CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Outils de veille : typologie

Claire FRANCOIS

Unité de recherche et Innovation

INIST / CNRS

Rencontres des professionnels de l'IST, 19 juin 2003, PARIS

Plan

- Introduction
- Les grands types d'outils de veille
- Les différentes fonctionnalités
- Les technologies mises en œuvre
- 2 exemples : Stanalyst ® et Leximine

Introduction

- Veille : différents besoins selon étapes du processus
  - Recherche d'information et surveillance
  - Collecte et gestion
  - Analyse :
    - Textuelle
    - Bibliométrie
    - Infométrie
    - Visualisation
  - itérations et interactions

Types d'outils de veille

- Recherche d'information
  - Bases de données ,
  - Internet :
    - portails, annuaires
    - Moteurs de recherche, Agents intelligents
- Surveillance
  - Outils de surveillance des différentes sources
  - Outils de push
- Gestion documentaire et gestion des connaissances
- Exploration du contenu :
  - Analyse texte
  - Classification et Cartographie
  - Évolutions temporelles

Keywatch
ApertoLibro

- Systèmes complets de gestion de l'information

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Méthodes utilisées

- Bibliométrie
  - différents modes de représentation des résultats
- Analyse textuelle
  - requêtes et textes collectés
  - Statistique ou linguistique
- Infométrie
  - Classement ou catégorisation
  - Classification automatique
  - Evolution temporelle
- Visualisation de l'information
  - Arborescence
  - graphes
  - Cartographie

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Les méthodes bibliométriques

Exemples :

Tétralogie, WordMapper, Simbad, Stanalyst

Rencontres des professionnels de l'IST, 19 juin 2003, PARIS

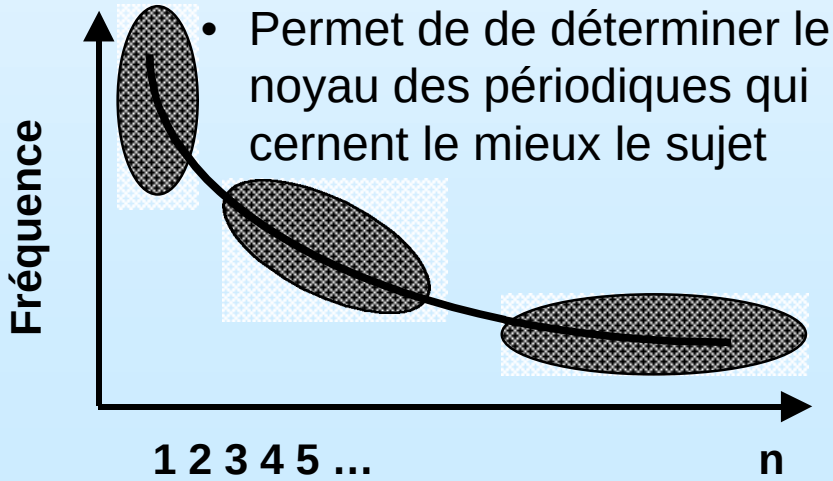
Quelles analyses ?

Construire des tableaux de répartition des références selon un ou plusieurs critères

Type d'Information	Type d'Analyse
Mots-clés, titres et résumés des documents **Code de classement**	**Analyse de contenu**
Auteurs **Affiliations (Établissement, pays)**	**Analyse des acteurs et de leurs relations**
Type de document : revues, rapports, communications de congrès **Pays éditeur** **Langue de communication**	**Analyse de leurs moyens de communication**
Date de publication	**Analyse de l'évolution de l'activité de différents domaines**

Les lois bibliométriques

- Bradford : productivité des revues scientifiques
  - Une relative petite proportion de périodiques peut satisfaire la requête d'une grande proportion d'articles.
  - Permet de de déterminer le noyau des périodiques qui cernent le mieux le sujet



- Lotka : productivité des auteurs
  - Il y a, par rapport au nombre d'auteurs qui publient un article, n^2 fois moins d'auteurs qui publient n articles.
- Zipf : répartition des mots-clés
  - Principe du moindre effort, on utilise plus facilement des mots familiers que des mots inhabituels

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Institut de l'Information Scientifique et Technique

Analyse textuelle

Rencontres des professionnels de l'IST, 19 juin 2003, PARIS

Approches statistiques

- Traitements de lématisation
- Fréquence et cooccurrence des termes
- Segments répétés
- Utilisation de ressources terminologiques :
  - Dictionnaires de mots vides,
  - Dictionnaires d'équivalence,
- Outils :
  - Alceste, Tétralogie, WordMapper, Sphynix Lexica …

Approches linguistiques (1)

- Utilisation de ressources terminologiques :
  - Reconnaissance des termes sous la forme d'origine ou sous des formes variantes
- Analyse des phrases :
  - basée sur les patrons syntaxiques et marqueurs linguistiques
- Analyse sémantique
- Résumé ou traduction automatique
  - Ex : Leximine, DigOut4U, Pertimm, Tropes, Insight Discoverer

Les variantes flexionnelles

- **La variation flexionnelle** permet d'identifier, pour chaque terme, les formes singulier / pluriel des noms

Anglais	Français
bacterium/bacteria	*cheval/chevaux*
activity/activities	*eau/eaux*
	corail/coraux

Les variantes syntaxiques (1)

- Variation d'insertion :
  - **ajout de tout mot à l'intérieur du groupe nominal**
    - *structural erm genes -> Structural genes*
    - *Système racinaire de surface -> Système de surface*

- Variation de coordination :
  - **formes coordonnées de mots (adjectifs ou noms) à l'intérieur du groupe nominal**
    - *structural and regulatory genes -> Structural genes*
    - *gène de régulation et de structure -> Gène de structure*

Les variantes syntaxiques (2)

- Variation de permutation :
  - **tout mot ou groupes de mots pouvant permuter autour d'un élément pivot (prépositions ou séquences verbales) :**
    - *mechanism of clarithromycin resistance -> Resistance mechanism*
    - Quasi-inexistante pour le français

Variantes morpho-derivationnelles

- Propriétés linguistiques de mots pouvant être dérivés en d'autres mots de catégories grammaticales différentes :
  - ***abdomen /abdominal***
    - dérivations nom adjectif
  - ***abort/abortion***
    - dérivation verbe nom
  - ***Rare species ->rarely encountered enterococcus species***
    - dérivation de l'adjectif *rare* en adverbe : *rarely*,
    - insertion de *encountered enterococcus*
  - ***Variation de alimentation -> Alimentation hydrique d'un arbre varie***
    - dérivation du nom *variation* en verbe *varie*

Exemple : Processus ILC (1)

- Le traitement automatique des ressources terminologiques :
  - **étiquetage grammatical du lexique (TreeTagger)**
  - **création/utilisation de ressources morpho-dérivationnelle à partir des mots du lexiques**
  - **formatage des lexiques en règles *PATRII***
    - règles sur les mots et règles sur les termes
  - **compilation des lexiques en *PATRII* par l 'analyseur *FASTR***

Exemple : Processus ILC (2)

- Le traitement du corpus
  - **étiquetage grammatical du corpus (TreeTagger)**
  - ***PATRII***
  - **indexation du corpus**

