INFORSID

Informatique des organisations
et systèmes d'information et de décision

Extraction, gestion de connaissance et Web (Amedeo Napoli, LORIA)

Tutoriel associé à Inforsid
Nancy, 3 juin 2003

INFORSID-2003

Extraction de connaissances,

gestion de connaissances et Web sémantique

Amedeo Napoli

Équipe Orpailleur

LORIA – UMR 7503

BP 239, 54506 Vandœuvre-lès-Nancy Cedex

(Email : Amedeo.Napoli@loria.fr)

http://www.loria.fr/ napoli/

http://www.loria.fr/LORIA/EXT/equipes/ORPAILLEUR/

Plan du cours

- Introduction générale à l'ECBD.
- Techniques symboliques d'ECBD : classification par treillis, extraction de motifs fréquents et de règles d'association.
- Un exemple de fouille de base de données de réactions en chimie organique.
- Quelques mots sur la fouille de textes.
- L'ECBD dans le cadre du Web sémantique.
- ...

Vers l’extraction de connaissances

dans les bases de données (ECBD)

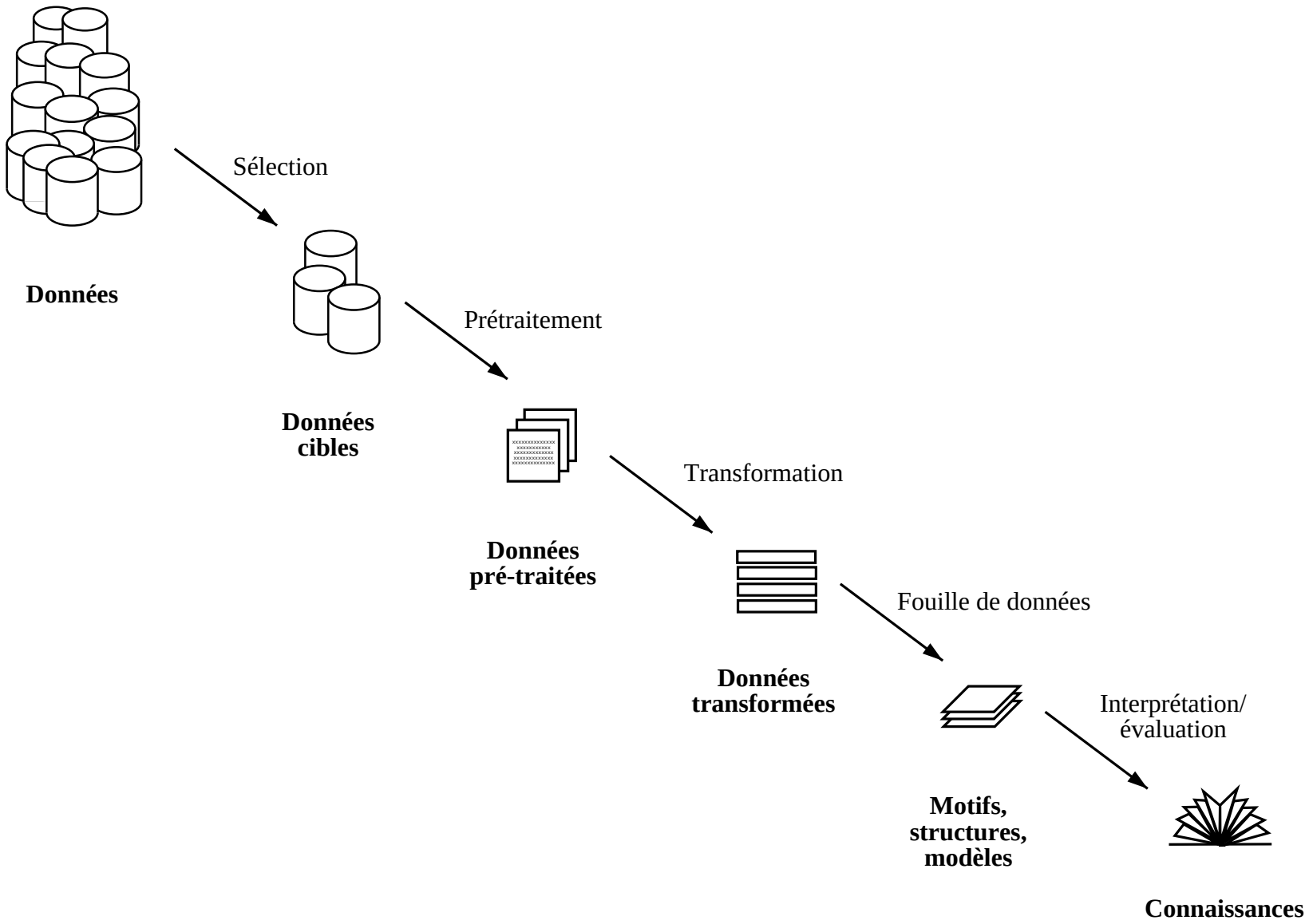
- Des données — documents sur le Web par exemple — sont disponibles en quantité importante et en qualité variable, sans utilisation particulière et précise a priori.
- Une question fondamentale est savoir s’il est possible d’extraire « quelque chose d’intéressant » de ces grandes bases de données, et comment.
- Une question parallèle est de pouvoir « manipuler les documents par leur contenu » : rechercher et classifier les documents, extraire des informations, exploiter les contenus dans des raisonnements, ...

Un exemple de requêtes sur le Web

- *Un livre sur Cartier-Bresson*.
- *Une biographie de Cartier-Bresson*.
- *Une autobiographie de Cartier-Bresson*.
- *Un livre écrit par Cartier-Bresson*.
- *Un livre illustré par Cartier-Bresson*.
- *Un livre de photos de Cartier-Bresson*.
- ...

Une définition du processus d’ECBD

- Le but du *processus d’ECBD* est d’extraire dans des grands volumes de données des *éléments de connaissances non triviaux* et *nouveaux* pouvant avoir un *sens* et un *intérêt* pour être réutilisés.


Sélection
Données
Prétraitement
Données cibles
Transformation
Données pré-traitées
Fouille de données
Données transformées
Interprétation/ évaluation
Motifs, structures, modèles
Connaissances

Quelques méthodes numériques en fouille de données

- Les méthodes statistiques et les méthodes d'analyse des données.
- Les modèles de Markov cachés d'ordre 1 et 2 (HMM 1 et 2), conçus et mis au point à l'origine pour la reconnaissance de formes (parole, image, caractères) : classification d'entités spatio-temporelles par recherche de régularités.
- Les réseaux bayésiens pour la recherche de causalités.
- Les réseaux de neurones.
- Les algorithmes génétiques.
- ...

Quelques méthodes symboliques en fouille de données

- La classification par arbres de décision.
- La classification par treillis.
- La recherche de motifs fréquents et l'extraction de règles d'association.
- Les méthodes inductives en apprentissage : à partir d'instances, à partir d'exemples, ou à partir de cas.
- Les méthodes de recherche d'information et d'interrogation de bases de données.
- Les ensembles approximatifs (*rough sets*) qui sont des ensembles qui s'expriment par l'intermédiaire d'une borne inférieure et supérieure, et d'un intérieur, structures à partir desquelles il est possible d'extraire des règles entre les propriétés des éléments des ensembles.
- ...

Éléments de bibliographie

- J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.
- D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, The MIT Press, Cambridge (MA), 2001.
- Machine Learning and Data Mining, R.S. Michalski, I. Bratko and M. Kubat editors, John Wiley & Sons LTD, Chichester, 1998.
- I.H. Witten and E. Franck, Data Mining, Morgan Kaufmann Publishers, San Francisco, California, 2000 (Practical machine learning tools and techniques with Java implementations – Weka `http://www.cs.waikato.ac.nz/ ml/weka/`).
- A. Cornéjuols et L. Miclet, Apprentissage artificiel : concepts et algorithmes, Eyrolles, Paris, 2002.
- ...

Une base de données formelle (contexte formel)

Objets / Propriétés	P1	P2	P3	P4	P5
O-1	1	0	0	0	1
O-2	1	1	1	1	1
O-3	1	0	1	0	0
O-4	0	0	1	0	1
O-5	0	1	1	1	1
O-6	1	1	1	0	1
O-7	1	0	1	1	1
O-8	1	1	1	0	0
O-9	1	0	0	1	0
O-10	0	1	1	0	1

Exemples d'éléments de connaissances extraits

- $\mathtt{X} = \{\mathtt{P2}, \mathtt{P3}, \mathtt{P5}\}$ est un *motif* dont la *fréquence* est $\phi(\mathtt{X}) = \mathtt{0.4}$.
  $\mathtt{X'} = \{\mathtt{P3}, \mathtt{P5}\}$ est un motif inclus dans $\mathtt{X}$ de fréquence $\phi(\mathtt{X'}) = \mathtt{0.6}$.

