

L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique

Luc GRIVEL

INIST/CNRS

CRRM

PLAN DE L'EXPOSÉ (1)

■ DE L'ANALYSE DE L'IST À L'HYPERTEXTE

- L'analyse de l'information scientifique et technique (IST) : contexte et définition
- Le processus d'analyse et la génération automatique d'hypertextes
- Vers un environnement d'analyse de l'IST

PLAN DE L'EXPOSÉ (2)

- UN ENVIRONNEMENT D'ANALYSE DE L'IST
 - La plate-forme infométrique de l'URI
 - SDOC (méthode des mots associés)
 - HENOCH, un générateur d'hypertextes pour analyser l'IST
 - » Ingénierie documentaire
 - » Interface utilisateur

PLAN DE L'EXPOSÉ (3)

- BILAN DES TRAVAUX
- PERSPECTIVES

L'analyse de l'IST

- Contexte social : veille et évaluation de la recherche à partir de la littérature S & T
- Contexte institutionnel : l'URI de l'INIST
- Contexte disciplinaire : l'infométrie

L'analyse de l'IST

Objectif

Caractériser un ensemble documentaire sur le plan cognitif et factuel, de façon à pouvoir en dégager le sens ou les aspects stratégiques.

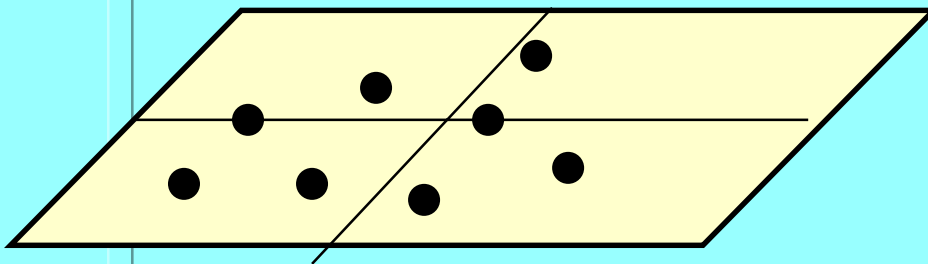
(‘Qui fait quoi, où, collabore avec qui, quand ?’)

Technologies contributives : linguistique, classification et cartographie

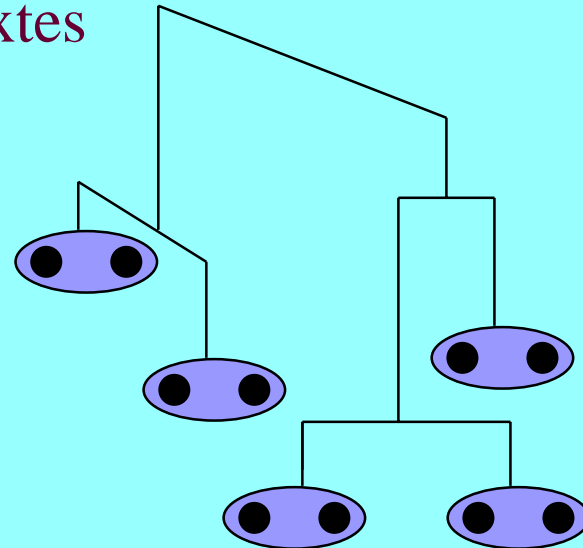
data storage
= *storage of data*

$$(N \rightarrow N_2 N_1) = (N_1 \text{ of } N_2)$$

Extraction de termes à partir de textes



Méthodes factorielles



Méthodes de classification

Analyse de l'IST (définition opérationnelle)

- **Traitement automatique du langage naturel,**
- **Classification automatique**
- **Représentation graphique (cartographie)**

De l'analyse de l'IST à l'hypertexte

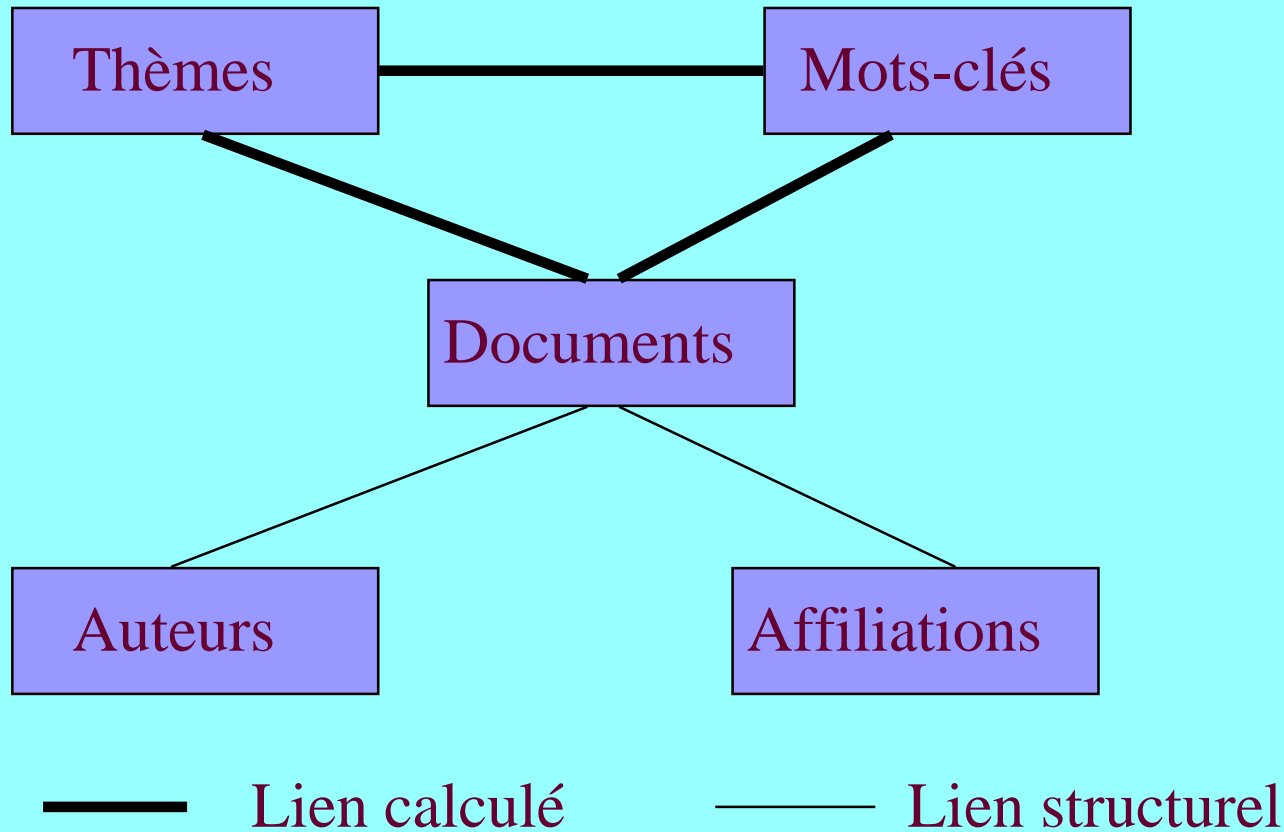
■ Le processus d'analyse

- exploration intuitive, par association d'idées
- exploration méthodique, tenant compte des caractéristiques des méthodes d'analyse utilisées,
- soutenue par un questionnement : Qui fait quoi, où, ... ?