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Infométrie

Rencontres des professionnels de l'IST, 19 juin 2003, PARIS

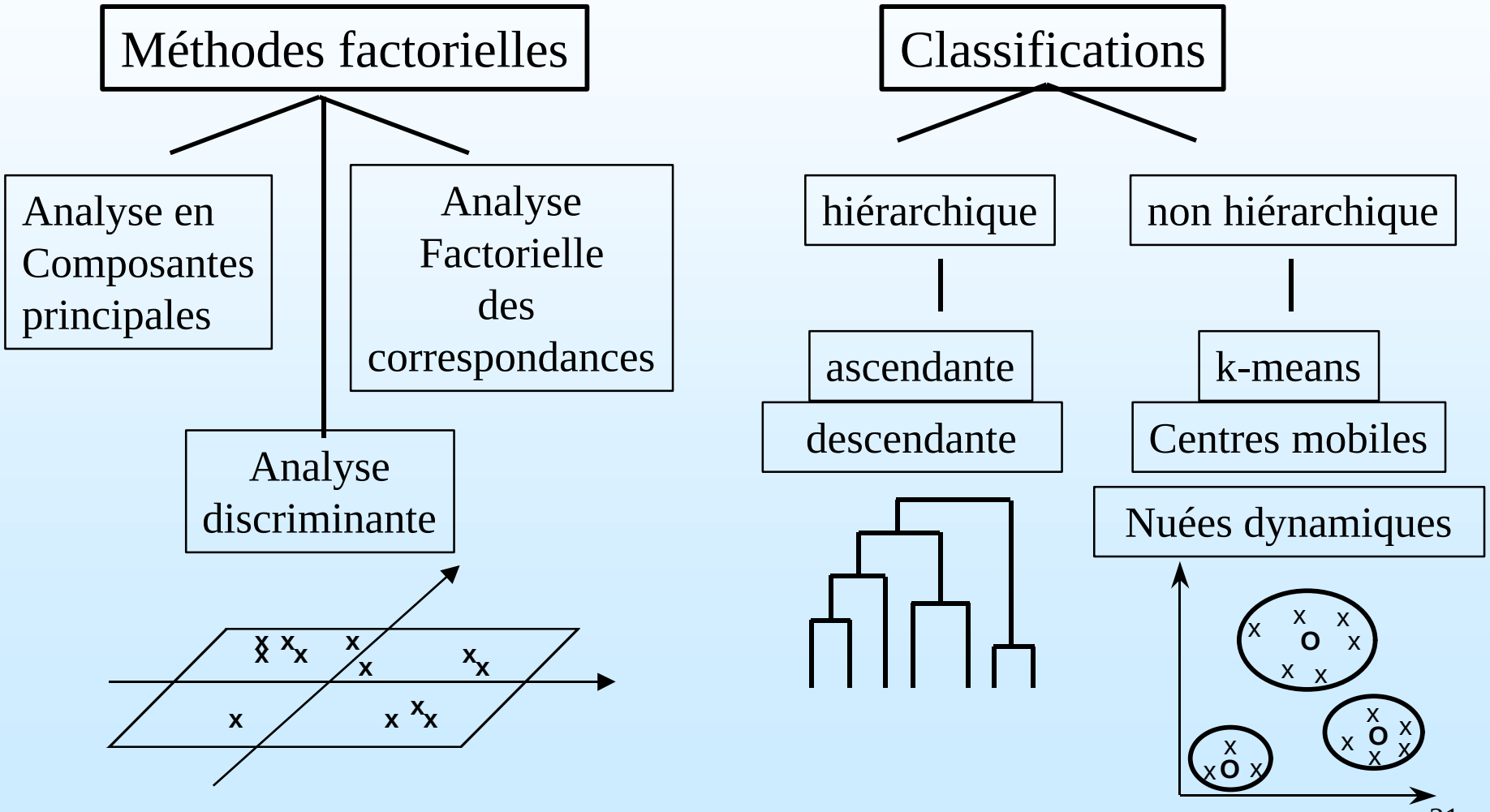
Classement - classification

- Classement ou catégorisation
  - Mode supervisé
  - Affectation des documents dans une arborescence
  - Ex : Exalead, Vivisimo, …
- Classification automatique
  - Méthodes d'analyse de données
  - Méthodes neuronales

Analyse de données (1)

- Méthodes factorielles
  - **Produisent des représentations graphiques permettant de positionner les objets sur un plan**
    = réduction dimensionnelle
  - **Les composantes des axes factoriels**
    = la solution d'une équation mathématique
- Méthodes de classification
  - **Produisent des classes permettant de grouper les objets à décrire**
  - **Définissent les classes à partir d'une formulation**

Analyse de données (2)



Classification automatique

- Mots associés
  - Distance : cooccurrence entre descripteurs
  - Classification hiérarchique
  - Ex : Leximine, WordMapper, SDOC
- K-means axiales
  - Classification non hiérarchique
  - Ex : CartoWeb, Neurodoc
- SOM : cartes auto-adaptatives de Kohonen
  - Méthode neuronale : apprentissage compétitif,
  - Contraintes topographiques
  - Ex : Websom, MultiSom

Evolution temporelle

- Courbe de suivi dans le temps d'un sujet ou d'un terme
  - Ex : Péricles, Leximine
- Evolution entre 2 analyses à T et T+1
  - WordMapper
- Distinction des thèmes émergents, stabilisés et déclinants.
  - Calliope

