Incremental Development of a Shared Urban Ontology: the Urbamet Experience

Jacques Guyot[1], Gilles Falquet[1], Jacques Teller[2]

[1] Université de Genève, Centre universitaire d'informatique,
7 route de Drize, CH-1227 Carouge
jacques.guyot@unige.ch, gilles.falquet@unige.ch

[2] Université de Liège – LEMA, Chemin des Chevreuils, 1 B52/3, B-4000 Liège
jacques.teller@ulg.ac.be

Abstract. Thesauruses are used for document referencing. They define hierarchies of domains. We show how document and domain contents can be used to validate and update a classification based on a thesaurus. We use document indexing and classification techniques to automate these operations. We also draft a methodology to systematically address those issues. Our techniques are applied to Urbamet, a thesaurus in the field of town planning.

Keywords: ontology, thesaurus, text mining, automated classifier, Urbamet,

1. Introduction

We shall briefly remind below the definition of a thesaurus. We shall also describe the concept and history of Urbamet, the collection which was used in our experiment. We shall then explain how we use automated classification with a view to creating or updating thesauruses.

1.1 Thesauruses

For years, thesauruses were the favorite tool used by librarians and documentalists to classify documents. This classification was meant to facilitate document search by users. A thesaurus includes a set of terms used in a given domain, which is distributed among a hierarchy of sub-domains. The set of chosen terms becomes a controlled vocabulary whose meaning is strictly defined by the thesaurus designers. The structure of thesauruses, in particular the types of relationships between terms, has been studied an normalized during the last decades (ANSI/ISO, 2005), (ISO, 1986). A thesaurus is used by the documentalist to assign one or several domain(s) to each document, and to assign keywords chosen from the terms of the assigned domain(s). User requests are then expressed through those keywords and domains.

Indexing documents with the help of thesauruses is a technique that allows building up classifications:

- Without using a computer
- Without knowing the document contents (esp. in the case of multimedia documents).

This is particularly useful when those elements cannot be requested. However, building up a thesaurus requires a major investment in terms of creation, learning and maintenance work. Indexing and search engines that operate directly on document contents have changed the habits of users, who would rather use directly their own vocabulary. Such a method is good enough when the corpus is very large (the complete Web in the case of Google or Yahoo!) but in that case the issue of polysemy (*i.e.* multiple meaning) creates "noise" in the results; this noise would require introducing domains to filter out the results. As for keywords, they are necessary to harmonize the vocabulary in the case of a middle-size corpus (*i.e.* a corpus for which all possible expressions are not available for a given piece of information).

Thesauruses are thus still necessary when it comes to indexing and searching documents on the basis of their content.

1.2 Urbamet

Urbamet (Urbamet, 2009) is a bibliographic database created and maintained by the French Centre for Urban Documentation. The corpus currently includes 280'000 documents and is fed with an additional 8'000 documents each year. Originally designed in 1969 with 2'300 terms, the Urbamet thesaurus currently includes 4'200 terms, which are used to index the document corpus. It is a hierarchy of terms with 24 main themes (top level categories) . Figure 1 shows the main themes and an excerpt of the hierarchy of sub-domains in the field of transportation.

It can be observed on the figure that the terms in Urbamet denote either concepts or sub-domains. For instance, the term “utility vehicle” may denote a concept that has an intension (the properties of a utility vehicle) and an extension (the set of all utility vehicles). Conversely, the term “road and traffic” can hardly denote a concept: it is difficult to figure out what is an instance of “road and traffic”. Moreover “road and traffic” cannot be considered as a specialization of its parent term “land transport”. Hence, the Urbamet thesaurus, at least on the first levels, is mostly a hierarchy of sub-domains. As a consequence, it does not provide a starting point or backbone for the construction of an urban ontology.