De l'analyse de l'IST à l'hypertexte

- Disposer d'une **carte**, de méthodes pour faire le point, se **positionner** et
- **S'interroger** 'Qui fait quoi, où, quand, ... ?'
- **Modéliser** cette organisation, ces interactions
- → **L'hypertexte**

La génération automatique d'hypertextes



Principes directeurs

- **Explorer** → une carte
- **Positionner** → partir de la question générique «Qui fait quoi, où, ...» → **croiser les informations, calcul d'indicateurs**
- **Combiner ces deux principes**

Vers un environnement d'analyse de l'IST

- Une plate-forme d'outils d'analyse + une base infométrique,
- Stocker, explorer et interroger méthodiquement à travers une interface conviviale

2ème Partie

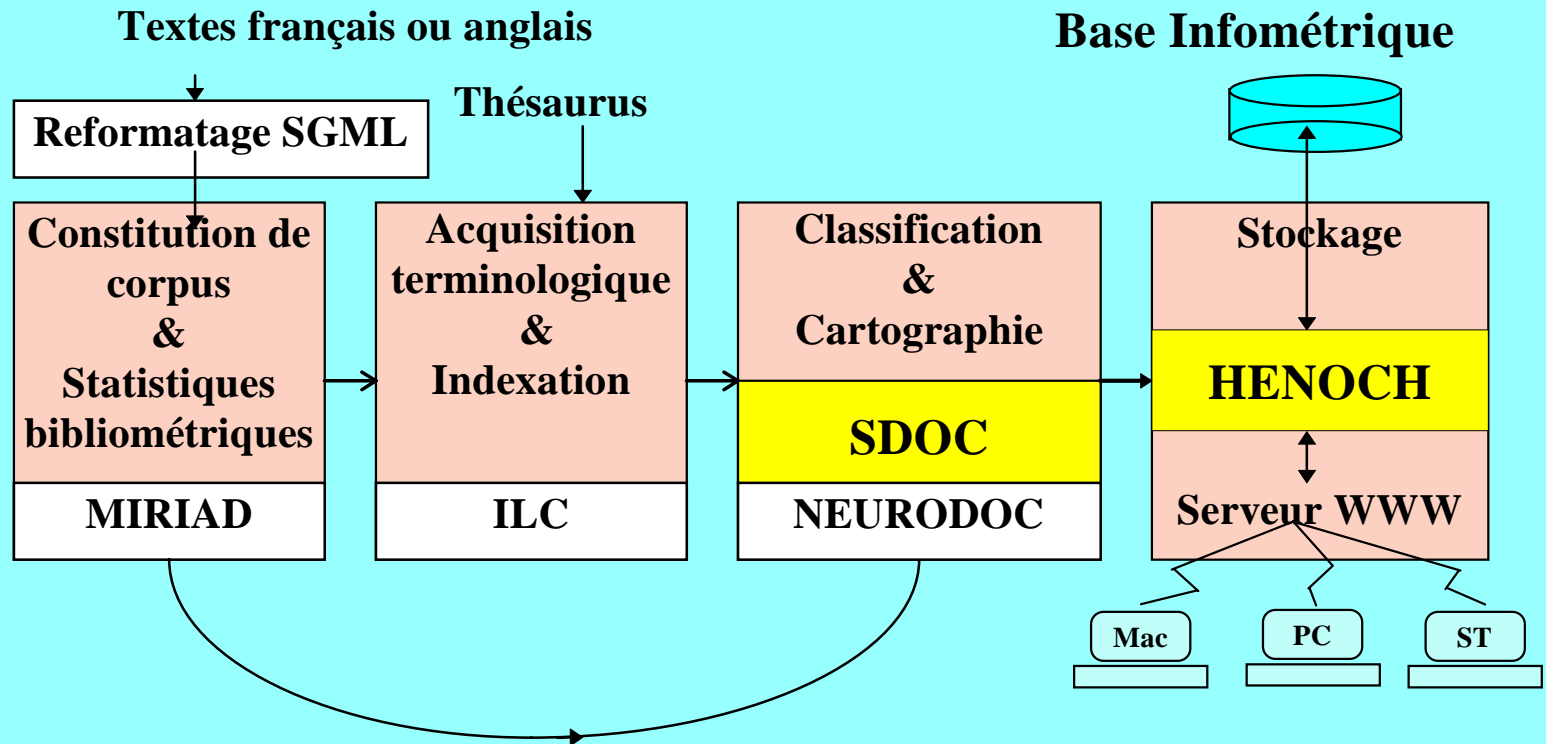
- UN ENVIRONNEMENT D'ANALYSE DE L'IST
 - La plate-forme infométrique de l'URI
 - SDOC (méthode des mots associés)
 - HENOCH, un générateur d'hypertextes pour analyser l'IST
 - » Ingénierie documentaire
 - » Interface utilisateur

La plate-forme infométrique de l'URI

(Contexte opérationnel)

- des mécanismes **d'extraction terminologique** sur du texte intégral en anglais et en français --> mots-clés
- des techniques de **tris simples ou croisés** (statistiques descriptives fondées sur les distributions bibliométriques),
- des techniques de **classification hiérarchique et non hiérarchique** et des techniques de **cartographie** (ACP, diagramme stratégique, réseaux neuronaux) pour la structuration de l'information.
- des techniques **d'ingénierie documentaire**.