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

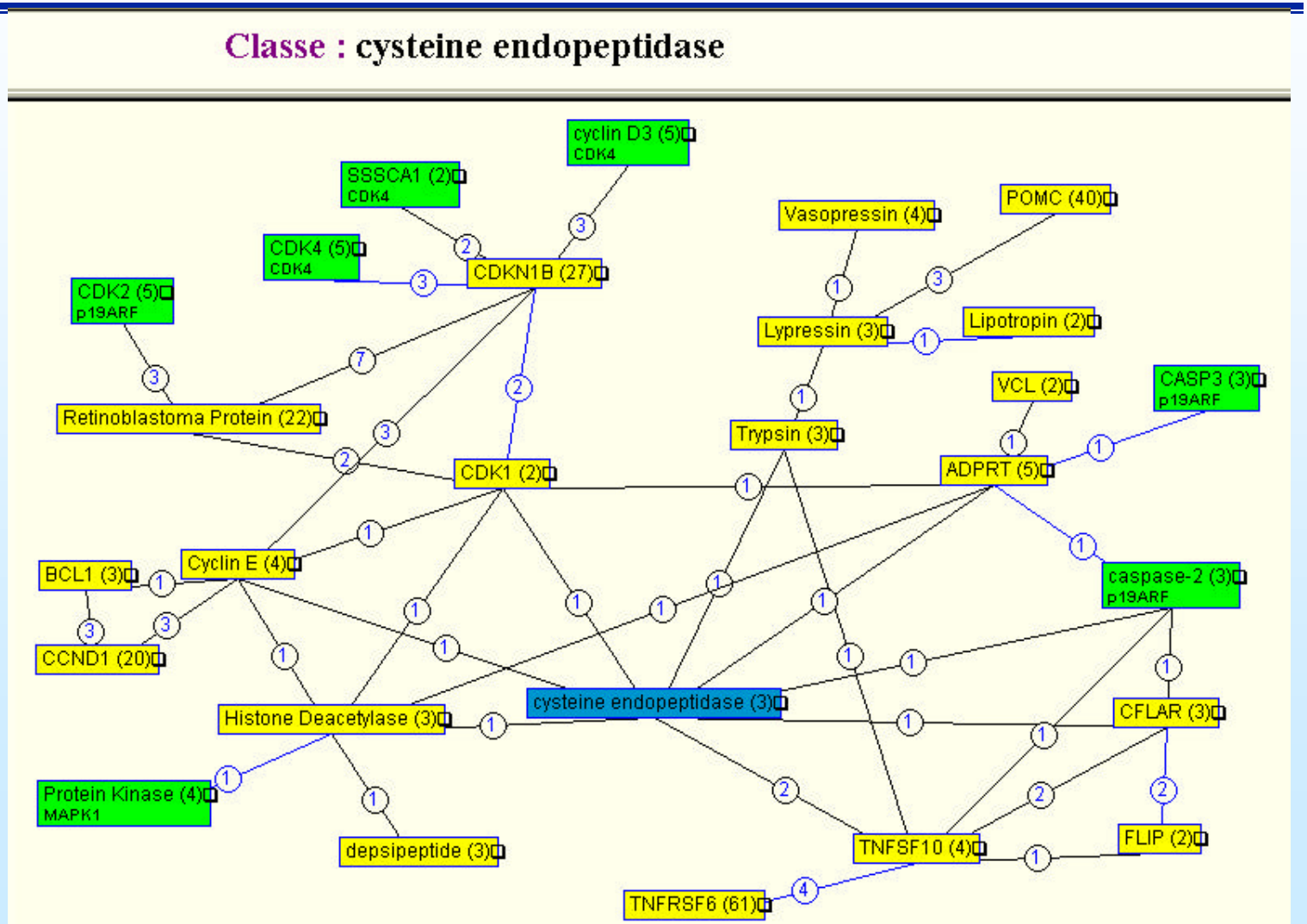
Visualisation

Rencontres des professionnels de l'IST, 19 juin 2003, PARIS

Approches utilisées

- Arborescence
  - Ex : Exalead, Vivisimo
- Graphes des classes, réseaux
  - Ex : Leximine, Wordmapper, Tétralogie
  - moteurs Webbrain, kartoo, Mapstan
- Cartographies :
  - Analyses factorielles
    - Ex : Tétralogie, CartoWeb, Neurodoc,
  - Centralité, densité, ex :Sdoc
  - Cartographie par pavage , ex : Miner 3D, Map.net

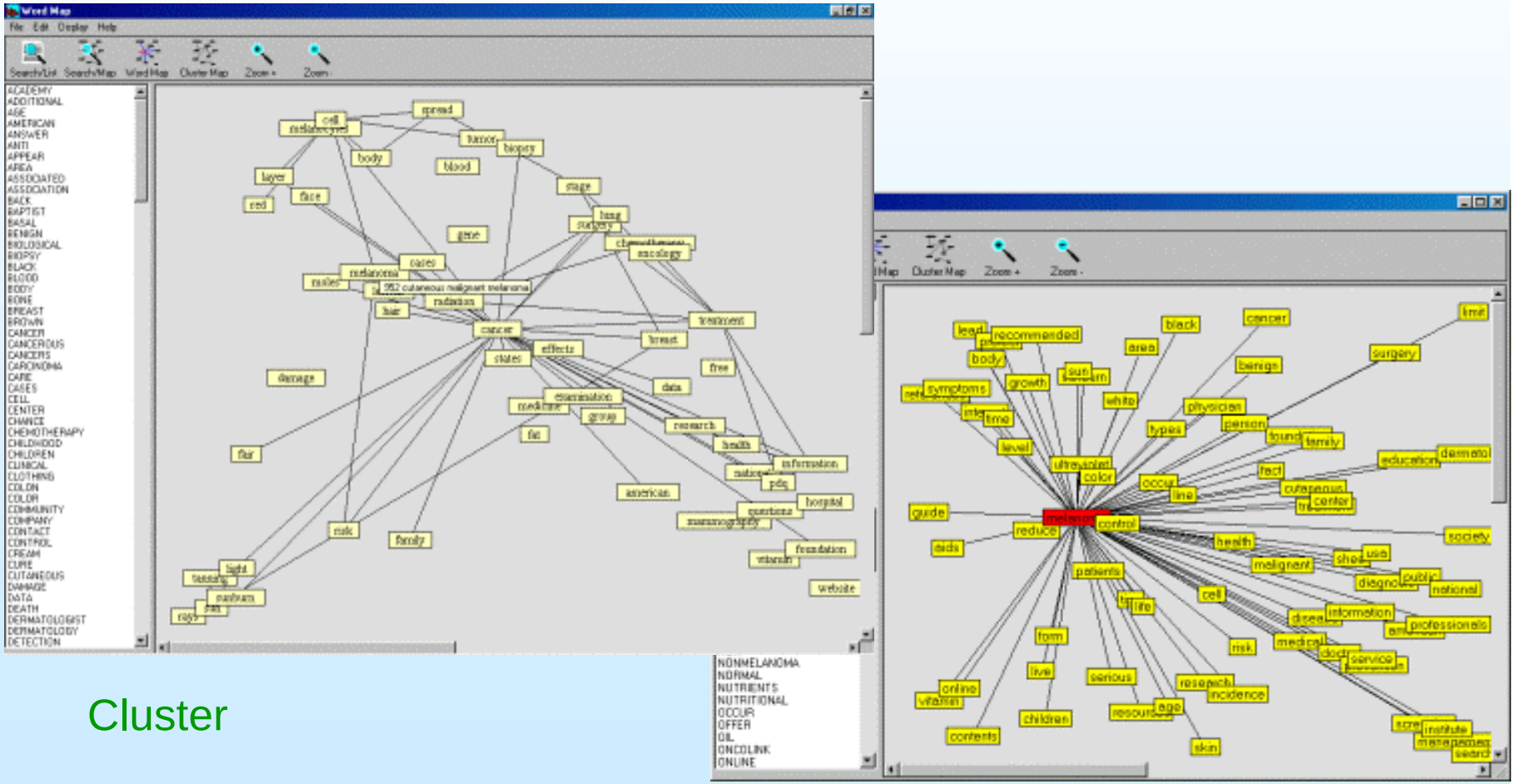
Mots associés : graphe de cluster



CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Institut de l'Information Scientifique et Technique

