIENTIFIQUE ET TECHNIQUE</a> ; <a href="#">FRANCE. MINISTERE DE L'EQUIPEMENT. DIRECTION DE LA RECHERCHE ET DES AFFAIRES SCIENTIFIQUES ET TECHNIQUES. CENTRE DE PROSPECTIVE ET DE VEILLE SCIENTIFIQUES ET TECHNOLOGIQUES</a> ; <a href="#">GROUPE D'ETUDES RESSOURCES PLANIFICATION AMENAGEMENT. PARIS</a> ; <a href="#">RECHERCHE DEVELOPPEMENT INTERNATIONAL. PARIS</a>
<b>Source bibliographique</b>	<i>Paris : DRAST, CPVS, fév. 2005.- 150 p., ann., tabl., graph.</i>
<b>Collection</b>	Dossiers du CPVS, n° 8 - février 2005
<b>Notes</b>	Coll. Les Dossiers du CPVS n°8. Coll. Dossiers du CPVS
<b>Cote</b>	<b>CDU 58938 ; RST RCPVS05-001</b>
<b>Résumé</b>	A l'été 2003, la DRAST a engagé une vaste consultation prospective sur les attentes en matière de recherche dans ses différents domaines d'intervention - en y incluant le logement, la ville, la gestion des risques et l'observation des milieux. Cette consultation, intitulée Agora 2020, entre dans sa phase finale. Le dossier constitue une première synthèse des travaux réalisés en 2003 et 2004. Il est divisé en deux grandes parties : une synthèse à mi-parcours qui reprend sous forme de " messages clefs " les deux premières phases d'Agora 2020 ; un dossier complémentaire rassemblant quelques comptes-rendus représentatifs de l'ensemble des travaux déjà réalisés (présentation d'Agora 2020, les attentes des acteurs, observation de la terre des milieux et gestion des risques, des visions du monde qui ne communiquent pas : une illustration à travers le thème de la ville).
<b>Lieu(x)</b>	<a href="#">France</a>
<b>Thème(s)</b>	<a href="#">Aménagement urbain</a> ; <a href="#">Habitat - Logement</a> ; <a href="#">Infrastructures - Ouvrages d'art</a> ; <a href="#">Sciences de la terre</a> ; <a href="#">Transports</a>
<b>Mot(s)-clé(s) français</b>	<a href="#">aménagement du territoire</a> ; <a href="#">besoin</a> ; <a href="#">habitat - logement</a> ; <a href="#">infrastructures - ouvrages d'art</a> ; <a href="#">ministère équipement logement</a> ; <a href="#">prospective</a> ; <a href="#">recherche</a> ; <a href="#">transports</a> ; <a href="#">ville</a>
<b>Mot(s)-clés anglais</b>	<a href="#">forecast</a> ; <a href="#">housing</a> ; <a href="#">infrastructures - engineering works</a> ; <a href="#">inter-regional planning</a> ; <a href="#">need</a> ; <a href="#">public works housing ministry</a> ; <a href="#">research</a> ; <a href="#">town</a> ; <a href="#">transportation</a>

# Main Themes (Domains)

- + PUBLIC ADMINISTRATION
- + ARCHITECTURE
- + GENERAL TRAFFIC
- + TERRITORIAL COMMUNITIES
- + CONSTRUCTION
- + ECONOMY
- + JOB - OCCUPATION - EDUCATION
- + ENVIRONMENT - LANDSCAPE
- + PUBLIC FACILITIES
- + LAND PROPERTY
- + EARTH SCIENCES
- + HOUSING
- + SOCIAL SCIENCES
- + INFORMATION - DOCUMENTATION - COMMUNICATION
- + INFRASTRUCTURES - ENGINEERING WORKS
- + LEGAL FRAMEWORK
- + METHODS - TECHNIQUES
- + INTER-REGIONAL PLANNING
- + RESSOURCES
- + COUNTRY PLANNING
- + HEALTH
- + HOLIDAY TOURISM - LEISURE
- + TRANSPORTATION
- URBAN PLANNING DEVELOPMENT