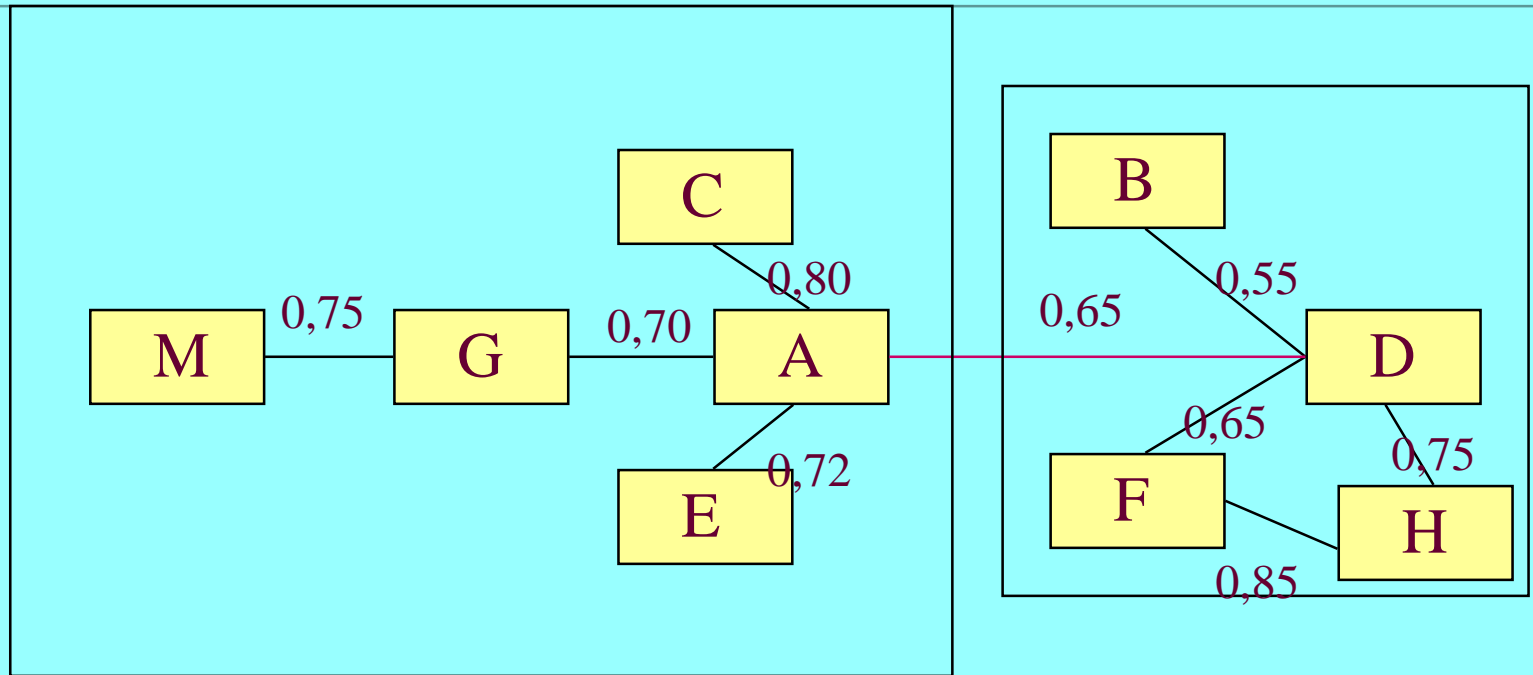
La plate-forme infométrique



Mots associés

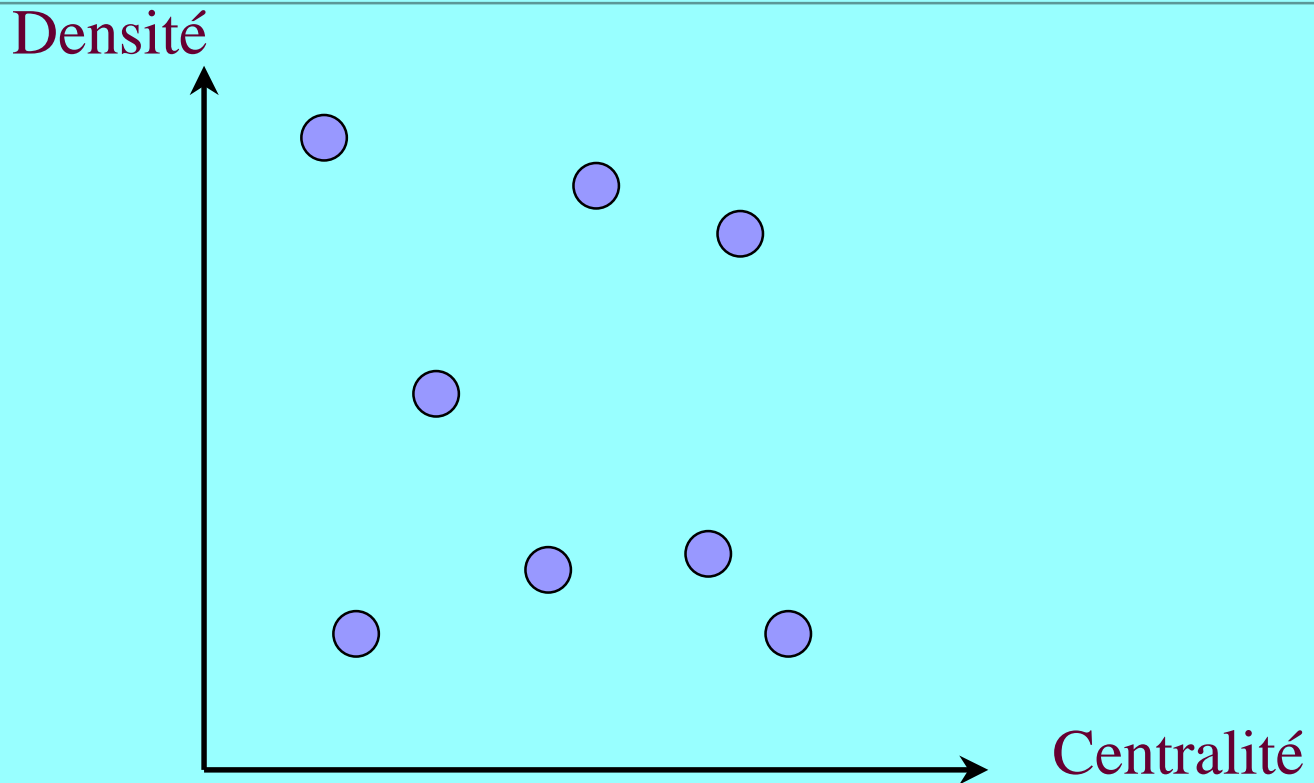
- **Cooccurrence des mots-clés**
- **Indice statistique : ex. $E_{ij} = C_{ij}^2 / C_i C_j$**
- **Découpage du réseaux d'associations en clusters (agrégats) de mots-clés**

Classification



Réseau d'associations découpé en deux clusters
de 5 mots maximum

Cartographie



L'outil SDOC, au service de l'analyse de l'IST

■ technologique :

- conception modulaire par décomposition en programmes indépendants,
- adoption de standards (SGML)

■ conceptuel :

- paramétrage de l'outil,
- production d'indicateurs, démarche d'analyse

Tableau caractéristique des clusters

[1]:Seuil de saturation, [2]:densité, [3]:centralité, [4]:Nombre de mots-clés internes, [5]:Nombre de mots-clés externes, [6]:Nombre d'associations internes, [7]:Nombre d'associations externes avec d'autres clusters, [8]:Nombre de citations du cluster par d'autres clusters, [9]:Nombre de docs construisant le cluster/Nombre de docs cluster, [10]:Nombre de docs propres au cluster.