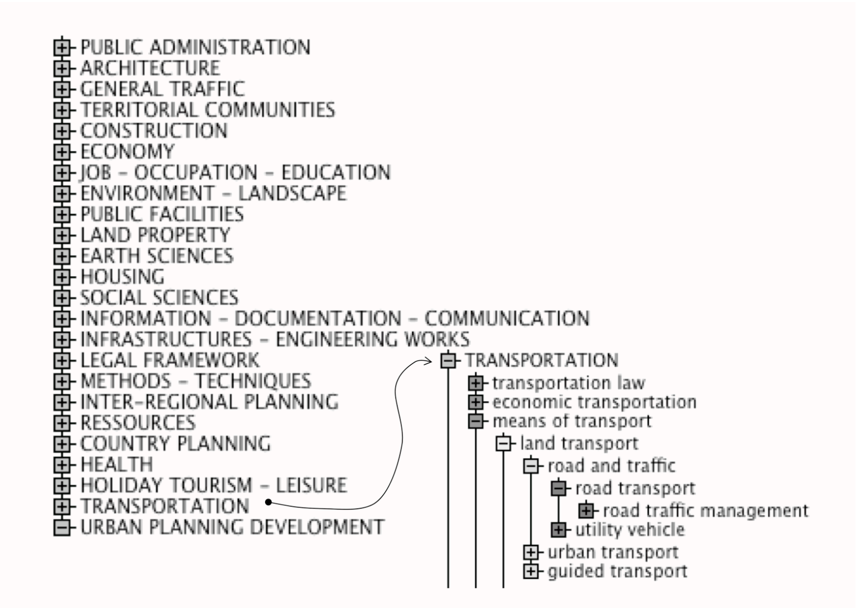


Fig. 1. Urbamet main themes and the *Transportation* sub-hierarchy

1.3 Methodology

Since the thesaurus cannot be directly used to build an ontology, the proposed methodology relies on the existing thesaurus and the indexed document corpus. The document classification induced by the thesaurus is analyzed with an automated document classifier. This tool operates on document contents. Initially a training corpus is used to teach the classifier the class concepts. Then the tool can start classifying other documents. The analysis is performed in the following steps:

1. Extracting the corpus
2. Building up the training catalogue
3. Training and validating the classifier
4. Generating and analyzing the confusion matrix (list of mistakes made by the classifier)
5. Generating the Top-50 terms (list of the most classifying terms)

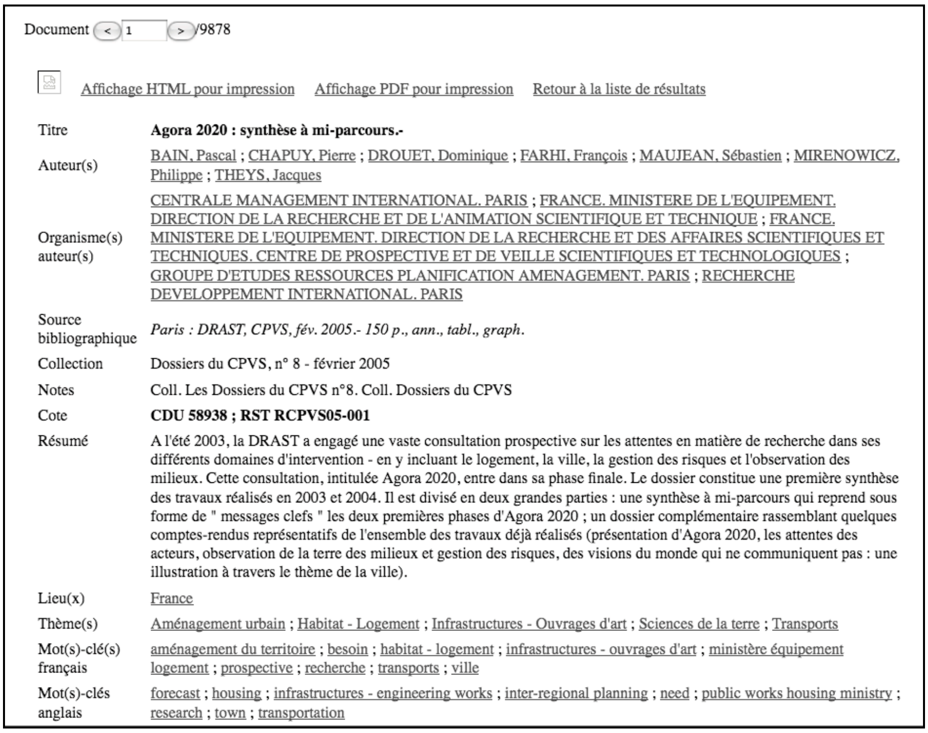
We shall show how the analysis of the confusion matrix and the Top-50 list helps us understand how the corpus is structured in terms of domains and how the thesaurus may be re-structured on the basis of these indications.