  — La règle « $\mathtt{P3\ \ ET\ \ P5} \Longrightarrow \mathtt{P2}$ » peut être extraite à partir des motifs $\mathtt{X}$ et $\mathtt{X'}$, avec l'*interprétation* suivante :

  — « Si un objet possède les propriétés P3 et P5, alors il possède la propriété P2 avec la probabilité $\mathtt{0.66}$. »

  — La *confiance* associée à la règle est $\mathtt{0.66}$: 2/3 des objets possédant P3 et P5 possèdent aussi P2.
- Le motif $\mathtt{Y} = \{\mathtt{P2}, \mathtt{P3}\}$ a pour fréquence $\phi(\mathtt{Y}) = \mathtt{0.5}$.
  Le motif $\mathtt{Y'} = \{\mathtt{P2}\}$ a pour fréquence $\phi(\mathtt{Y'}) = \mathtt{0.5}$.

  — La règle « $\mathtt{P2} \Longrightarrow \mathtt{P3}$ » peut être extraite à partir des motifs $\mathtt{Y}$ et $\mathtt{Y'}$ et a pour confiance $\mathtt{1}$.

Les connaissances dans le processus de l'EBCD

- **La dualité de l'ECBD guidée par les connaissances** :

  $\longrightarrow$ S'appuyer sur des ontologies pour fouiller les bases de données.

  $\longrightarrow$ Fouiller les bases de données pour alimenter des ontologies.

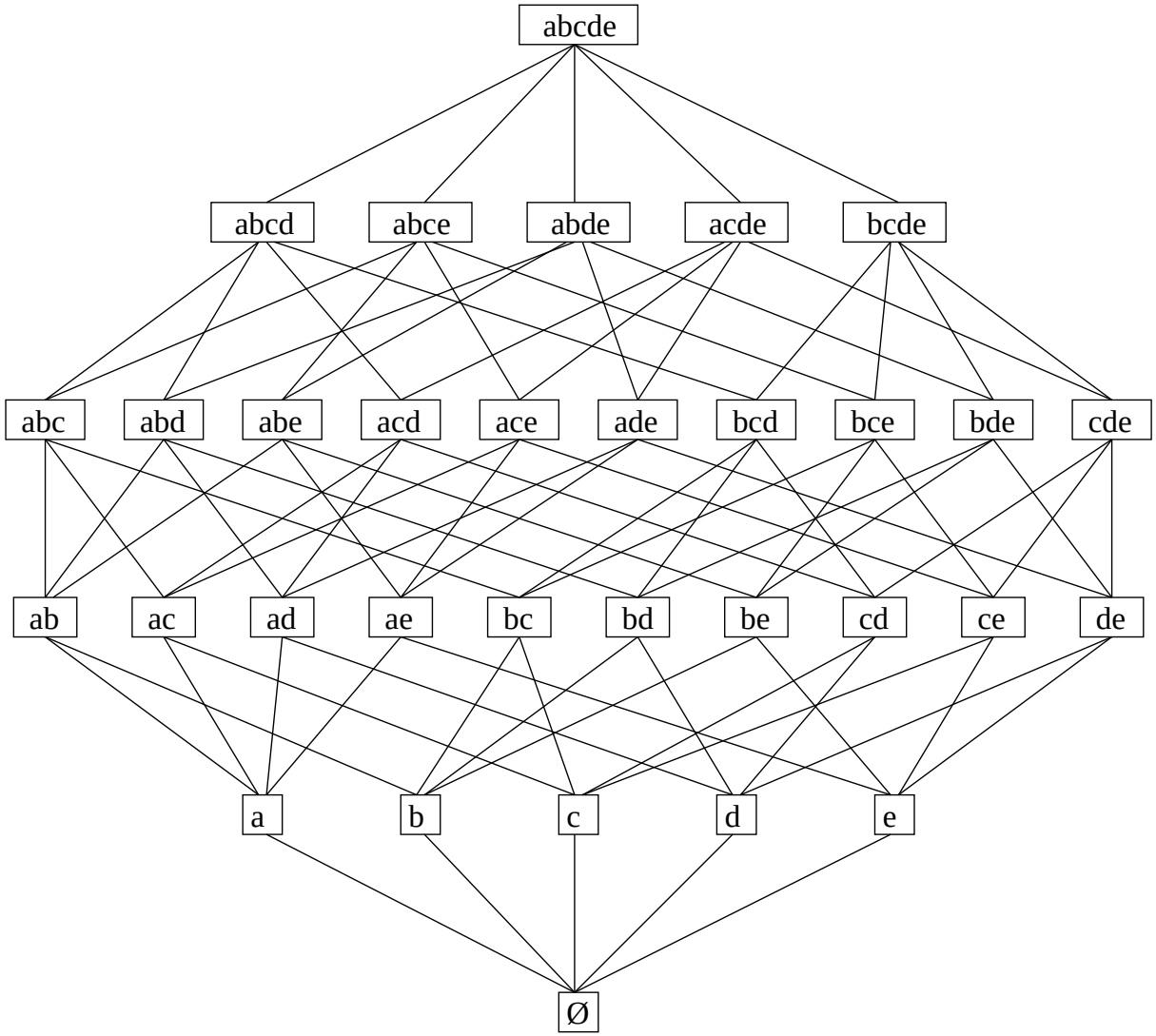
- La représentation des connaissances et l'EBCD sont deux processus complémentaires : *pas de fouille des données sans modèle du domaine des données* !

- Un *analyste* a la charge de contrôler le processus d'ECBD : il oriente — selon ses *propres connaissances* et selon une *représentation du modèle du domaine* (*ontologie*) — le processus et interprète les éléments extraits pour faire émerger des connaissances.

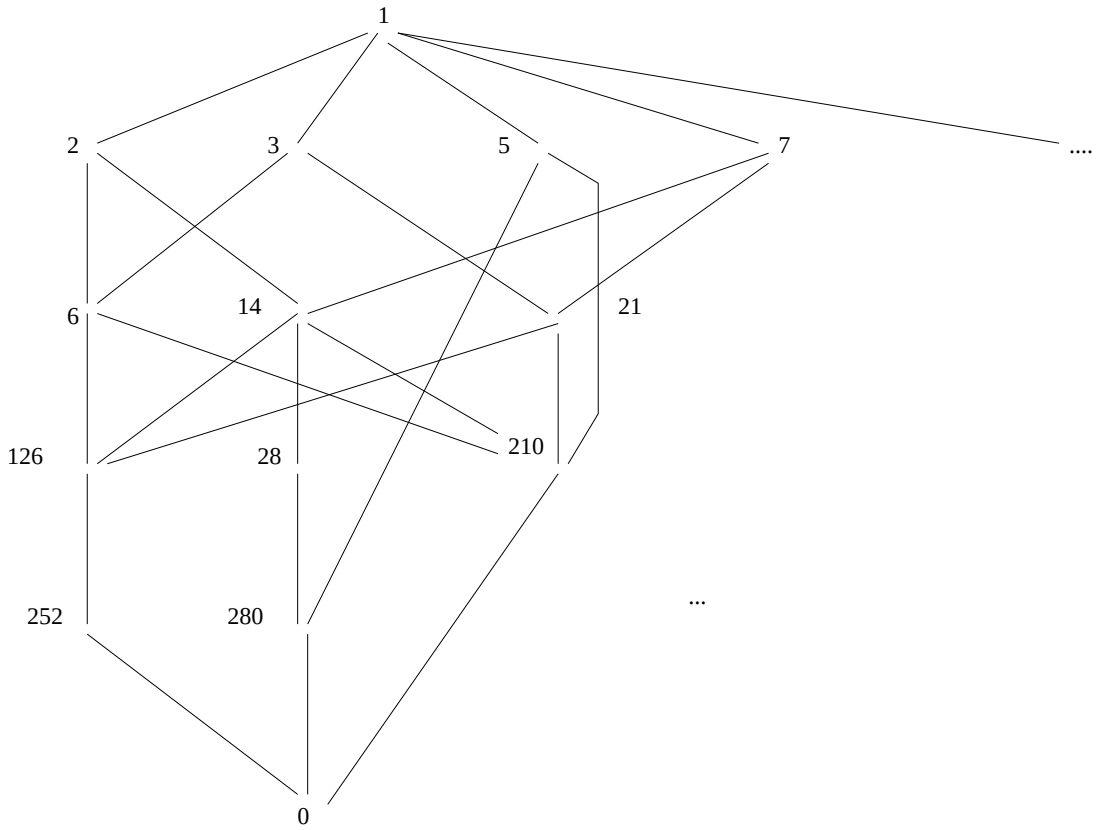
La classification par treillis pour l’ECBD