Nom	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
...										
Analyse statistique	0.062	0.082	0.041	16	2	26	2	3	58/90	20
Enseignement superieur	0.067	0.127	0.051	14	3	25	3	6	21/43	8
Organisation travail	0.067	0.156	0.065	14	4	16	4	6	26/31	4
Reproduction document	0.071	0.075	0.075	13	4	25	5	10	45/73	16
Fourniture electronique document	0.083	0.072	0.073	13	4	34	4	25	67/84	3



sommaire



carte



thèmes



revues



congrès



organismes



auteurs

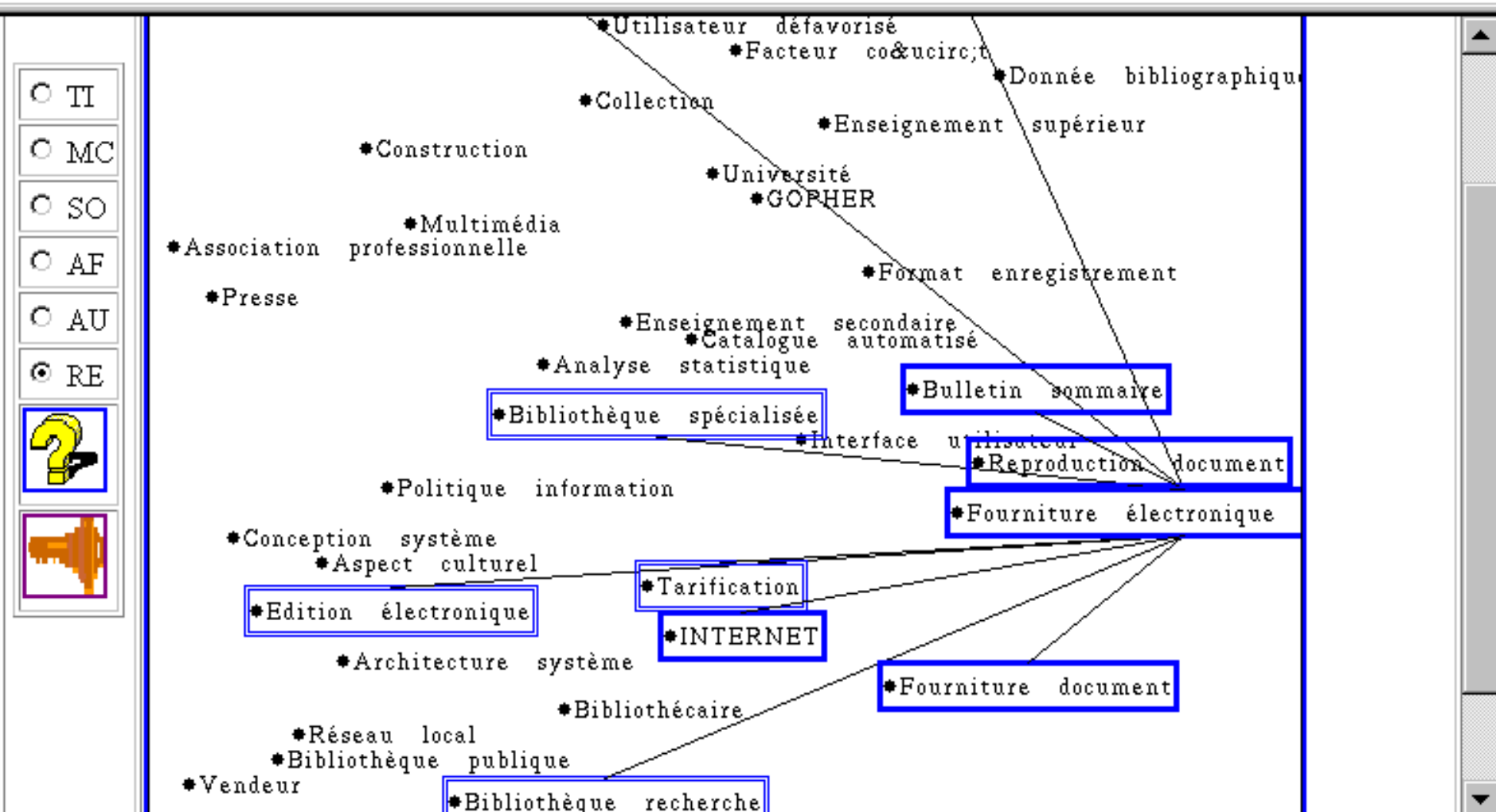