2. Analyzing the Urbamet

We use a classifier based on the neural network technique. The goal of the classifier is, for each document, to predict the class to which it belongs. The input is the list of document terms[1] and the output is a score (between 0 and 1) for each class. In this case the classes are the 24 top-level Urbamet domains.

2.1 Training the classifier

To build the training and test corpus we extracted from the Urbamet web site about 10'000 information sheets similar to the one shown in 0 (which is mostly in French):


Document < 1 > /9878
Affichage HTML pour impression
Affichage PDF pour impression
Retour à la liste de résultats

Titre	**Agora 2020 : synthèse à mi-parcours.-**
Auteur(s)	BAIN, Pascal ; CHAPUY, Pierre ; DROUET, Dominique ; FARHI, François ; MAUJEAN, Sébastien ; MIRENOWICZ, Philippe ; THEYS, Jacques
Organisme(s) auteur(s)	CENTRALE MANAGEMENT INTERNATIONAL. PARIS ; FRANCE. MINISTERE DE L'EQUIPEMENT. DIRECTION DE LA RECHERCHE ET DE L'ANIMATION SCIENTIFIQUE ET TECHNIQUE ; FRANCE. MINISTERE DE L'EQUIPEMENT. DIRECTION DE LA RECHERCHE ET DES AFFAIRES SCIENTIFIQUES ET TECHNIQUES. CENTRE DE PROSPECTIVE ET DE VEILLE SCIENTIFIQUES ET TECHNOLOGIQUES ; GROUPE D'ETUDES RESSOURCES PLANIFICATION AMENAGEMENT. PARIS ; RECHERCHE DEVELOPPEMENT INTERNATIONAL. PARIS
Source bibliographique	*Paris : DRAST, CPVS, fév. 2005.- 150 p., ann., tabl., graph.*
Collection	Dossiers du CPVS, n° 8 - février 2005
Notes	Coll. Les Dossiers du CPVS n°8. Coll. Dossiers du CPVS
Cote	**CDU 58938 ; RST RCPVS05-001**
Résumé	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).
Lieu(x)	France
Thème(s)	Aménagement urbain ; Habitat - Logement ; Infrastructures - Ouvrages d'art ; Sciences de la terre ; Transports
Mot(s)-clé(s) français	aménagement du territoire ; besoin ; habitat - logement ; infrastructures - ouvrages d'art ; ministère équipement logement ; prospective ; recherche ; transports ; ville
Mot(s)-clés anglais	forecast ; housing ; infrastructures - engineering works ; inter-regional planning ; need ; public works housing ministry ; research ; town ; transportation

Fig. 2. Sample Urbamet information sheet

[1] Terms are the text words and co-occurrences of two words appearing with a given frequency. For instance, "engineering works" is recognized as a single term. We can also use term stemming techniques by grouping under a single tag all words from the same family. For example, "manage, managed, management, managing" can be considered as a single term. A list of so-called "stop-words" is used to get rid of noisy words such as "the, it, and, …"

We used the Winnow algorithm to set the neuron weights that link the terms to the classes. Initially all the weights have the same value. For each document in the training corpus, and according to the terms used in the text, we sum up the values in each class. Then we sort out the results according to the sums. For the classes which are above a given threshold and which are not correctly predicted, we lower the weight of the document terms. Conversely, for the classes which are below the threshold and which are correctly predicted, we raise the term weights. Thus by simulating a "punishment/reward" heuristic, we produce a neural network that has learned the underlying classification of the training corpus.