La notion de treillis

- Un treillis $(\mathtt{E}, \leqslant)$ est un ensemble ordonné tel que chaque couple d'éléments $(\mathtt{x},\mathtt{y})$ possède un *supremum* noté $\mathtt{x} \vee \mathtt{y}$ et un *infimum* noté $\mathtt{x} \wedge \mathtt{y}$.
- L'ensemble $2^{\mathtt{E}}$ des parties d'un ensemble E muni de la relation d'inclusion est un exemple de treillis.
- L'ensemble $\mathbb{N}$ des entiers naturels muni de la relation de divisibilité est un treillis : $\mathtt{x} \sqsubseteq \mathtt{y}$ ssi $\mathtt{y}$ est un diviseur de $\mathtt{x}$ dans $\mathbb{N}$.
  $\mathtt{x} \vee \mathtt{y} = \mathtt{pgcd}(\mathtt{x},\mathtt{y})$ et $\mathtt{x} \wedge \mathtt{y} = \mathtt{ppmc}(\mathtt{x},\mathtt{y})$.


abcde
abcd
abce
abde
acde
bcde
abc
abd
abe
acd
ace
ade
bcd
bce
bde
cde
ab
ac
ad
ae
bc
bd
be
cd
ce
de
a
b
c
d
e
Ø

Le treillis des diviseurs



La notion de connexion de Galois (1)

La *connexion de Galois* $(\mathtt{f},\mathtt{g})$ d'un *contexte* $\mathtt{I} \times \mathtt{P}$ se définit comme suit :

- La fonction $\mathtt{f} : 2^{\mathtt{I}} \longrightarrow 2^{\mathtt{P}}$ associe à un sous-ensemble d'individus $\mathtt{X}$ de $\mathtt{I}$ le sous-ensemble de propriétés communes qu'ils partagent $\mathtt{Y}$ de $\mathtt{P}$ (*intension* de $\mathtt{X}$).

  La fonction $\mathtt{f} : 2^{\mathtt{I}} \longrightarrow 2^{\mathtt{P}}$ construit l'intension (la plus spécifique) décrivant un ensemble d'individus de $\mathtt{I}$.
- La fonction $\mathtt{g} : 2^{\mathtt{P}} \longrightarrow 2^{\mathtt{I}}$ associe à un sous-ensemble de propriétés $\mathtt{Y}$ de $\mathtt{P}$ le sous-ensemble d'individus $\mathtt{Y}$ de $\mathtt{P}$ qui possèdent ces propriétés (*extension* de $\mathtt{Y}$).

  La fonction $\mathtt{g} : 2^{\mathtt{P}} \longrightarrow 2^{\mathtt{P}}$ détermine l'extension (la plus générale) décrivant un ensemble de propriétés de $\mathtt{P}$.

La notion de connexion de Galois (2, exemple)

Objets / Items	a	b	c	d	e
O-1	0	1	1	0	1
O-2	1	0	1	1	0
O-3	1	1	1	1	0
O-4	1	0	0	1	0
O-5	1	1	1	1	0
O-6	1	0	1	1	0

$f(\{O\text{-}1\}) = \{b,c,e\}$ et $g(\{b,c,e\}) = \{O\text{-}1\}$

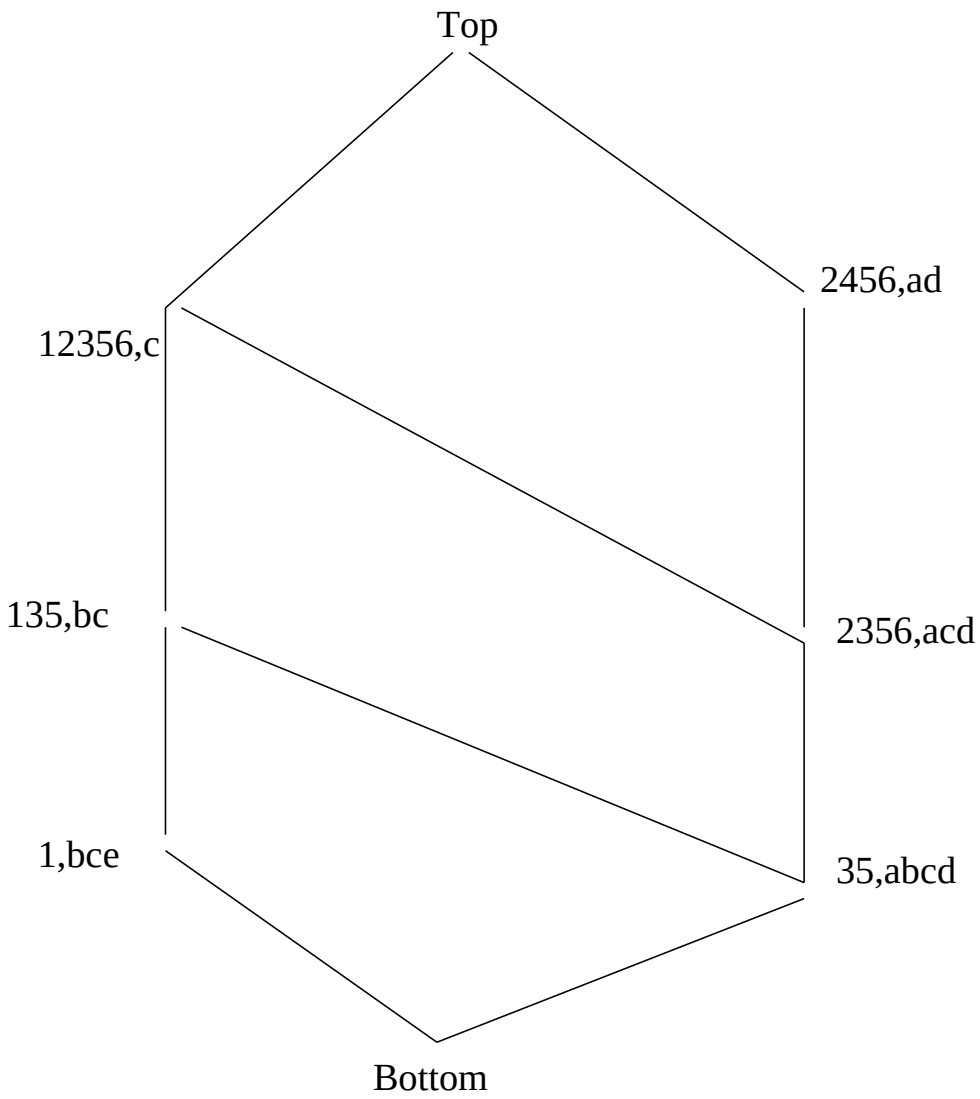
$f(\{O\text{-}1,O\text{-}2\}) = \{c\}$ et $g(\{c\}) = \{O\text{-}1,O\text{-}2,O\text{-}3,O\text{-}5,O\text{-}6\}$

$g(\{a,c\}) = \{O\text{-}2,03\text{-},O\text{-}5,O\text{-}6\}$ et $f(\{O\text{-}2,O\text{-}3,O\text{-}5,O\text{-}6\}) = \{a,c,d\}$

La notion de connexion de Galois (3)

- Les fonctions $\mathtt{h} = \mathtt{g}(\mathtt{f})$ et $\mathtt{h}' = \mathtt{f}(\mathtt{g})$ sont des opérateurs de *fermeture* : ils sont croissants, extensifs, et idempotents.
- Un *fermé* $\mathtt{X}$ de $\mathtt{h}$ vérifie $\mathtt{h}(\mathtt{X}) = \mathtt{X}$; un *fermé* $\mathtt{Y}$ de $\mathtt{h}'$ vérifie $\mathtt{h}'(\mathtt{Y}) = \mathtt{Y}$.
- $\mathtt{h}$ et $\mathtt{h}'$ permettent de construire le *treillis de Galois* du contexte $\mathtt{I} \times \mathtt{P}$.
- Le treillis de Galois est défini comme le produit des deux treillis isomorphes $\mathtt{L}_\mathtt{I} \times \mathtt{L}_\mathtt{P}$, où $\mathtt{L}_\mathtt{I}$ est le treillis des fermés pour $\mathtt{h}$ — ou *treillis des extensions* — et $\mathtt{L}_\mathtt{P}$ le treillis des fermés pour $\mathtt{h}'$, ou *treillis des intensions*.