mots-clés

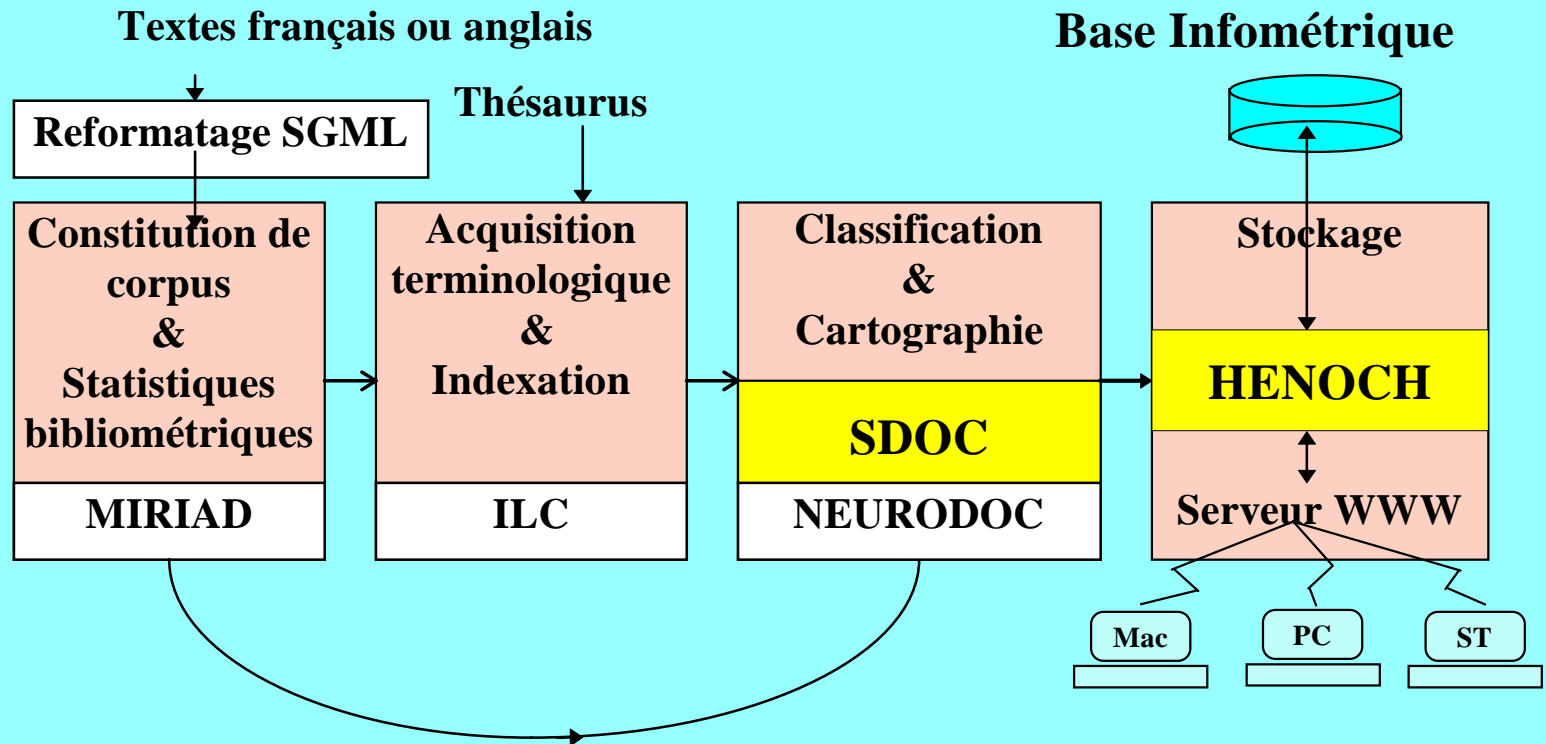


Aide

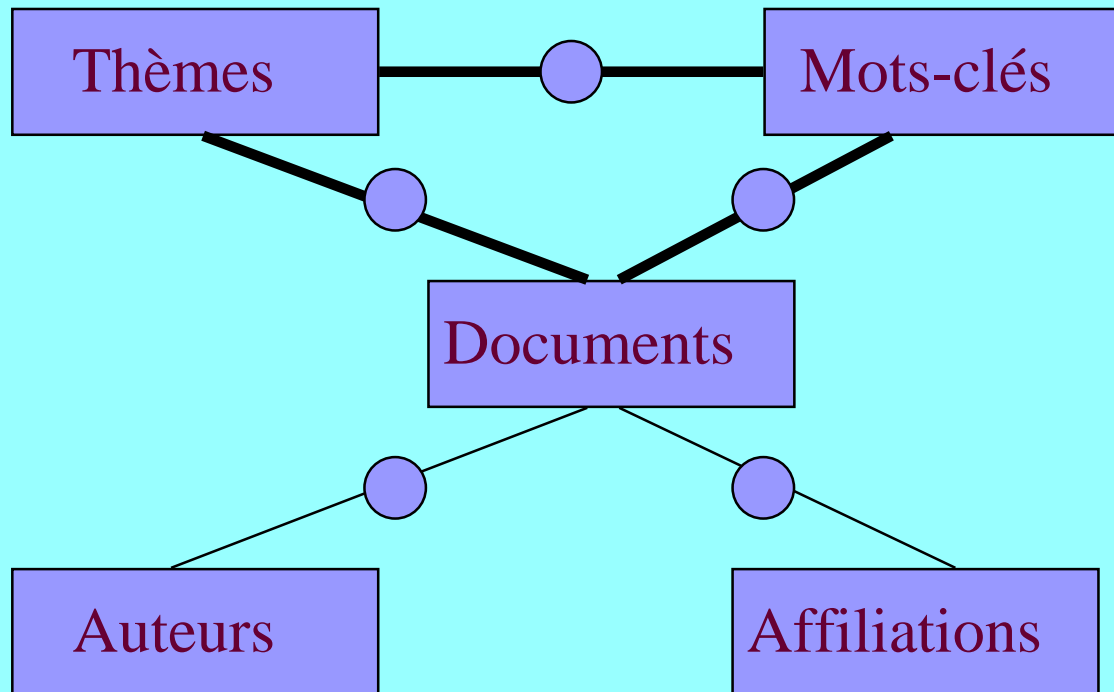
- TI
- MC
- SO
- AF
- AU
- RE
-
-



La plate-forme infométrique

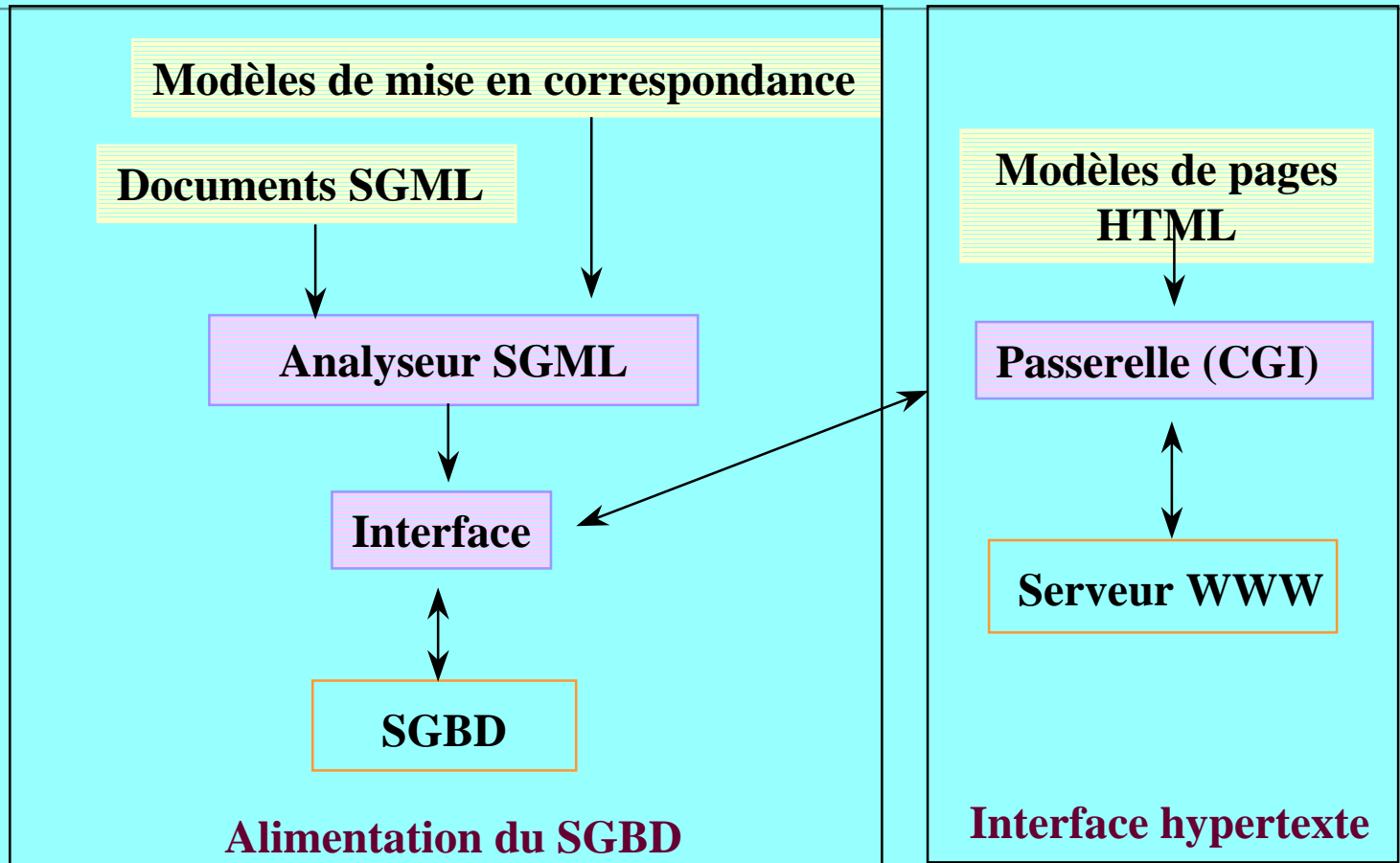


Modèle relationnel



● Relations n-m

Architecture d'HENOCH : SGBD, WWW, SGML



Document SGML

<record>

<NO>12508319 </NO>

**<TI>AMYOTROPHIC-LATERAL-SCLEROSIS AND STRUCTURAL
DEFECTS IN CU,ZN SUPEROXIDE-DISMUTASE </TI>**

**<AU>DENG HX</AU><AU>HENTATI A</AU><AU>TAINER
JA</AU><AU> IQBAL Z</AU><AU> CAYABYAB A</AU><AU>