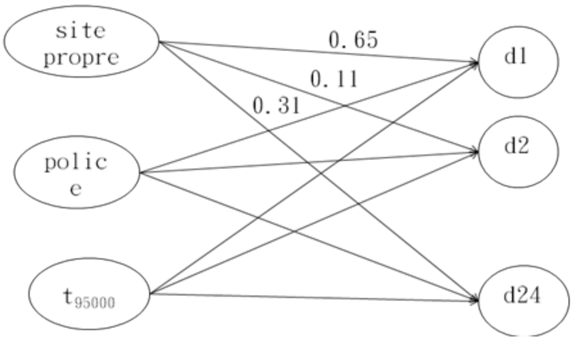


Fig. 3. Neural network with weights linking terms to classes

2.2 Testing the Classifier

The quality of the classification must be tested. It depends essentially upon whether the condition of class inclusion can be deducted from the document content, and thus from the terms (for instance, dividing a corpus in two parts on the basis of whether the ideas in a document are ethical or not is an example where the terms are of little use). To perform such a test, we separate a part (20%) of the training corpus which is not used in the training phase; it is used later on to evaluate the classifier performance. In the table below, we see that the first class predicted by the classifier is correct in 59% of cases, while it would have been predicted in only 4% of cases had the prediction been made randomly. The following lines show results where the second and third choices were added up to the first one.

Table 1. Performance of the classifier

Number of Predicted Classes	Classifier Performance	Random Choice Performance
1	59 %	4 %
2	75 %	12 %
3	81 %	23 %

From the results above, we can see that the Urbamet classifier is effective and that the Urbamet classification can be deduced from the document contents. Therefore there is a relationship between the document terms and the document classes.

2.3 The Confusion Matrix

It is interesting to examine the classifier's mistakes. These mistakes are due to the fact that it is difficult to distinguish the classes on the basis of their respective vocabulary. A confusion matrix may be built up on these mistakes: Each row of the matrix represents the class which should be found, while each column represents the predicted class. Ideally, only the diagonal should be filled up to 100%. The complete confusion matrix can be found in appendix. Let us take a look at its exceptions.

2.4 Vocabulary shared between two or several domains

Although the *Transportation* and *Traffic* domains are relatively well separated from the other domains, 24% of the documents which should have been classified in *Transportation* were actually classified in *Traffic* and 10% of documents which should have been in *Traffic* were in *Transportation.*

This is due to the fact that the vocabulary is common to both the *Transportation* and *Traffic* domains and thus makes the separation difficult.

Table 2. Confusion matrix for *Transportation*, *Traffic* and *Tourism*

In \ out	Transportation	Traffic	Tourism	…
Transportation	45%	24%	3%	
Traffic	10%	40%	1%	
Tourism	1%	1%	49%	

2.5 Orthogonality of domains

The *Legal* and *Methods* domains are not well distinguished from the other domains. The documents that should have been classified in those two classes were in fact scattered across all domains.

Table 3. Confusion matrix for *Legal*, *Methods*, *Urbanism*, and *Infrastructure*.

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

This is due to the fact that *Legal* and *Methods* are not domains of urbanism, but rather aspects of it. One could equally speak of the legal aspects of transportation, of real estate, or of environment. Such domains are said to be "orthogonal" to the other ones.

2.6 Top-50 Term List of a Domain

Neural networks can be criticized because of the lack of explanation on the fact that the classifier chose a particular class (as compared to rule-based engines which can explain their reasoning). However, it is always interesting to analyze, for each class, the list of the most heavily weighted terms. This list is a selection of the "champion terms" of the domain.