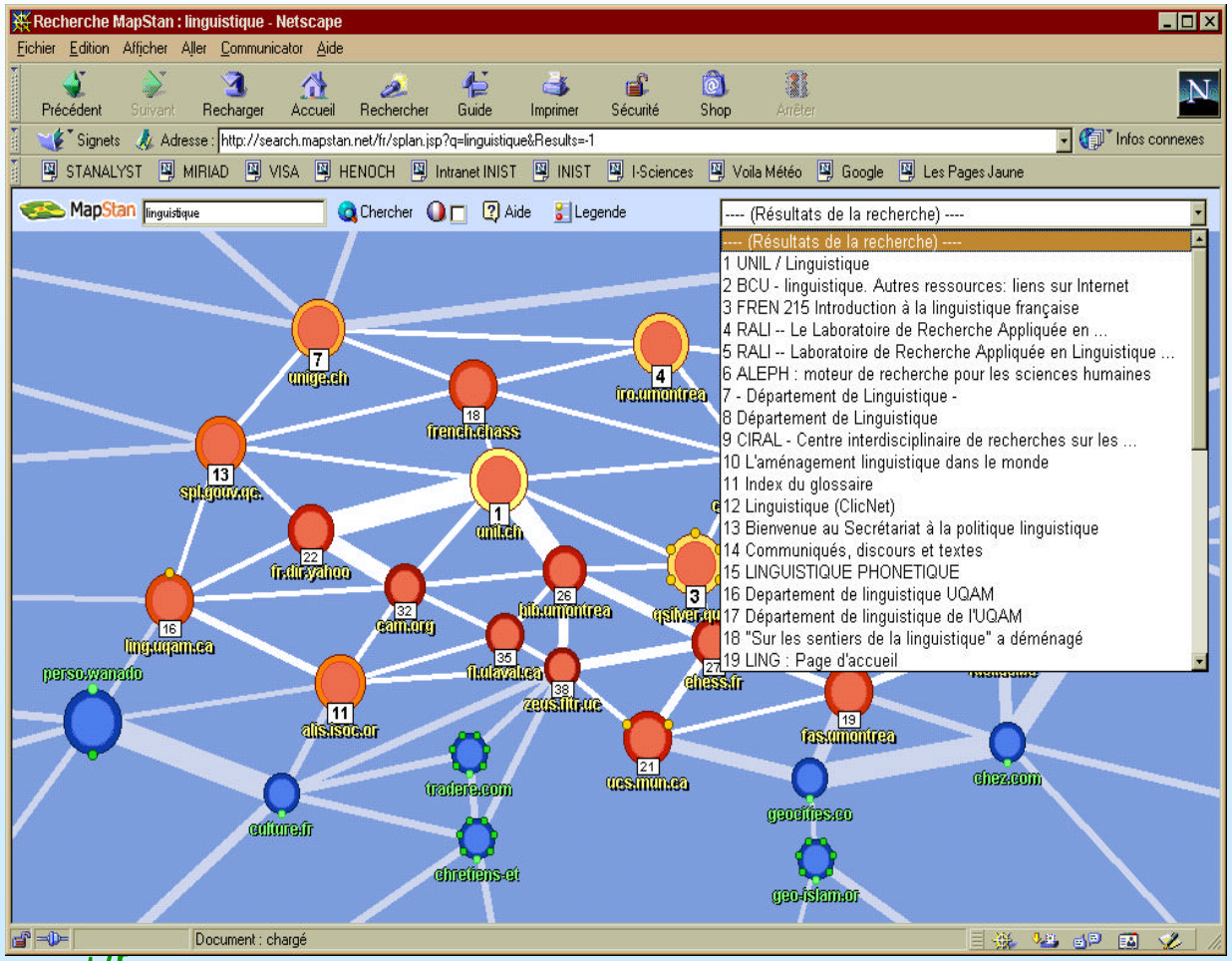
WordMapper



Cluster

MapStanSearch

Cartographie de sites, capitalisation des requêtes des autres internautes.



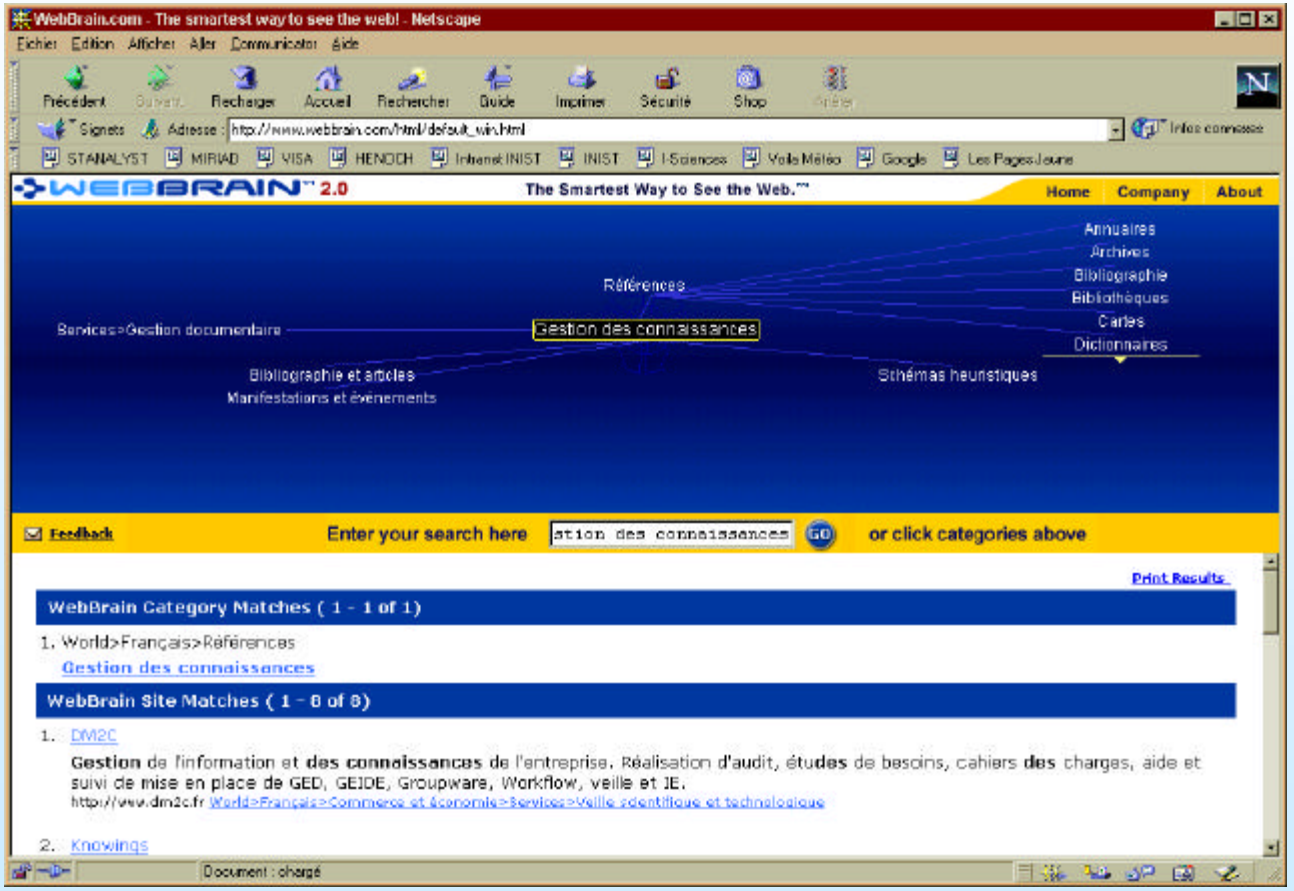
http://search.mapstan.net/fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Webbrain