HUNG WY</AU><AU> GETZOFF ED</AU>...**

<AF>

**<NA> NORTHWESTERN UNIV,SCH MED,DEPT NEUROL,300 E
SUPERSTNEUROL</NA><TO>CHICAGO</TO><CO>IL</CO>**

</AF> ...

</record>

Un fichier de mise en correspondance

Variable	Chemin d'accès	Symbole d'occurrence
Name	record/AF/NA	repeat
Town	record/AF/TO	repeat
Country	record/AF/CO	repeat

query :

begin

/* the insertion procedure to execute */

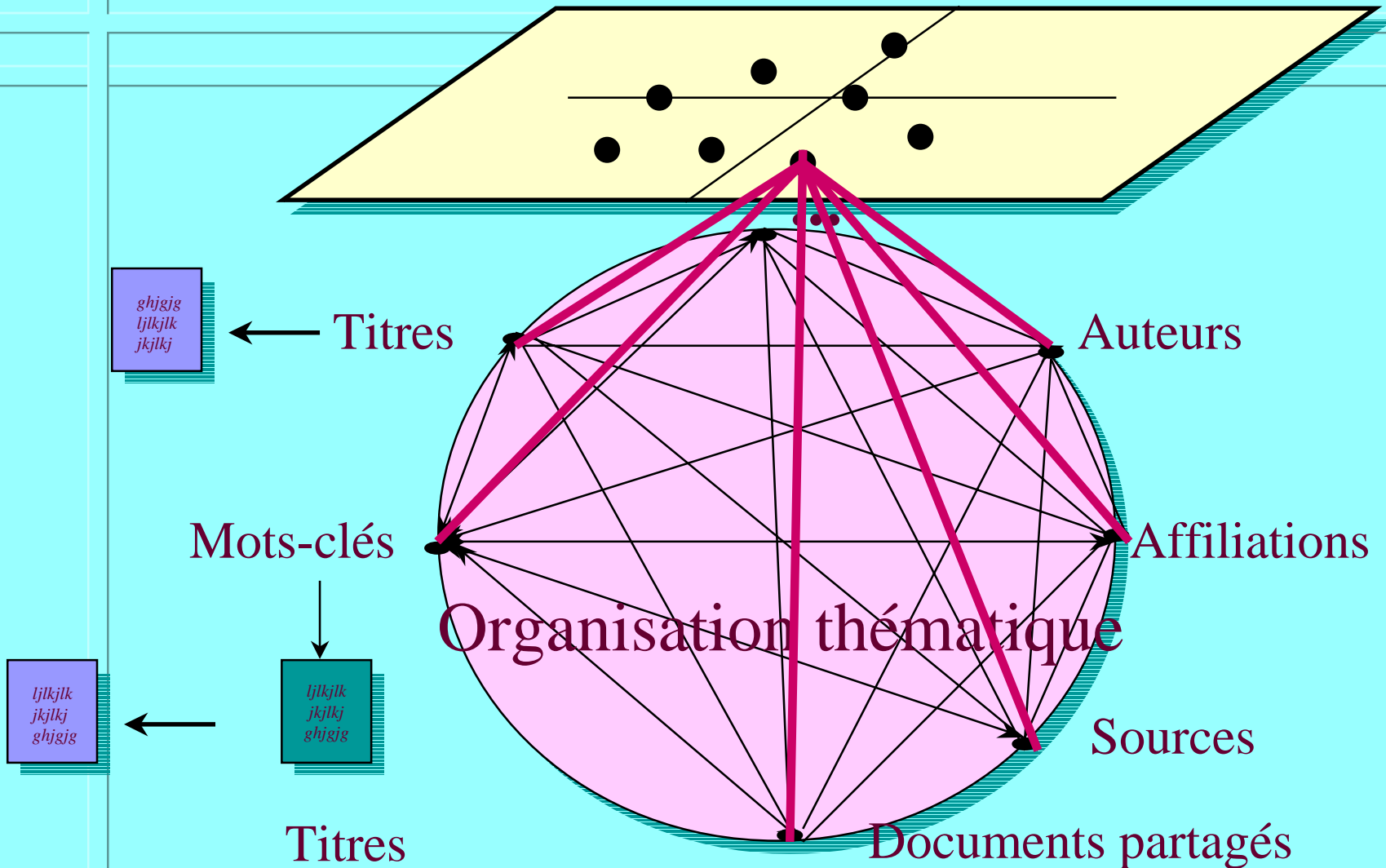
INS_AFFILIATION(:{NAME}, :{TOWN}, :{COUNTRY})

end;