# Training file and learning

Build a training file with records of the form

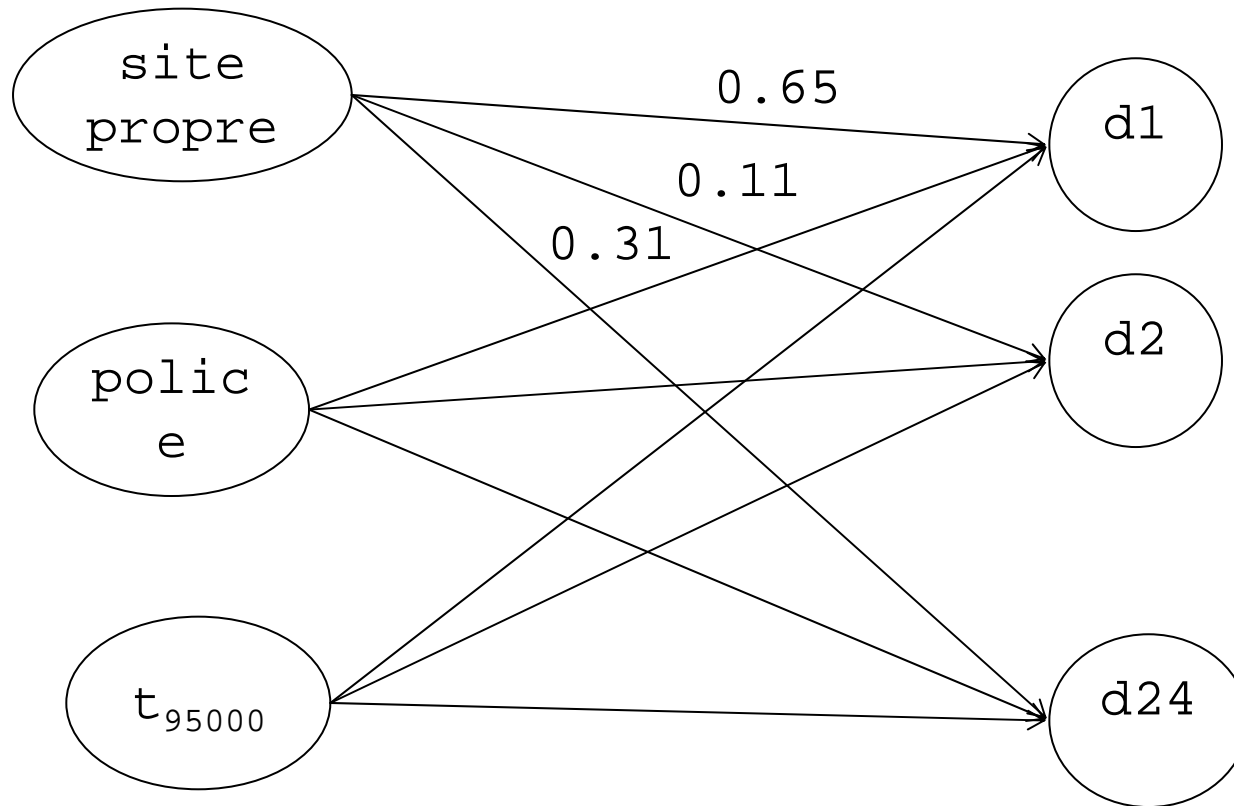
```
<doc name> <dom1><dom2> ... <domk>
```

Training phase:

The classifier builds a neural network by reading the training file and applying the Winnow learning technique.

# Neural network

Weighted arcs from a word or pair of words to a domain.



Weight of term  $i$  for domain  $j$  = how strongly  $i$  draws to  $j$

# Performance of the generated classifier

Training with 80% of the corpus. Test with 20%

The classifier discovers the main domain of each tested document with probability

59% for the first proposed domain

16% for 2<sup>nd</sup> choice

7 % for 3<sup>rd</sup> choice

82% in first 3 proposals (random choices = 23%)

The classifier is effective => **The Urbamet classification corresponds to the text contents**

# Confusion matrix analysis

Objective: find domains which are poorly classified.