We shall only analyze here one domain, namely *Environment* (which includes 326 documents). The list of the most heavily weighted terms (in French) is the following:

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 [2]

These terms are the "champions" of this specific class for this specific corpus.

[2] In this experience we did not use any stemming technique, which explains why some terms are found both in singular and plural forms.

2.7 Top-50 Terms Versus Thesaurus Terms

We then compared the top-50 terms with the terms of the Urbamet thesaurus related to the *Environment* 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***

The terms that were not included in the Urbamet thesaurus are displayed in bold and underlined. It appears that 34 terms out of the 50 were not in the thesaurus. The hypothesis we make to explain this fact is the following:

- The documents which are classified in the *Environment* domain are correctly classified
- The *Environment* domain has changed since 1969
- The thesaurus updates do not reflect those changes
- The *Environment* domain includes in fact two domains: one is *Urban Environment* and the other one is *Ecology*.

3. Towards a Methodology to Update Thesauruses

The examples provided in the previous sections show that automated analysis tools are relevant. Yet a detailed analysis of the confusion matrix and the Top-50 list definitely requires a corpus expert (a documentalist who knows well his/her corpus and thesaurus).

We suggest some elements of methodology when using an automated classifier to validate the domains. It should be assessed initially that the classifier is globally able to reach a given level of efficiency. Then the confusion matrix allows to:

1. Analyze the domains that are not clearly separated (such as *Traffic* and *Transportation*). In such a situation the following steps should be applied:
   - Check out the quality of the classification for both domains
   - Possibly merge both domains into a single one and then separate them into two sub-domains.
2. Look for orthogonal domains that would be distributed across all domains (such as *Legal* and *Method*). In that case it could be necessary to:
   - Build up a hierarchy of domains. For example *Legal* and *Method* are sub-domains of all the other domains.

A hierarchy of domains may be used to train a classifier by building up a neural network (*i.e.* a classifier) for each node in the hierarchy. In our example, we can see that confusion may be avoided by removing the *Legal* and *Method* domains from the first level of the hierarchy. Indeed, the terms related to legal and methodological issues will be scattered across the various other domains and as such will be lightly weighted. Conversely, at the second level of the classification, the domain-related terms will be lightly weighted while the legal and methodological terms will be heavier.

With the Top-50 list of terms we can:

1. Analyze the highly classifying terms:
   - To discover an emerging new domain which was covered by another one (such as *Urban Environment* and *Ecology*)
   - To discover an emerging new domain which was distributed among several other domains (a typical example is computer science, which before the 1970s did not exist as an independent domain but was considered to be either mathematics, automatics or electronics)
2. Turn the classifying terms into concepts of an ontology:
   - These concepts are the seeds on which an ontology may be grown up.
3. Repeat the previous steps on a regular basis (every *x* years):
   - Repeating these steps allows monitoring how the confusion matrix and the Top-50 list evolve, which is an indication of how the domains themselves evolve.

4. Conclusion

Thesauruses such as the Urbamet thesaurus are not ontologies. In particular, their hierarchical structure is not an "*is a*" relationship: therefore they should be considered as a hierarchy of domains which is connected with a corpus of documents.

We have shown that in such situations, text mining techniques, and more particularly automated document classification techniques may be used to support the following actions:

- Analyzing the thesaurus
- Maintaining the thesaurus and restructuring domains
- Finding new domain terms to build up ontologies.

Typical text mining based ontology extraction tools, such as the OntoLearn system (Velardi et al., 2001), rely on statistical analysis of the corpus terms, together with syntactic analysis. The methodology we propose takes advantage of an existing classification scheme and, to a certain extent, discovers how and why it works (e.g. it finds the most classifying terms). This discovery process yields insights into the structure of the domain and thus provides a basis for building an ontology.

The methodology we propose must still be evaluated on other test cases. However, as far as we know, the evolution of knowledge resources has not been documented in the urban field. Thus we intended to test our approach on a physics corpus for which there exists a classification (the Annual Classification of the Physikalische Berichte) that has already been studied (Hurni, 2009).