Navigation dans une hiérarchie de concepts

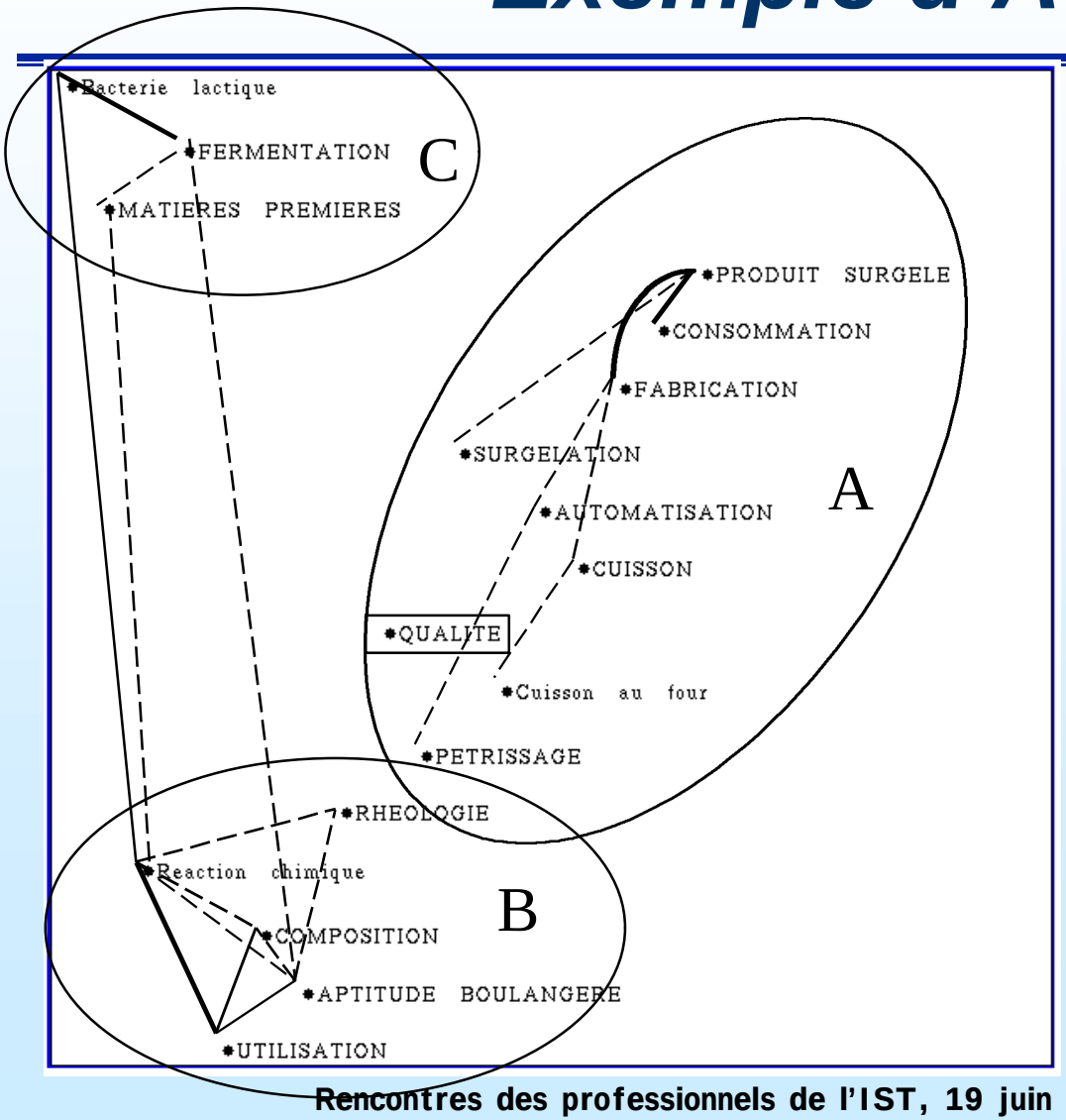


http://www.webbrain.com/html/default_win.html

Les méthodes factorielles

- Analyse en Composantes principales
  - **S'applique aux tableaux de mesures**
    - Objets * caractères quantitatifs
- Analyse factorielle des Correspondances
  - **S'applique aux tableaux de contingence**
    - Caractères qualitatifs * Caractères qualitatifs
    - Ce tableau contient des effectifs
- Analyse factorielle discriminante
  - **Expliquer un caractère qualitatif par un ensemble de caractères quantitatifs.**

Exemple d'ACP



Corrélations entre classes

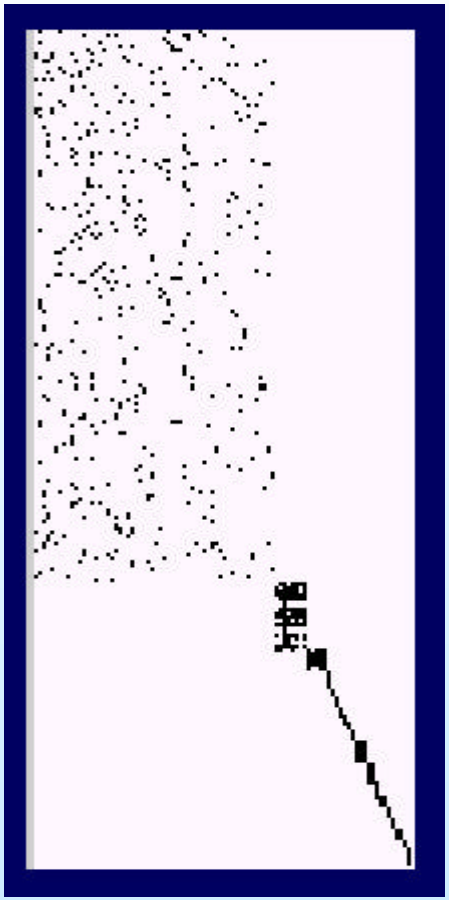
- >= 0,3
- >= 0,2 et < 0,3
- >= 0,1 et < 0,2
- classe isolée