Complete in-out 24x24 matrix

$M_{ij}$  = percent of document in domain  $i$  classified in  $j$

Ideally  $M_{ii} = 100\%$

What are the exceptions ?

# Transportation - Traffic confusion

In \ out	Transportation	Traffic	Tourism	...
Transport	45%	24%	3%	
Circulation	10%	40%	1%	
Tourism	1%	1%	49%	
---				

Cause: both domains share a large common vocabulary

# Legal Framework, Methods confusion

In \ out	Legal	Methods	Urbanism	Infra...
Legal	8%	3%	5%	3%
Methods	2%	4%	4%	13%
Urbanism	17%	14%	24%	4%
Infrastruct ure	2%	11%	1%	22%

Legal framework and Methods are orthogonal to the other domains.

Documents are rarely only about Law or Methods, they present legal aspects of Urbanism, Transportation, etc.

# TOP50 words analysis

For each domain, select the “most classifying” (highly weighted) terms in the neural network.

Domain = `Environment`, nbdoc = 326

paysagiste, écologique, paysagères, écologiques, biodiversité, jardins, paysagistes, marais, parcs-naturels, jardin, directive, environnementales, naturel, paysages, pnr, protection, espèces, berges, paysagère, naturels-régionaux, paysage, arbres, précaution, faune, éco, forestier, protection-nature, environnemental, environnementale, green, pédagogiques, charte, écologie, patrimoine-naturel, vertes, ceinture, naturelles, verts, landscape, utilisé, principe-précaution, ceinture-verte, empreinte-écologique, durables, littoral, parcs, baie, conservation, participer, plans-programmes

Yields a set of domain terms (for this corpus)

# TOP50 terms vs. Thesaurus

Compare the top 50 terms of a domain with the thesaurus terms for this domain.

paysagiste, écologique, paysagères, écologiques, biodiversité, jardins, paysagistes, marais, parcs-naturels, jardin, directive, environnementales, naturel, paysages, pnr, protection, espèces, berges, paysagère, naturels-régionaux, paysage, arbres, précaution, faune, éco, forestier, protection-nature, environnemental, environnementale, green, pédagogiques, charte, écologie, patrimoine-naturel, vertes, ceinture, naturelles, verts, landscape, utilisé, principe-précaution, ceinture-verte, empreinte-écologique, durables, littoral, parcs, baie, conservation, participer, plans-programmes

34 terms not in Urbamet (*in red*)

**Hypothesis:** The *Environment* domain has changed since 1969. Thesaurus updates were not able to reflect this change. Environment now denotes two domains : *Urban environment* and *Ecology*.

# Toward a Methodology for Thesaurus Evolution

Use automated classification to validate the domains

1. Analyze not clearly separated domains (e.g. traffic - transportation)
  - Check the quality of document classification
  - Merge the domains and create new subdomains
  
2. Find orthogonal domains
  - Build a hierarchy of domains
    - (e.g. Law and Method as subdomains of all others)

### 3. Analyze the classifying terms

- Discover the emergence of new subdomains
- Discover new domains which span other domains

Example: Computer science emerged from Mathematics, Automation, Electronics (< 1970)

### 4. Create concepts form TOP50 terms

- Use them to create domain ontologies

### 5. Repeat every n years

- Compare confusion matrices
- Compare top 50 terms

# Conclusion

Thesauruses are not ontologies. In particular, their structure is not a "is a" hierarchy.

It is interesting to consider them as hierarchies of **domains** connected with **corpuses of documents**.

Text mining techniques can be used on indexed documents to

- analyze the thesaurus
- update it (restructure the domains)
- find (new) domain terms (to build ontologies)

# From Thesauruses + Documents to Ontologies