Le treillis de Galois associé



Treillis de Galois (ou treillis de concepts)

- Dans un treillis de Galois ou *treillis de concepts*, les concepts sont des couples $\mathtt{C_k} = (\mathtt{E_k}, \mathtt{I_k})$, définis par une une *extension* $\mathtt{E_k}$ et une *intension* $\mathtt{I_k}$:

  $\longrightarrow \mathtt{E_k}$ représente l'ensemble des éléments recouverts par le concept, ou éléments qui possèdent les propriétés de $\mathtt{I_k}$,

  $\longrightarrow \mathtt{I_k}$ représente l'ensemble dual des propriétés du concept, ou les propriétés possédées par les éléments de $\mathtt{E_k}$.

- L'ordre partiel dans un treillis de concepts est défini par :

  $(\mathtt{E_1}, \mathtt{I_1}) \sqsubseteq (\mathtt{E_2}, \mathtt{I_2})$ ssi $\mathtt{E_1} \subseteq \mathtt{E_2}$, ou de façon duale, $\mathtt{I_2} \subseteq \mathtt{I_1}$.

Treillis de Galois (bibliographie)

Éléments de bibliographie :

- M. Barbut et B. Monjardet, Ordre et classification (2 tomes), Hachette, Paris, 1970.
- B.A. Davey and H.A. Priestley, Introduction to Lattices and Order, Cambridge University Press, Cambridge, 1990.
- V. Duquenne, Latticial structures in data analysis, Theoretical Computer Science, 217:407–436, 1999.

L’analyse formelle de concepts

- La notion de treillis de Galois a donné naissance à la *classification par treillis* et à l’*analyse de concepts formels*

  Voir : B. Ganter and R. Wille, Formal Concept Analysis, Springer, Berlin, 1999, et `fca-list@aifb.uni-karlsruhe.de` `http://www.aifb.uni-karlsruhe.de/mailman/listinfo/fca-list`
- L’*analyse de concepts formels* est utilisée dans des contextes divers, pour extraire des hiérarchies de concepts (à partir de contextes formels), mais aussi pour concevoir et ajuster des hiérarchies de classes en programmation par objets (classification et re-classification).
- Un des enjeux actuels de la classification par treillis est de construire des treillis à partir de données complexes multi-valuées et relationnelles

  (voir G. Polaillon, Organisation et interprétation par les treillis de Galois de données de type multi-valué, intervalle ou histogramme, Thèse d’informatique, Université Paris IX Dauphine, 1998).

L’extraction de motifs fréquents et de règles d’association

à partir d’un contexte formel

La notion de motif fréquent

- Un **motif** est un ensemble d'items qui apparaît dans une population d'objets donnés (*un objet possède l'item*) ; le nombre d'items donne la **longueur** du motif.
- L'**image** d'un motif correspond à l'ensemble d'objets qui possèdent le motif.
- Le **support** d'un motif correspond à la proportion d'objets qui possèdent le motif (par rapport à la population totale) d'objets.
- Un motif est **fréquent** si son support est supérieur à un *seuil de fréquence* σ_S donné :

  une proportion au moins égale à σ_S des objets possèdent *tous* les items présents dans le motif.

Exemple de motifs et motifs fréquents

Objets / Items	a	b	c	d	e
O-1	0	1	1	0	1
O-2	1	0	1	1	0
O-3	1	1	1	1	0
O-4	1	0	0	1	0
O-5	1	1	1	1	0
O-6	1	0	1	1	0

$\{a\}$ est un motif de longueur 1 et de support 5/6 ;

$\{ac\}$ est un motif de longueur 2 et de support 4/6 ;

$\{abc\}$ est un motif de longueur 3 et de support 2/6 ;

$\{abcde\}$ est un motif de longueur 5 et de support est 0/6.

Si le seuil de fréquence σ_S est 3/6, alors $\{a\}$ et $\{ac\}$ sont fréquents mais $\{abc\}$ et $\{abcde\}$ ne le sont pas.

La recherche par niveaux de motifs fréquents

Algorithme APriori

- La recherche commence en partant des motifs de longueur 1.
- Les motifs fréquents sont retenus et combinés entre eux pour former les motifs *candidats* de longueur supérieure, et le processus est réitéré.
- **Deux principes duaux et fondamentaux** :

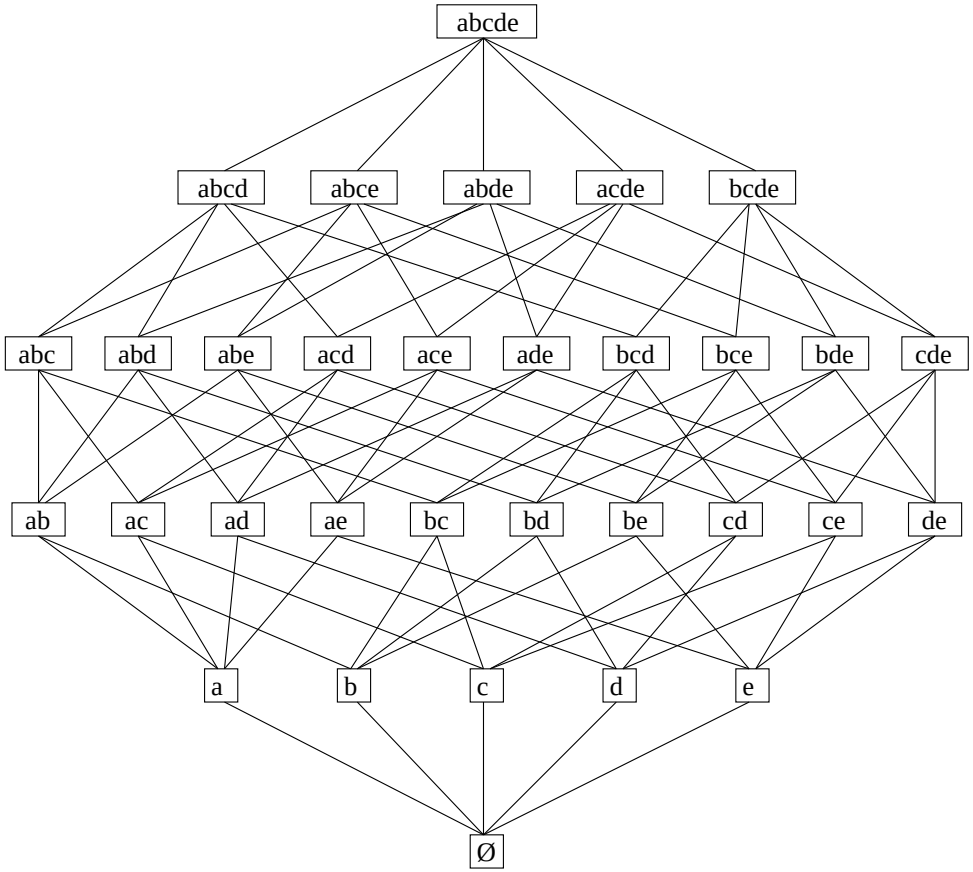
  $\longrightarrow$ *Tous les sous-motifs d'un motif fréquent sont fréquents*.

  $\longrightarrow$ *Tous les super-motifs d'un motif non fréquent sont non fréquents*.
- Treillis et motifs sont liés : la recherche de motifs fréquents revient à un parcours du treillis sous-jacent.
- L'appariement et l'accès à la base des données sont des opérations fondamentales dans le processus, et leur emploi doit donc être minimisé.

Les motifs fréquents extraits du contexte formel

σ_S = 2/6

- **Motifs de taille 1** : {a} (3/6), {b} (5/6), {c} (5/6), {d} (5/6).
- **Motifs de taille 2** : {ab} (2/6), {ac} (4/6), {ad} (5/6), {bc} (3/6), {bd} (2/6), {cd} (4/6).
- **Motifs de taille 3** : {abc} (2/6), {abd} (2/6), {acd} (4/6), {bcd} (2/6).
- **Motifs de taille 4** : {abcde} (2/6).



Les motifs :

{a},{b},{c},{d},{ab},{ac},{ad},{bc},{bd},{cd},
{abc},{abd},{acd},{bcd},{abcde}.

Les principaux algorithmes de recherche de motifs

- **Close** (recherche de motifs fermés maximaux) et **Pascal** (recherche de motifs clés minimaux) :

  — Y. Bastide, Data mining : algorithmes par niveau, techniques d'implantation et applications, Thèse d'informatique, Université Blaise Pascal, Clermont-Ferrand, 2000.