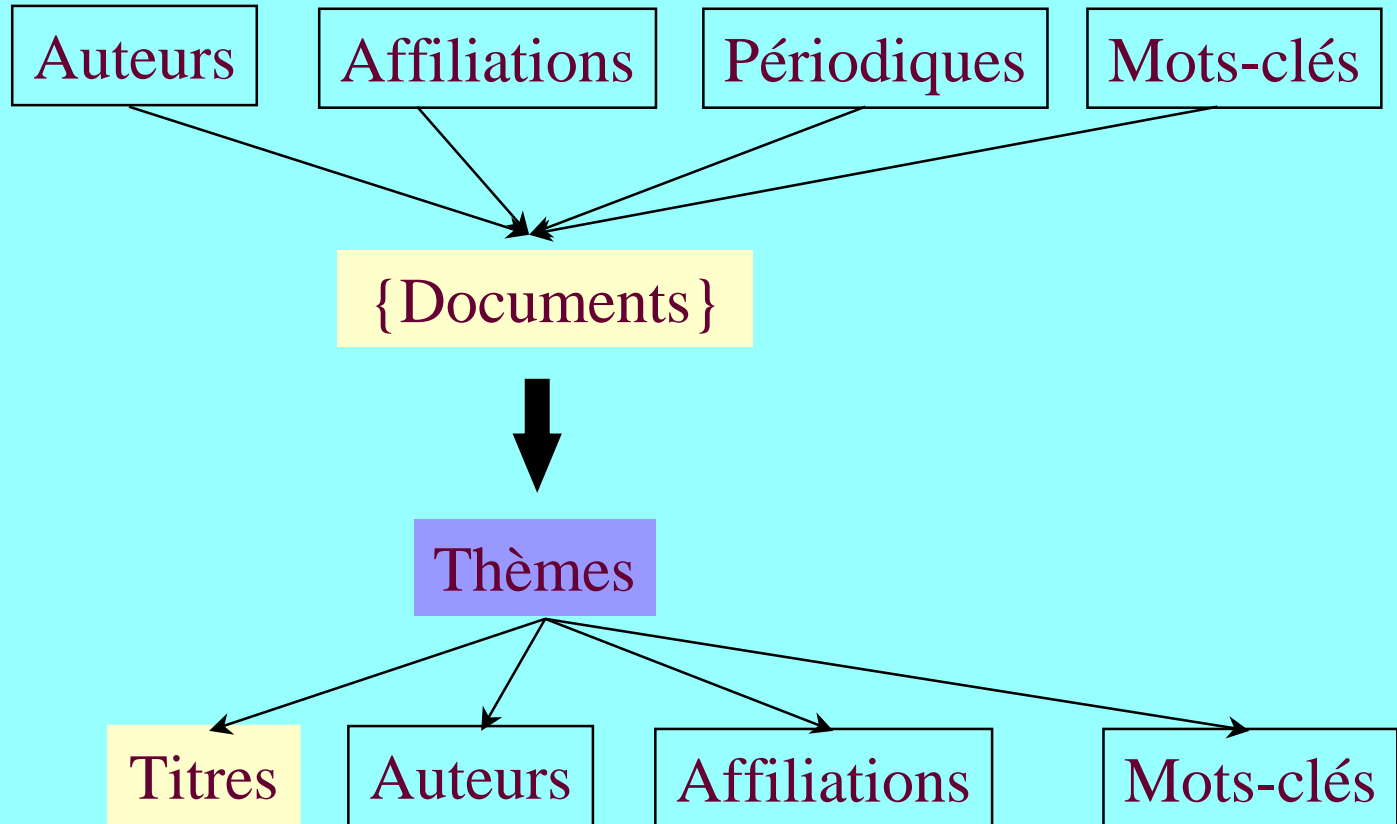
Objectifs de l'interface

- Avoir une vue d'ensemble du corpus et de son **organisation thématique et factuelle**,
- Evaluer le **positionnement thématique** d'un auteur, d'une institution, ...

Une interface métaphorique



Des fonctions de positionnement



The screenshot shows a web browser window titled "automatisation-parSDOC - Note". The menu bar includes "File", "Edit", "View", "Go", and "Commu". A purple callout bubble at the top right contains the text "Fonctions de positionnement". Below the menu is a navigation bar with icons and labels: "sommaire", "carte", "thèmes", "revues", "congrès", "organismes", "auteurs", "mots-clés", and "Aide". A blue oval highlights the "revues", "congrès", "organismes", and "mots-clés" icons. The main content area contains the following text:

Dans ce corpus nommé **automatisation-parSDOC**, il y a **39** thèmes où se répartissent **1763** documents sur un total de **1839** documents. Le nom du corpus rappelle dans son suffixe le nom de la méthode utilisée pour la classification. Pour visualiser les thèmes, vous pouvez cliquer sur la carte thématique ou le tableau des thèmes.

Below the text are two small icons: a landscape icon labeled "carte" and a table icon labeled "thèmes". A purple callout bubble on the right side of the page contains the text "Voies d'exploration". The status bar at the bottom shows "Document: Done" and a system tray with various icons.



sommaire



carte



thèmes



revues



congrès



organismes



auteurs



mots-clés



Aide

TI

MC

SO

AF

AU










RE



- *Donnée bibliographique
- *Collection
- *Enseignement supérieur
- *Construction
- *Université
- *GOPHER
- *Multimédia
- *Association professionnelle
- *Format enregistrement
- *Presse
- *Enseignement secondaire
- *Catalogue automatisé
- *Analyse statistique
- *Bulletin sommaire
- *Bibliothèque spécialisée
- *Interface utilisateur
- *Reproduction document
- *Politique information
- *Fourniture électronique
- *Conception système
- *Aspect culturel
- *Tarification
- *Edition électronique
- *INTERNET
- *Architecture système
- *Fourniture document
- *Bibliothécaire
- *Réseau local
- *Bibliothèque publique
- *Vendeur
- *Bibliothèque recherche

automatisation-parSDOC - Netscape

File Edit View Go Communicator Help

[mots-clés](#)
 [84 titres](#)
 [61 affiliations](#)
 [104 auteurs](#)
 [49 sources](#)

[Documents partagés avec d'autres thèmes](#)

[A propos des thèmes SDOC](#)

Description du thème '*Fourniture électronique document*'

Poids	Fréquence relative / Fréquence globale	Mots-clés
.26	23 /24	Fourniture électronique document
.24	13 /13	Article
.21	28 /39	Opération pilote
.21	20 /30	Périodique
.18	22 /30	Journal électronique
.18	15 /20	Editeur
.16	16 /20	Littérature scientifique
.16	13 /16	Commande document

Document Done

Typologie thématique d'une revue (1)

10 janvier

The screenshot shows a Netscape browser window titled "automatisation-parSDOC - Netscape". The menu bar includes "File", "Edit", "View", "Go", "Communicator", and "Help". The address bar is empty. Below the menu bar is a navigation bar with icons and labels: "sommaire", "carte", "thèmes", "revues", "congrès", "organismes", "auteurs", "mots-clés", and "Aide".

The main content area displays the heading "Recherche par periodiques sur le corpus". Below this is a search form with the label "Nom du periodique" and a "Filtrer" button. A list of periodicals is shown in a scrollable area, with "OK" and "Déselectionner" buttons above it.