Bibliography

ANSI/ISO (2005). Guidelines for the Construction, Format, and Management of Mono-lingual Thesauri. ANSI/NISO Z39.19-2005, American National Standards Institute (ANSI). Revision of Z39.19-1983.

Gómez-Pérez A, Fernández-López M, Corcho O (2003). Ontological Engineering, chapter Methodologies and Methods for Building Ontologies. Springer-Verlag, London (United Kingdom).

Gómez-Pérez A, Manzano-Macho D (2003). A survey of ontology learning methods and techniques. OntoWeb Consortium, Deliverable 1.5.

Hurni, J.-P (2009). Cornelius Lanczos et les Physikalische Berichte (1923–1933), ISRI Technical Report ISRI-Ge-09-01, Genève. (in preparation).

ISO (1986). Guidelines for the establishment and development of monolingual thesauri. ISO 2788, International Organization for Standardization (ISO).

Kietz JU, Maedche A, , Volz. R (2000). A method for semi-automatic ontology acquisition from a corporate intranet. In Proceedings of Workshop Ontologies and Text, EKAW'2000.

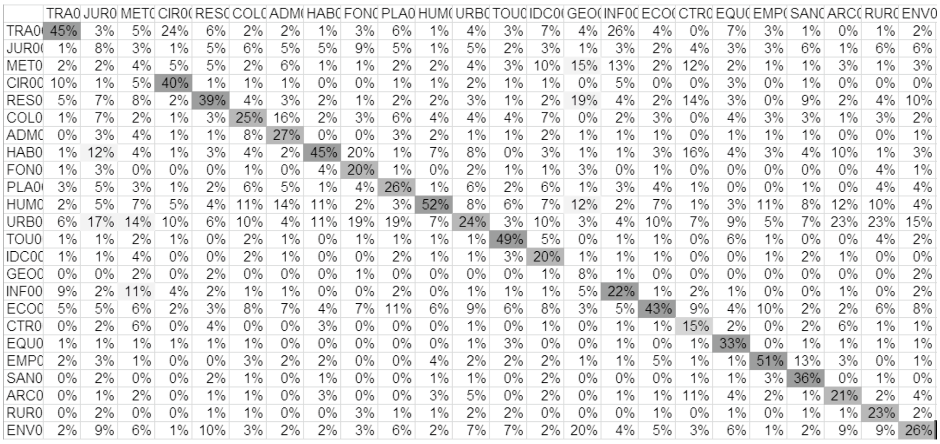
Littlestone, N. (1988) "Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm" Machine Learning 285-318(2)

Urbamet (2009). Urbamet home page. http://www.urbamet.com/ Retrieved on 10 Feb. 2009

Velardi P, Fabriani P, Missikoff M (2001). Using text processing techniques to automatically enrich a domain ontology. In Proceedings of the international conference on Formal Ontology in Information Systems, FOIS 2001, pp 270–284.