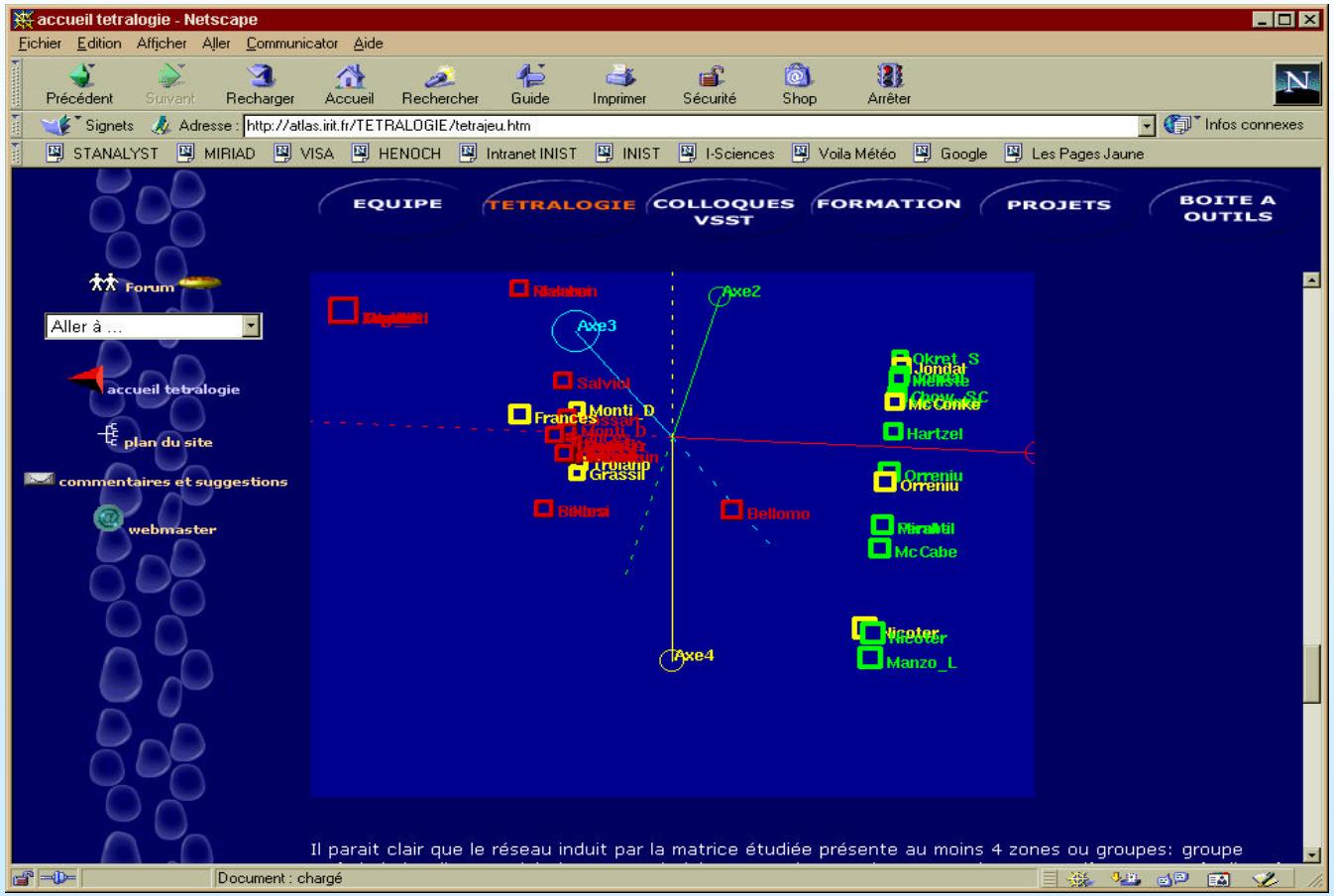
Neurodoc

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

inist Institut de l'Information Scientifique et Technique

Tétralogie

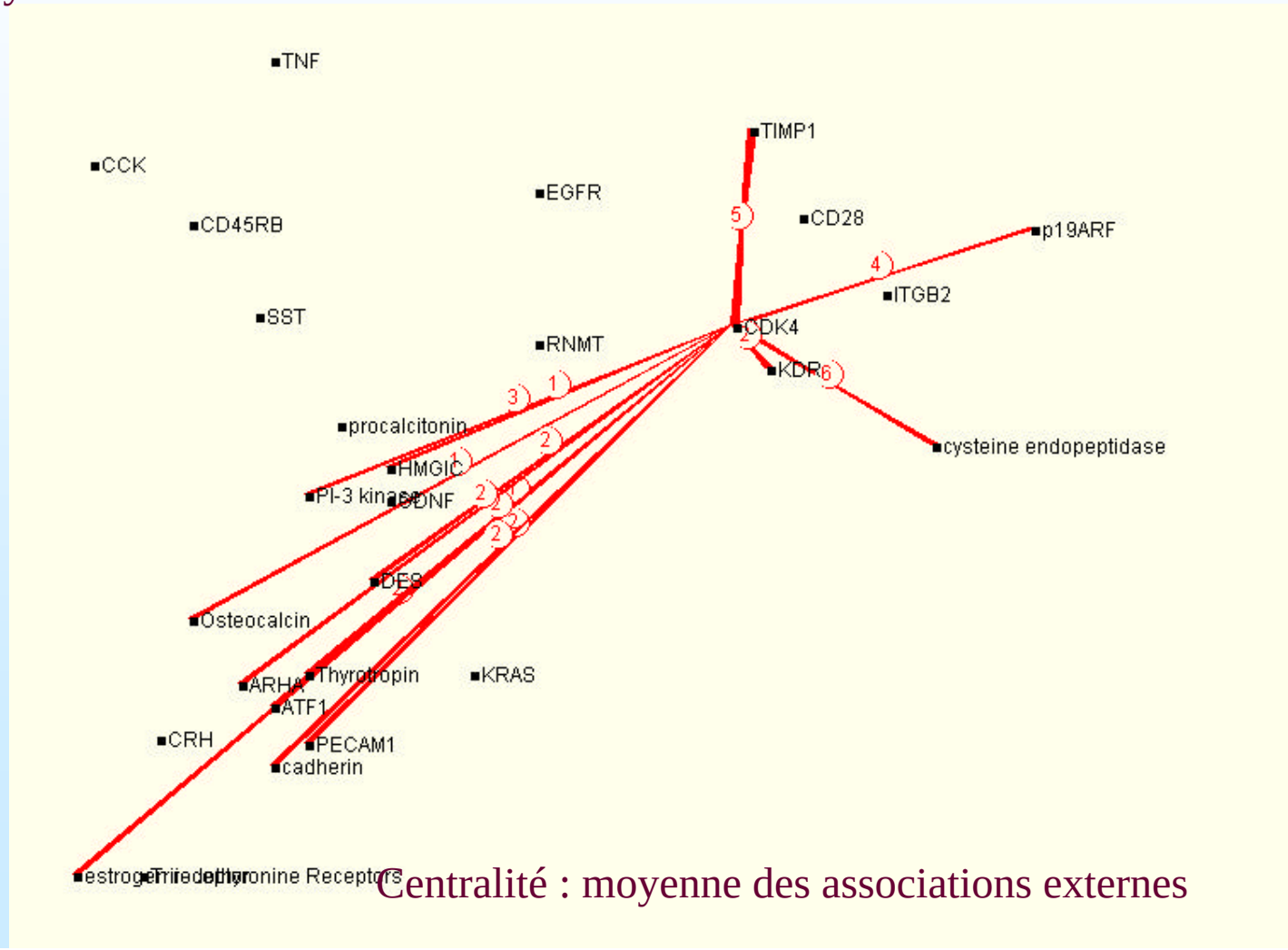




atlas.irit.fr

Mots associés : cartographie