  — Y. Bastide, R. Taouil, N. Pasquier, G. Stumme et L. Lakhal, Pascal : un algorithme d'extraction des motifs fréquents, TSI, 21(1):65–95, 2002. Close

- **Titanic** (profondeur d'exploration) :

  — G. Stumme, R. Taouil, Y. Bastide, N. Pasquier and L. Lakhal, Computing Iceberg Concept Lattices with Titanic, Journal of Data and Knowledge Engineering, 42(2):189–222, 2002.

L’extraction de règles d’association

à partir de motifs fréquents

L'extraction de règles d'association à partir de motifs fréquents

- Une *règle d'association* est de la forme $\mathtt{A} \longrightarrow \mathtt{B}$ où $\mathtt{A}$ et $\mathtt{B}$ sont deux ensembles d'items.
- Le *support d'une règle* $\mathtt{A} \longrightarrow \mathtt{B}$ est défini comme le support de $(\mathtt{A} \cup \mathtt{B})$.
  La *confiance d'une règle* $\mathtt{A} \longrightarrow \mathtt{B}$ est définie comme le rapport $\mathtt{support}(\mathtt{A} \cup \mathtt{B})$ / $\mathtt{support}(\mathtt{A})$.
- Une règle est dite *valide* si sa confiance est supérieure à un *seuil de confiance* σ_C, et son support est supérieur au *seuil de fréquence* σ_S (de fréquence des motifs).
  $\longrightarrow$ Une règle valide ne peut être extraite qu'à partir d'un motif fréquent.
- Une règle est dite *exacte* si sa confiance est de 1, ce qui revient à $\mathtt{support}(\mathtt{A} \cup \mathtt{B}) = \mathtt{support}(\mathtt{A})$, sinon la règle est *partielle*.

Exemples de règles extraites de motifs fréquents

Objets / Items	a	b	c	d	e
O-1	0	1	1	0	1
O-2	1	0	1	1	0
O-3	1	1	1	1	0
O-4	1	0	0	1	0
O-5	1	1	1	1	0
O-6	1	0	1	1	0

Avec σ_S = 3/6 et σ_C = 3/5, $\{\mathtt{ac}\}$ est fréquent ; la règle $\mathtt{a} \longrightarrow \mathtt{c}$ est valide (support 4/6 et confiance 4/5) ; la règle $\mathtt{c} \longrightarrow \mathtt{a}$ est valide (support 4/6 et confiance 4/5) ;

Avec σ_S = 2/6 et σ_C = 3/5, $\{\mathtt{abd}\}$ est fréquent ; la règle $\mathtt{b} \longrightarrow \mathtt{ad}$ est valide (support 2/6 et confiance 2/3) ; la règle $\mathtt{ad} \longrightarrow \mathtt{b}$ n'est pas valide (support 2/6 et confiance 2/5).

La forme des règles d'association extraites

- Soit la règle : $\mathtt{A} \longrightarrow \mathtt{B} - \mathtt{A}$, où avec $\mathtt{A} \subseteq \mathtt{B}$, alors :

  $\mathtt{support}(\mathtt{A} \longrightarrow \mathtt{B} - \mathtt{A}) = \mathtt{support}(\mathtt{A} \cup (\mathtt{B} - \mathtt{A})) = \mathtt{support}(\mathtt{B})$

  Pour que $\mathtt{A} \longrightarrow \mathtt{B} - \mathtt{A}$ soit valide, il faut que $\mathtt{B}$ soit un motif fréquent, et donc $\mathtt{A}$ en est un aussi.

- Si l'on sait qu'une règle à prémisse minimale est valide :

  si $\mathtt{A} \longrightarrow \mathtt{B} - \mathtt{A}$ est valide, alors $\mathtt{A}' \longrightarrow \mathtt{B} - \mathtt{A}'$, où $\mathtt{A} \subseteq \mathtt{A}' \subseteq \mathtt{B}$ est valide.

  Ainsi, si $\{\mathtt{ab}\} \longrightarrow \{\mathtt{cd}\}$ est valide, alors $\{\mathtt{abc}\} \longrightarrow \{\mathtt{d}\}$ et $\{\mathtt{abd}\} \longrightarrow \{\mathtt{c}\}$ sont valides.

Un algorithme d'extraction de règles d'association (1)

σ_S = 2/6 et σ_C = 2/5

- **Initialisation** :

  Génération des règles valides de la forme : $\mathtt{P} - \{\mathtt{i}\} \longrightarrow \{\mathtt{i}\}$ à partir des motifs fréquents $\mathtt{P}$ de taille 2 : $\{\mathtt{ab}\}$ (2/6), $\{\mathtt{ac}\}$ (4/6), $\{\mathtt{ad}\}$ (5/6), $\{\mathtt{bc}\}$ (3/6), $\{\mathtt{bd}\}$ (2/6), $\{\mathtt{cd}\}$ (4/6).

  Si $\mathtt{P} = \{\mathtt{ab}\}$, les règles engendrées sont :

  $\{\mathtt{a}\} \longrightarrow \{\mathtt{b}\}$ (2/6,2/5)

  $\{\mathtt{b}\} \longrightarrow \{\mathtt{a}\}$ (2/6,2/3)

Un algorithme d'extraction de règles d'association (2)

- **Suite de la génération** :

  Génération des règles valides de la forme : $P - \{i\} \longrightarrow \{i\}$ à partir des motifs fréquents P de taille 3 : $\{abc\}$ (2/6), $\{abd\}$ (2/6), $\{acd\}$ (4/6), $\{bcd\}$ (2/6).

  Si $P = \{abc\}$, les règles engendrées sont :

  $\{ab\} \longrightarrow \{c\}$ (2/6,1)

  $\{ac\} \longrightarrow \{b\}$ (2/6,1/2)

  $\{bc\} \longrightarrow \{a\}$ (2/6,2/3)

  et $\{a,b,c\}$ sont trois conclusions valides qui peuvent se combiner pour donner $\{ab,ac,bc\}$ et :

  $\{c\} \longrightarrow \{ab\}$ (2/6,2/5)

  $\{b\} \longrightarrow \{ac\}$ (2/6,2/3)

  $\{a\} \longrightarrow \{bc\}$ (2/6,2/5)

  ...

Autres mesures associées aux règles d'association

- Parallèlement au support et à la confiance, il existe d'autres mesures associées aux règles d'association : *généralité*, *découverte*, *nouveauté*, *oubli*, etc.
- $\mathtt{A} \longrightarrow \mathtt{B}$ se réécrit en logique $\left(\neg \mathtt{A} \vee \mathtt{B}\right)$: l'*indice de l'analyse statistique implicative*, mesure la probabilité d'avoir $\left(\mathtt{A} \wedge \neg \mathtt{B}\right)$.

Un exemple d’application de l’extraction de motifs fréquents dans une base de données de réactions (données évolutives)

- S. Berasaluce, Fouille de données et acquisition de connaissances à partir de bases de données de réactions chimiques, Thèse de chimie informatique et théorique, Université Henri Poincaré Nancy 1, 2002.
- S. Berasaluce, C. Laurenço et A. Napoli, Extraction de connaissances à partir de bases de données de réactions en chimie organique, Treizième journées francophones d’ingénierie des connaissances — IC 2002, Rouen, France, B. Bachimont éditeur, pages 151–162, 2002.

La fouille de base de données de réactions en chimie organique

- En chimie organique, il existe de très grandes bases de données : plusieurs millions de substances décrites avec leurs propriétés chimiques, physiques et biologiques, et plusieurs millions de réactions.
- Pour aider le chimiste dans la résolution de problèmes de synthèse, l'idée est d'exploiter l'ECBD pour découvrir des régularités dans les données réactionnelles et faire émerger des schémas réactionnels génériques.
- Ces schémas réactionnels peuvent ensuite être réutilisés pour optimiser les requêtes et l'indexation dans les bases de réactions, en combinaison avec un système de connaissances, mais aussi pour mettre au point des plans de synthèse.