Appendix

Confusion matrix for the 24 domains



	TRA0	JUR0	MET(	CIR0(	RES(	COL(	ADM(	HAB(	FON(	PLA0	HUM(	URB(	TOU(	IDC0(	GEO(	INF0(	ECO(	CTR(	EQU(	EMP(	SAN(	ARC(	RUR(	ENV0
TRA0(	45%	3%	5%	24%	6%	2%	2%	1%	3%	6%	1%	4%	3%	7%	4%	26%	4%	0%	7%	3%	1%	0%	1%	2%
JUR0(	1%	8%	3%	1%	5%	6%	5%	5%	9%	5%	1%	5%	2%	3%	1%	3%	2%	4%	3%	3%	6%	1%	6%	6%
MET0	2%	2%	4%	5%	5%	2%	6%	1%	1%	2%	2%	4%	3%	10%	15%	13%	2%	12%	2%	1%	1%	3%	1%	3%
CIR0(	10%	1%	5%	40%	1%	1%	1%	0%	0%	1%	1%	2%	1%	1%	0%	5%	0%	0%	3%	0%	1%	0%	0%	0%
RES0	5%	7%	8%	2%	39%	4%	3%	2%	1%	2%	2%	3%	1%	2%	19%	4%	2%	14%	3%	0%	9%	2%	4%	10%
COL0	1%	7%	2%	1%	3%	25%	16%	2%	3%	6%	4%	4%	4%	7%	0%	2%	3%	0%	4%	3%	3%	1%	3%	2%
ADM(	0%	3%	4%	1%	1%	8%	27%	0%	0%	3%	2%	1%	1%	2%	1%	1%	1%	0%	1%	1%	1%	0%	0%	1%
HAB0	1%	12%	4%	1%	3%	4%	2%	45%	20%	1%	7%	8%	0%	3%	1%	1%	3%	16%	4%	3%	4%	10%	1%	3%
FON0	1%	3%	0%	0%	0%	1%	0%	4%	20%	1%	0%	2%	1%	1%	3%	0%	1%	0%	0%	0%	0%	0%	4%	1%
PLA0(	3%	5%	3%	1%	2%	6%	5%	1%	4%	26%	1%	6%	2%	6%	1%	3%	4%	1%	0%	0%	1%	0%	4%	4%
HUM(	2%	5%	7%	5%	4%	11%	14%	11%	2%	3%	52%	8%	6%	7%	12%	2%	7%	1%	3%	11%	8%	12%	10%	4%
URB0	6%	17%	14%	10%	6%	10%	4%	11%	19%	19%	7%	24%	3%	10%	3%	4%	10%	7%	9%	5%	7%	23%	23%	15%
TOU0	1%	1%	2%	1%	0%	2%	1%	0%	1%	1%	1%	1%	49%	5%	0%	1%	1%	0%	6%	1%	0%	0%	4%	2%
IDC0(	1%	1%	4%	0%	0%	2%	1%	0%	0%	2%	1%	1%	3%	20%	1%	1%	1%	0%	0%	1%	2%	1%	0%	0%
GEO(	0%	0%	2%	0%	2%	0%	0%	0%	1%	0%	0%	0%	0%	1%	8%	1%	0%	0%	0%	0%	0%	0%	0%	2%
INF00	9%	2%	11%	4%	2%	1%	1%	0%	0%	2%	0%	1%	1%	1%	5%	22%	1%	2%	1%	0%	0%	1%	0%	2%
ECO(	5%	5%	6%	2%	3%	8%	7%	4%	7%	11%	6%	9%	6%	8%	3%	5%	43%	9%	4%	10%	2%	2%	6%	8%
CTR0	0%	2%	6%	0%	4%	0%	0%	3%	0%	0%	0%	1%	0%	1%	0%	1%	1%	15%	2%	0%	2%	6%	1%	1%
EQU0	1%	1%	1%	1%	1%	1%	0%	0%	0%	0%	0%	1%	3%	0%	0%	1%	0%	1%	33%	0%	1%	1%	1%	1%
EMP(	2%	3%	1%	0%	0%	3%	2%	2%	0%	0%	4%	2%	2%	2%	1%	1%	5%	1%	1%	51%	13%	3%	0%	1%
SAN0	0%	2%	0%	0%	2%	1%	0%	1%	0%	0%	1%	1%	0%	2%	0%	0%	0%	1%	1%	3%	36%	0%	1%	0%
ARC0	0%	1%	2%	0%	1%	1%	0%	3%	0%	0%	3%	5%	0%	2%	0%	1%	1%	11%	4%	2%	1%	21%	2%	4%
RUR0	0%	2%	0%	0%	1%	1%	0%	0%	3%	1%	1%	2%	2%	0%	0%	0%	1%	0%	1%	0%	1%	1%	23%	2%
ENV0	2%	9%	6%	1%	10%	3%	2%	2%	3%	6%	2%	7%	7%	2%	20%	4%	5%	3%	6%	1%	2%	9%	9%	26%