Densité : moyenne des associations internes

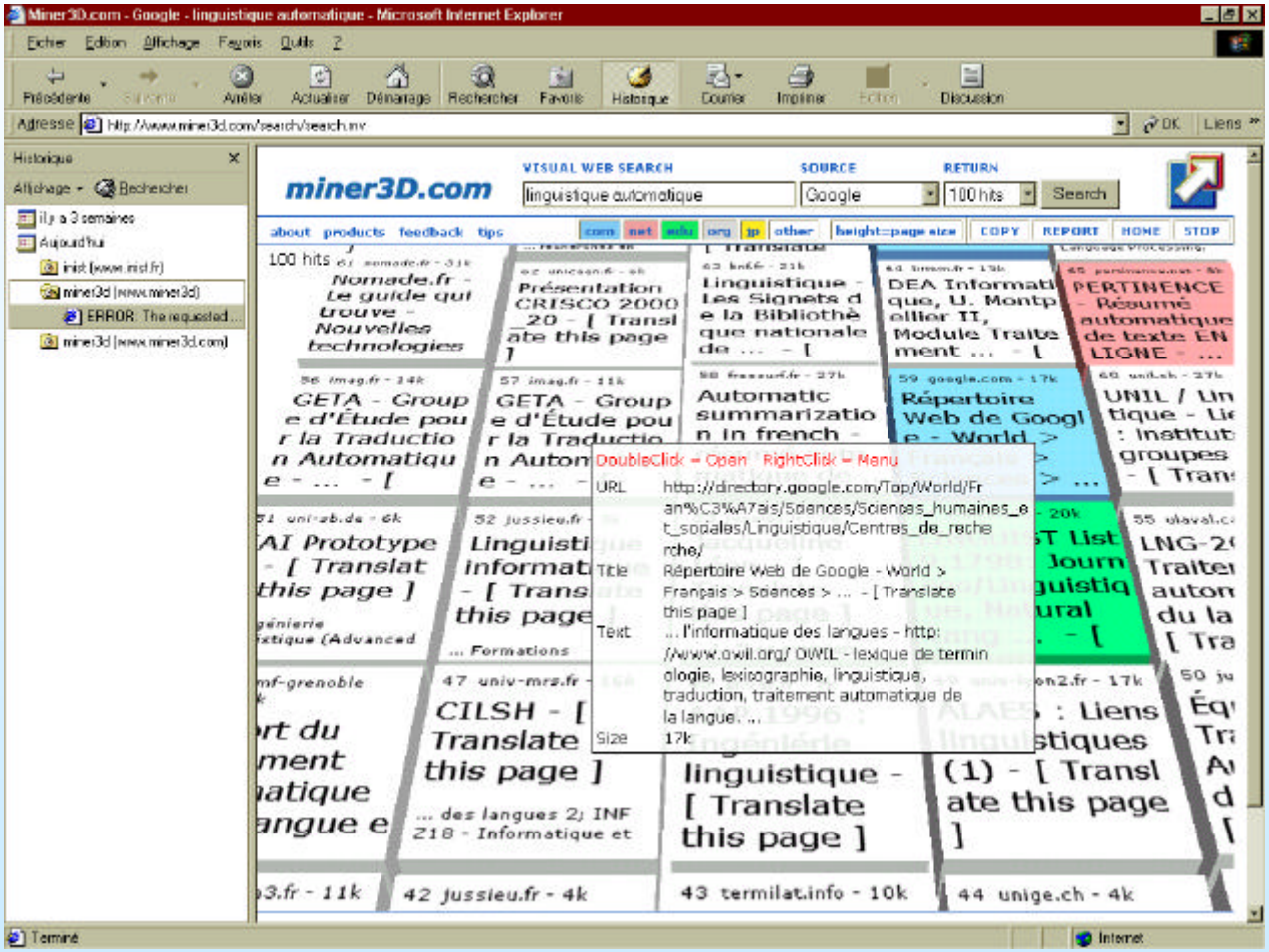


SDOC

Centralité : moyenne des associations externes

mimer3D

Visualisation des sites sous un pavage 3D coloré

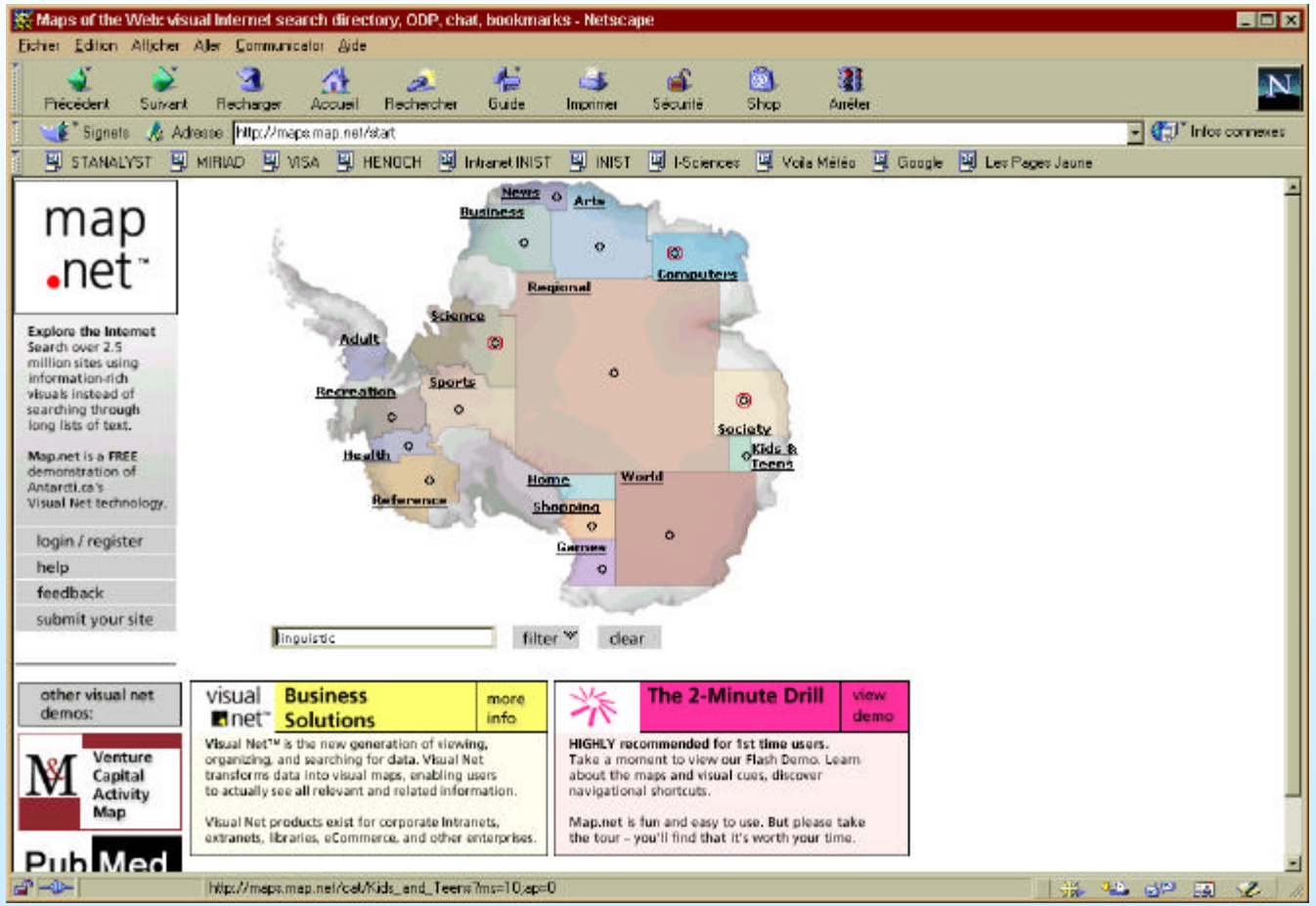


http://www.miner3d.com

Map.net (1)