La représentation des données relatives aux réactions

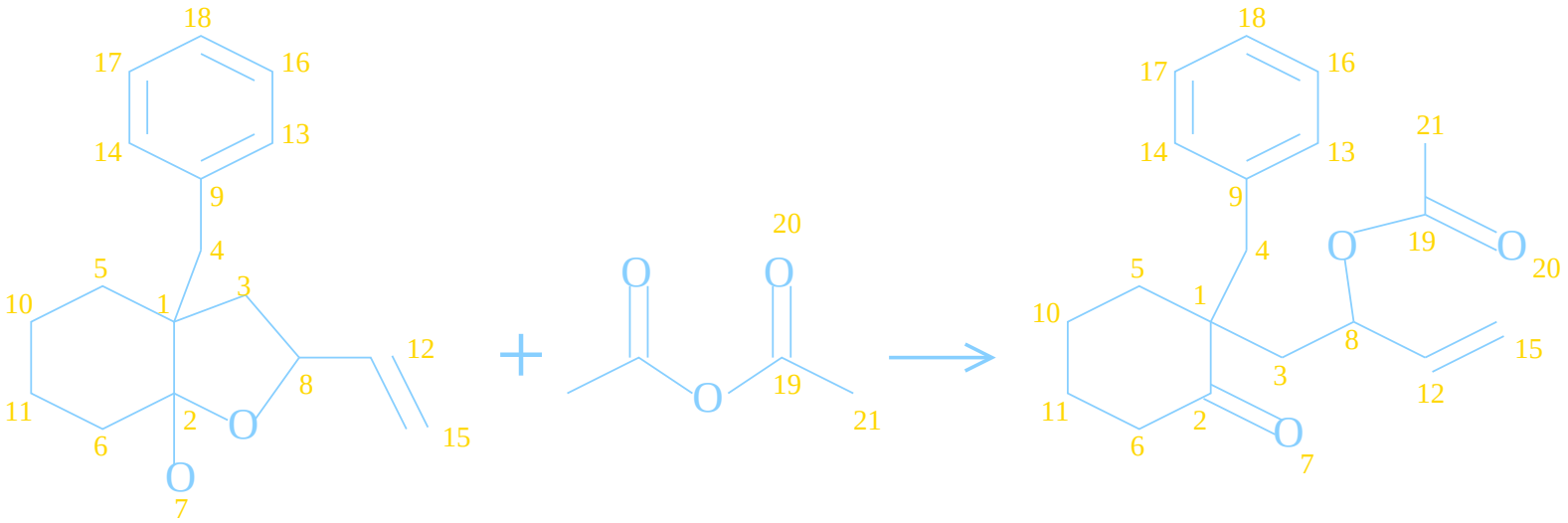
- Une réaction s'écrit sous la forme $\mathtt{A} + \mathtt{B} \longrightarrow \mathtt{C}$, où $\mathtt{A}$, $\mathtt{B}$, et $\mathtt{C}$ sont des molécules qui se composent de blocs.
- Une réaction est décrite par les blocs fonctionnels qu'elle met en jeu :

  $\Longrightarrow$ *blocs détruits* ($\mathtt{F}_\mathtt{d}$) : qui sont dans $\mathtt{A}$ ou $\mathtt{B}$ mais pas dans $\mathtt{C}$.

  $\Longrightarrow$ *blocs inchangés* ($\mathtt{F}_\mathtt{i}$), qui sont dans $\mathtt{A}$ ou $\mathtt{B}$ et dans $\mathtt{C}$.

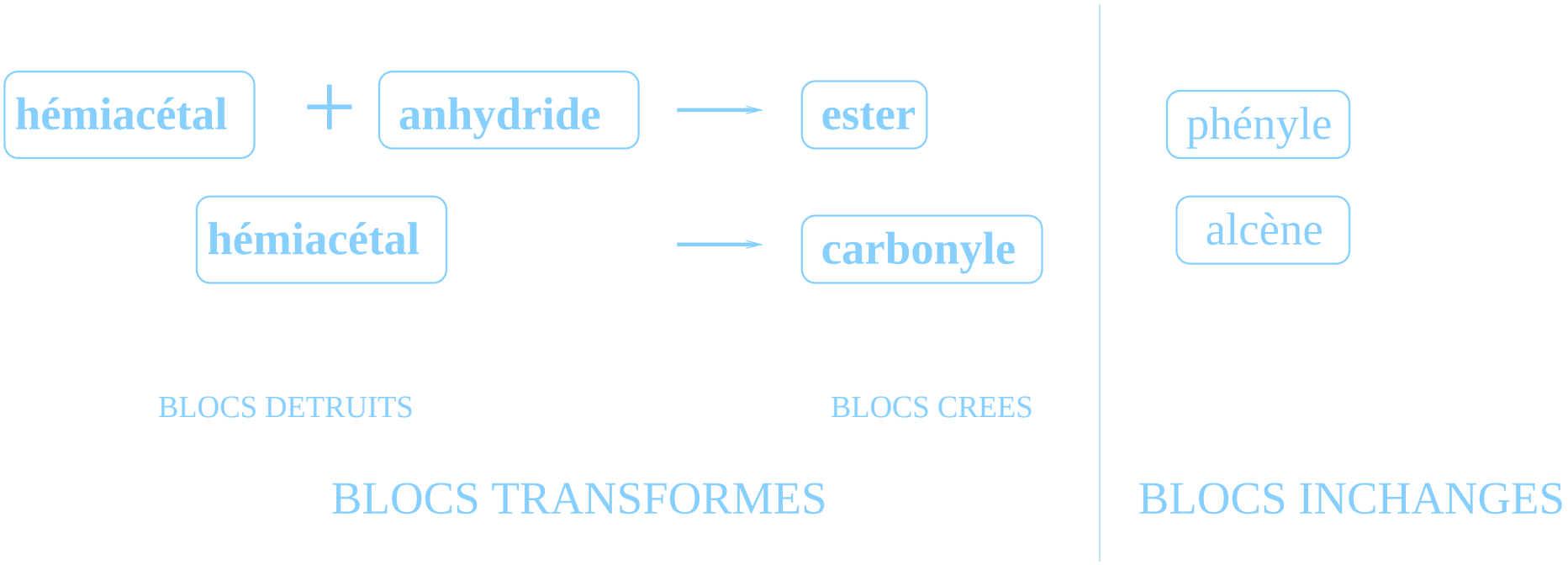
  $\Longrightarrow$ *blocs créés* ou $\mathtt{F}_\mathtt{c}$. qui ne sont ni dans $\mathtt{A}$ ni dans $\mathtt{B}$ mais dans $\mathtt{C}$.
- La base est transformée sous la forme d'un tableau booléen avec ces trois types de blocs.

La perception des composants d'une méthodes de synthèse



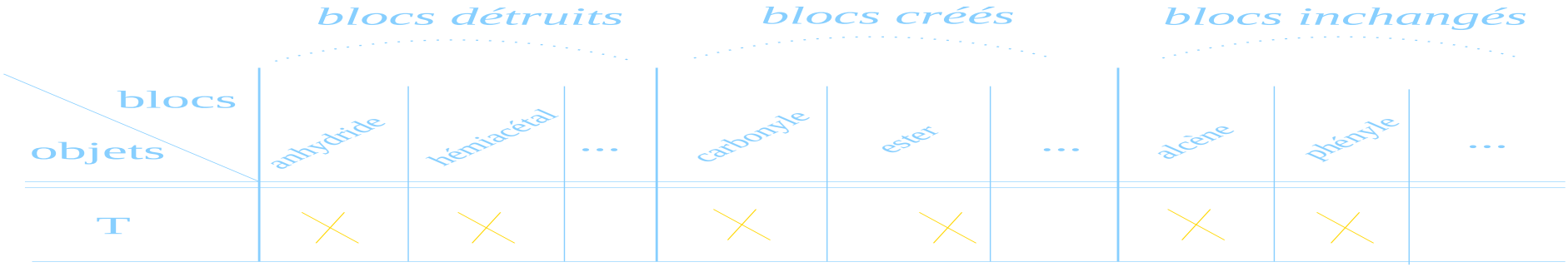
↓ perception

La perception en termes de blocs d’une méthode de synthèse

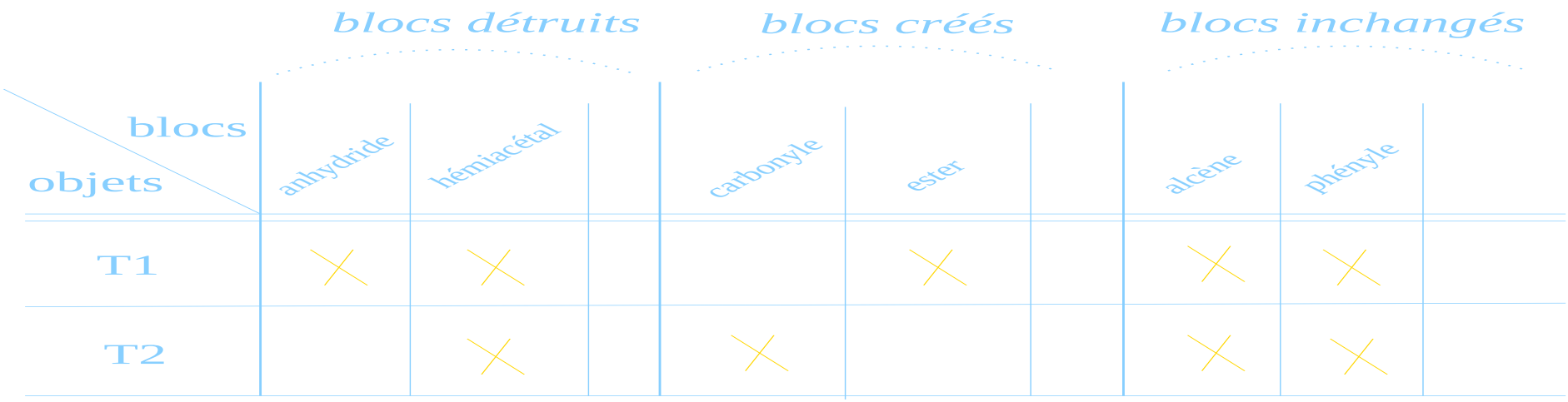


La mise en forme des données relatives à une réaction

Sans prise en compte explicite de la correspondance



Avec prise en compte explicite de la correspondance



Une interprétation des règles d'association

- La règle d'association « $? \longrightarrow \mathtt{C}$ » peut s'interpréter comme :