Nom du periodique

OK Déselectionner

- ABI - Technik
- ACM transactions on mathematical software
- Advances in librarianship
- Alexandria : (Aldershot)
- Annals of library science and documentation
- Annual review of information science and technolog
- Annual review of OCLC research
- Arbido
- ARBIDO - R
- Archimag : (Vincennes)
- Argus : (Montréal)
- Argus: (Montréal)
- Aslib proceedings
- Audiovisual librarian
- Bibliotekovedenie i bibliografija za rubežo
- Bibliotheek- en archiefgids - Vlaamse vereniging v
- Bibliothek
- Bibliotheksdienst
- Biblos : (Wien)
- Bolletino d'informazioni - Associazione italia

The status bar at the bottom shows "Document: Done" and various system icons.

automatisation-parSDOC - Netscape

File Edit View Go Communicator Help

sommaire carte thèmes revues congrès organismes auteurs mots-clés Aide

Typologie thématique de(s) périodique(s)

'Bibliotheksdienst'

Accès aux documents par thématique

OK Déselectionner

- Fourniture document (21)
- Bibliothèque publique (14)
- INTERNET (13)
- Bibliothèque spécialisée (8)
- Langage documentaire (8)
- Catalogue automatisé (7)
- Bibliothécaire (6)
- Fourniture électronique document (6)
- Tarification (6)
- Analyse statistique (4)

- [1. A test of the acquisition system SIERA 2.0 in the academic library of Erlangen- Nürnberg](#)
- [2. From union catalog to Gopher searcher](#)
- [3. Neue Strukturen in der Informationsvermittlung and der Universität Freiburg](#)
- [4. Endspurt - Stand der Einführung der EDV bei den Hamburger Öffentlichen Bücherhallen](#)
- [5. Automation of special library](#)
- [6. Erfahrungen bei der Anwendung von BIS-LOK in der Stadtbibliothek «Heinrich Heine» Halberstadt](#)

Document Done

Conclusion de l'exposé

- BILAN DES TRAVAUX
- PERSPECTIVES

Bilan opérationnel

- **SDOC → Etudes de veille, outil de recherche à l'INRIA Lorraine**
- **HENOCH → consulté par des partenaires de l'INIST (CNRS, BVD, CVT, ...) et support d'enseignement de la veille technologique**

Evaluation interne



- Interface agréable, pas d'apprentissage de commandes, une vue d'ensemble du corpus et de son organisation thématique et factuelle



- La détection et l'analyse des évolutions thématiques dans le temps

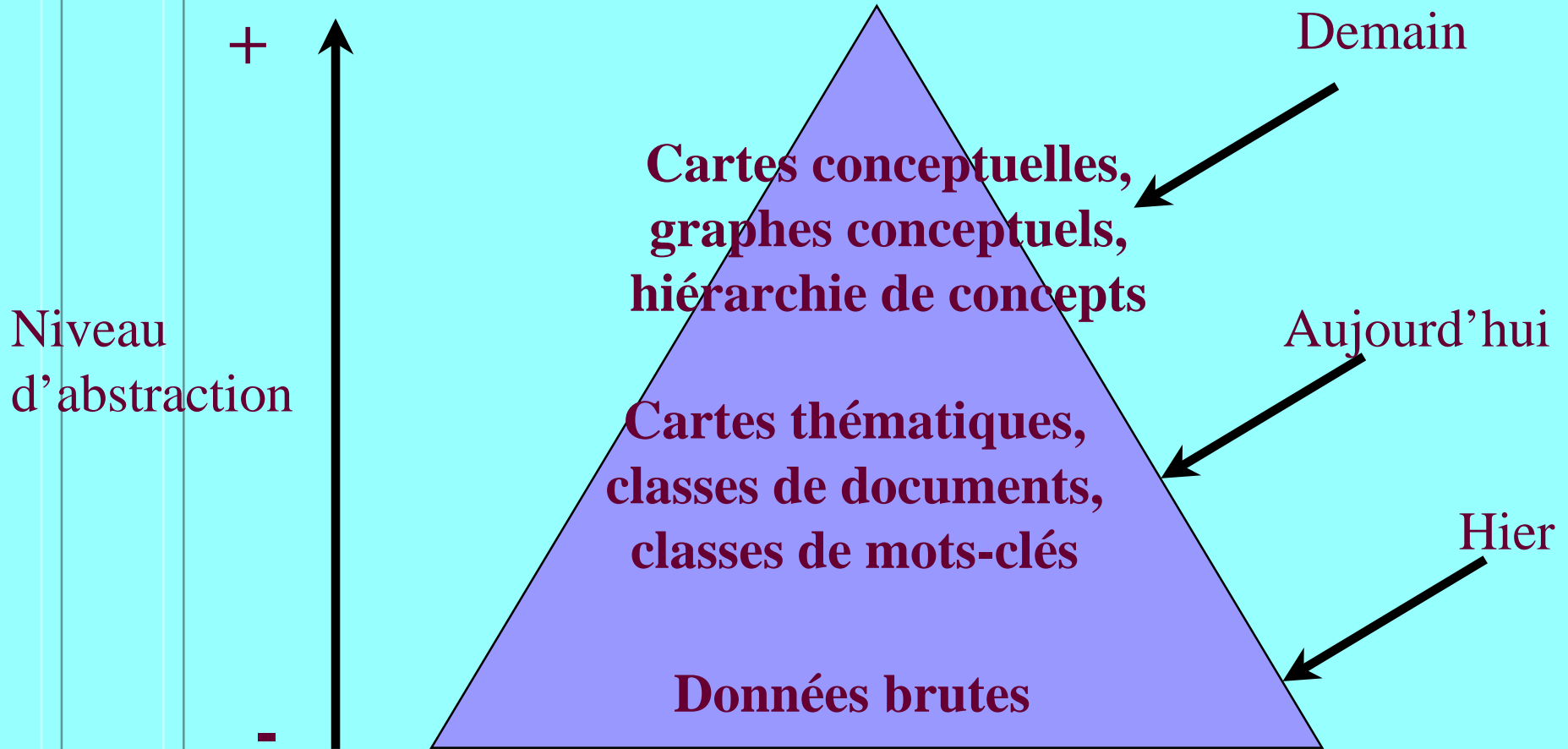
Applications

- Aide à la construction de vocabulaires, plan de classement, thésaurus, bases de connaissances partielles
- Aide à la découverte et à l'analyse stratégique d'un domaine de recherches ou d'applications tel qu'il transparaît à travers un corpus de documents

Conclusion

- **Assister l'interprétation par un opérateur humain des résultats d'une plate-forme d'analyse de l'IST**
- **Séparer clairement l'aspect 'traitement de l'IST' des aspects 'stockage' et 'visualisation'**

Perspectives



Perspectives

- Augmenter le niveau d'abstraction,
- Capitaliser le travail intellectuel, ...
- → Des technologies avancées (XML, SRCO, etc.)
- → **Une recherche de nature transversale**