Navigation dans les classes de premier niveau

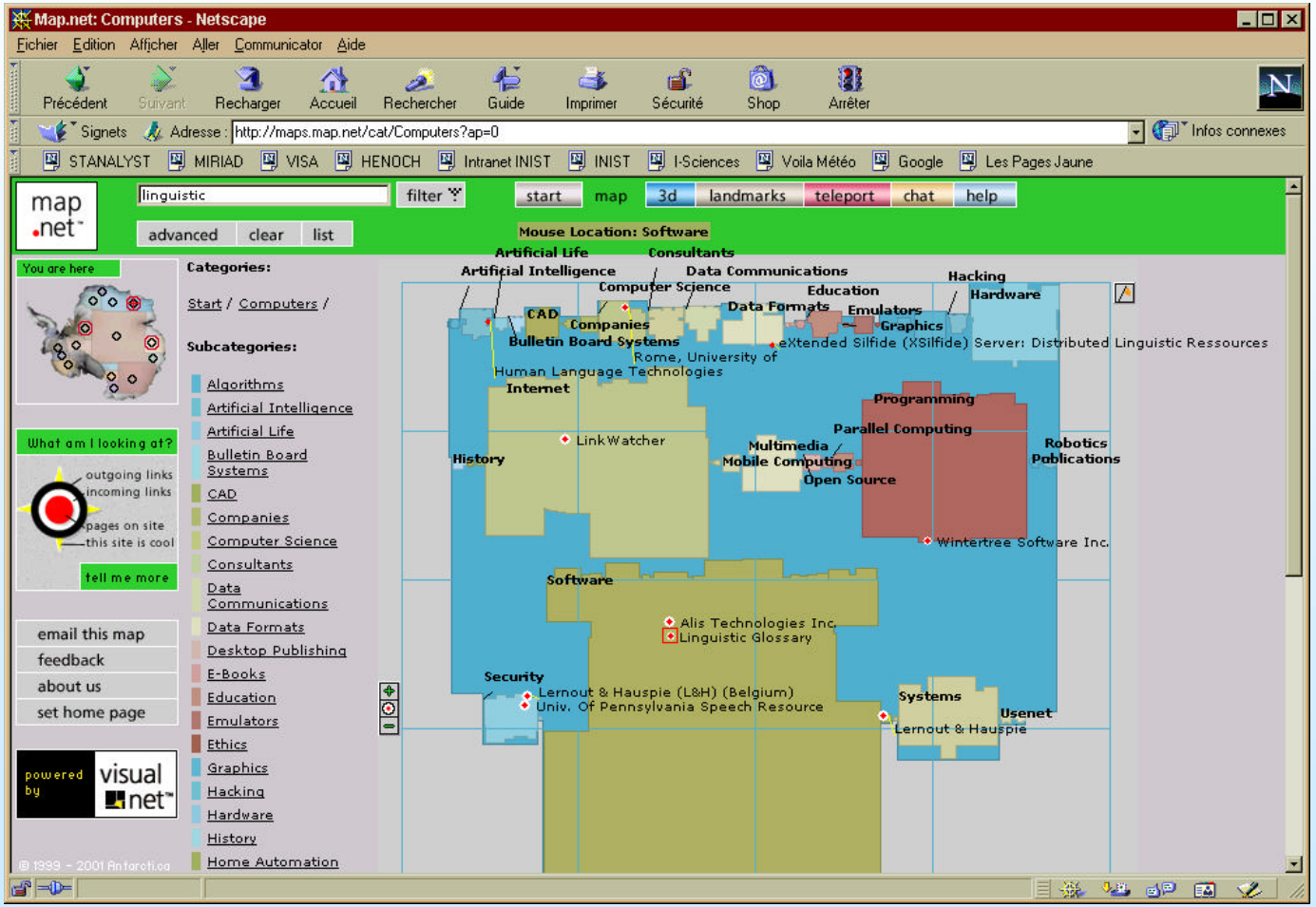
Requête : linguistique



http://maps.map.net

Map.net (2)

Navigation dans les sous-classes d'une classe premier niveau : la catégorie informatique



Intégré dans le site de PubMed

Kartoo

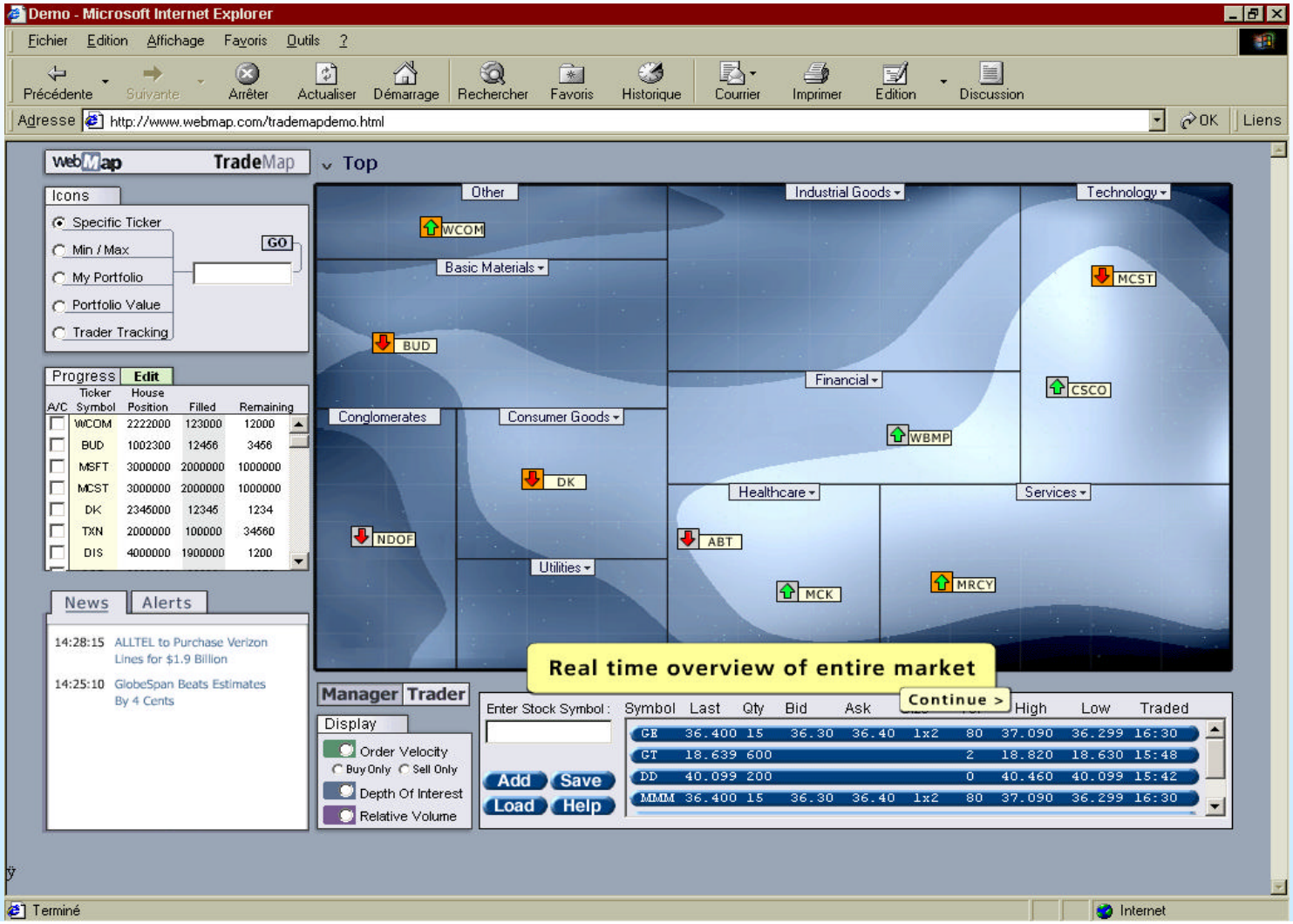
Métamoteur,
Cartographie
de sites

http://www.kartoo.com

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Inist Institut de l'Information Scientifique et Technique

WebMap



CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Claude Aschenbrenner Miste Paris 2001 - 2002

inist Institut de l'Information Scientifique et Technique