  « quelles propriétés faut-il satisfaire pour obtenir ou favoriser $\mathtt{C}$? »

- La règle d'association « $\mathtt{A} \longrightarrow ?$ » peut s'interpréter comme :

  ou « quelles propriétés peut-on posséder lorsque $\mathtt{A}$ est satisfait ? »

Interprétation des réactions en termes de motifs

- – À partir de quelle fonction $\mathtt{F_d}$ peut-on obtenir la fonction $\mathtt{F_c}$: recherche des motifs de la forme $\mathtt{F_d} \wedge \mathtt{F_c}$.

  – Un ester dérive (fréquemment) d'un alcool si $\mathtt{alcool_d} \wedge \mathtt{ester_c}$ est fréquent.
- Quelle est la fonction la plus fréquente qui permet d'obtenir un ester : recherche de motifs de la forme $\mathtt{ester_c} \wedge \mathtt{F_d}$.
- Quelles fonctions se combinent pour former un ester : recherche de motifs de la forme $\mathtt{F_{d1}} \wedge \mathtt{F_{d2}} \wedge \mathtt{ester_c}$.
- Quelles fonctions donnent un ester en laissant un éther inchangé : recherche de motifs de la forme $\mathtt{F_d} \wedge \mathtt{ether_i} \wedge \mathtt{ester_c}$.

Éléments sur la fouille de textes

- R. Feldman and H. Hirsh, Exploiting Background Information in Knowledge Discovery from Text, Journal of Intelligent Information Systems, 9(1):83–97, 1997.
- Y. Kodratoff, Knowledge Discovery in Texts: A Definition, and Applications, Foundations of Intelligent Systems (ISMIS'99), LNAI 1609, Z.W. Ras and A. Skowron editors, Springer, Berlin, pages 16–29, 1999.
- L. Lebart, A. Salem and E. Berry, Exploring Textual Data, Kluwer, Dordrecht, 1998.
- H. Cherfi, A. Napoli, et Y. Toussaint, Vers une méthodologie de fouille de textes en s'appuyant sur l'extraction de motifs fréquents et de règles d'association, Conférence CAp 2003, Plate-forme AFIA, Laval, 2003.

La fouille de textes (1)

- La *fouille de textes* consiste à analyser un volume important de textes pour construire une *synthèse interprétable* du contenu des textes.
- Les textes sont décomposés en groupes syntaxiques cohérents, et le processus de fouille de textes est appliqué pour fournir des *éléments synthétiques* permettant d'appréhender et de manipuler globalement les textes étudiés.
- Ces éléments synthétiques peuvent être un treillis de concepts, un ensemble de motifs et de règles d'association (qui peuvent jouer le rôle d'*explications* associées aux textes).
- Le processus de fouille de textes est contrôlé par un analyste et peut aussi s'appuyer sur une ontologie du domaine.

La fouille de textes (2)

Les résultats de la fouille de textes peuvent servir à :

- L'acquisition (semi-automatique) de connaissances ontologiques et linguistiques à partir de textes.
- La recherche d'information (par le contenu) et la veille technologique, pour l'analyse des tendances dans le domaine des textes.
- La construction d'une *synthèse conceptuelle* d'un ensemble de textes qui comprend : des termes clés et un réseau de règles permettant de situer et comparer les textes les uns par rapport aux autres.

Les données textuelles : un exemple

Corpus : 1361 documents (240 000 mots), indexés par 14 374 termes dont 632 sont différents.

Document 391

Titre : Sequencing of gyrase and topoisomerase IV quinolone-resistance-determining regions of Chlamydia trachomatis and characterization of quinolone-resistant mutants obtained In vitro.

Auteur(s) : Dessus-Babus-S ; Bebear-CM ; Charron-A ; Bebear-C ; de-Barbeyrac-B

Résumé : The L2 reference strain of Chlamydia trachomatis was exposed to subinhibitory concentrations of ofloxacin (0.5 microg/ml) and sparfloxacin (0.015 microg/ml) to select fluoroquinolone-resistant mutants. In this study, two resistant strains were isolated after four rounds of selection **[...]** A point mutation was found in the gyrA quinolone-resistance-determining region (QRDR) of both resistant strains, leading to a Ser83$\longrightarrow$Ile substitution (Escherichia coli numbering) in the corresponding protein. The gyrB, parC, and parE QRDRs of the resistant strains were identical to those of the reference strain. These results suggest that in C. trachomatis, DNA gyrase is the primary target of ofloxacin and sparfloxacin.

Les règles extraites des textes

- **Règle 120** : ”determine region” ”gyrA gene” ”gyrase” ”mutation” $\longrightarrow$ ”Quinolone” (*Support* = ”11” et *Confiance* = ”1.000”).

  Cette règle reflète le phénomène de résistance : les textes étudiés décrivent la mutation du gène ”gyrA” qui contrôle le comportement de l’enzyme ”gyrase” dans une zone précise de l’ADN, enzyme responsable de la résistance aux antibiotiques de la famille des ”Quinolones”.

- **Règle 202** : ”grlA gene” $\longrightarrow$ ”mutation” ”Staphylococcus Aureus” (*Support* = ”12” et *Confiance* = ”0.917”).

  Règle 270 : ”mecA” ”meticillin” $\longrightarrow$ ”mecA gene” ”Staphylococcus Aureus” (*Support* = ”12” et *Confiance* = ”1.000”).

  Ces deux règles indiquent que la ”meticillin” inhibe le gène ”mecA” des bactéries et permet de guérir des infections dues à la mutation du gène ”grlA” causé par la bactérie ”Staphylococcus Aureus”.

L’interprétation des règles extraites

- L’interprétation des règles est un processus très difficile et nécessite une connaissance approfondie du domaine des données.
- Toutes les règles extraites ne sont pas intéressantes ...

Le cadre du Web sémantique

- *Les machines parlent aux machines et sont au service des personnes*.
- Demain, le Web sera exploité — en priorité — par des machines, qui traiteront les problèmes posés par des personnes, et qui délivreront les résultats qu'elles auront obtenus à ces mêmes personnes.
- Le Web sera un *espace partagé déclaratif et navigable* : un espace de discussion pour les machines qui en exploitent toutes les ressources pour résoudre des problèmes.
- Une technologie pour le Web sémantique : XML + RDF(S) + ontologies/connaissances + services + recherche/fouille (moteurs intelligents).
- Le Web un peu partout : dans le frigo, le four, la voiture, la montre ...

La manipulation de documents pour le Web sémantique

- Donner un sens aux documents avec XML, RDF(S), et des langages de représentation des connaissances (logiques de descriptions) pour attacher une *sémantique* aux éléments des documents.
- Manipuler des documents en fonction de leur contenu et de leur sémantique : recherche d'information avec prise en compte de connaissances du domaine.
- Combiner l'extraction d'information et la fouille de textes : extraire des termes pour extraire des termes clés, et s'en servir pour fouiller et classifier les textes en fonction des contenus.

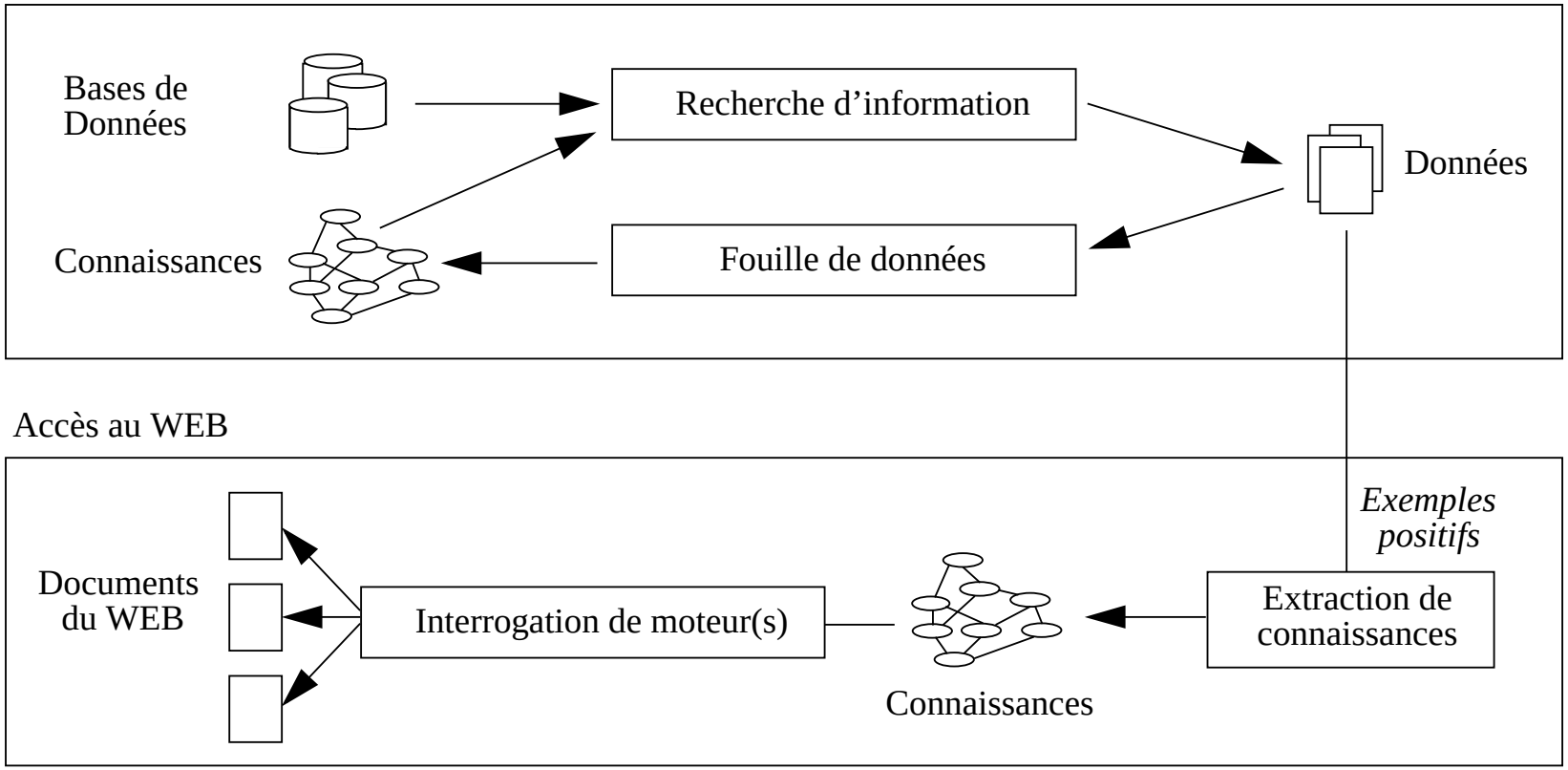
La notion d’ontologie

- La *sémantique* des documents sur le Web doit être accessible aux machines.
- Un élément majeur de cette sémantique est constitué par un *modèle explicite* du domaine des données.
- Un tel modèle décrit le *vocabulaire* et la *structure* des informations relatives au domaine d’intérêt, qui doit être *communément admis et partageable* : c’est là l’essence même de la notion d’*ontologie*, telle qu’elle est considérée en général en intelligence artificielle.

La construction d’ontologies

- Acquisition de connaissances et recueil d’expertise : « méthode manuelle ».
- Méthodes (semi-automatiques) en apprentissage (supervisées et non supervisées) : induction, inférence grammaticale, méthodes classificatoires, ...
- Extraction de connaissances et fouille de données.
- *Recherche d’information guidée par la fouille de données* : extraction de termes clés et construction (enrichissement) d’ontologies à partir du contenu des documents, puis exploitation de ces ontologies pour fouiller les textes, ce qui produit de nouvelles informations, qui vont enrichir l’ontologie ...

Système hypertexte de fouille de données bibliographiques



ECBD et Web sémantique (1)

- La *fouille de documents sur le Web* ou *fouille du Web* peut participer à la mise en œuvre du Web sémantique : les ontologies doivent être construites de façon (semi-)automatique pour envisager un passage à l'échelle.
- Les ontologies sont utilisées pour *annoter* les documents et améliorer en retour le processus de fouille en permettant d'exploiter la sémantique et le contenu des documents.
- **Bibliographie** : B. Berendt, A. Hotho and G. Stumme, Towards Semantic Web Mining, in The Semantic Web - ISWC 2002, LNAI 2342, I. Horrocks and J. Hendler editors, Springer, Berlin, pages 264–278, 2002.

ECBD et Web sémantique (2)

La fouille du Web peut se pratiquer selon trois points de vue principaux :

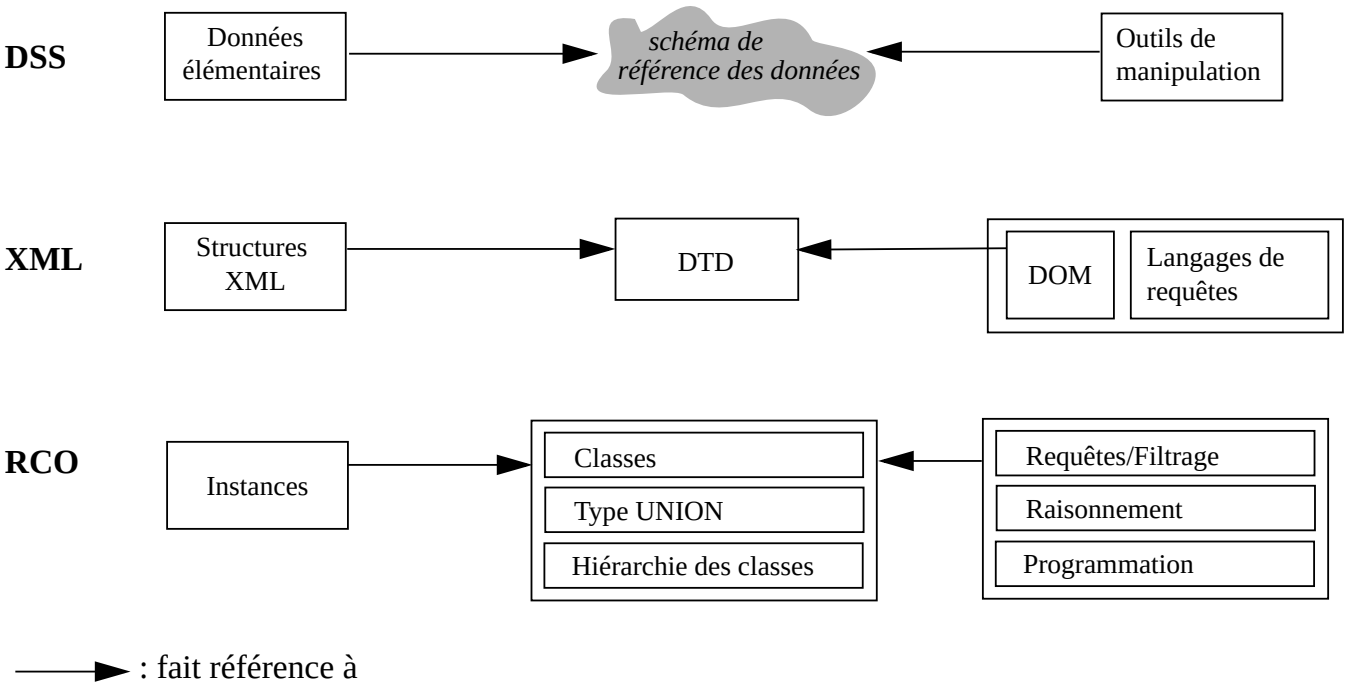
- La *fouille du contenu* (des documents) qui est en rapport avec la fouille de textes.
- La *fouille de la structure* (des pages) et des liens hypertextuels.
- La *fouille des usages* ou des ensembles d’opérations effectuées sur les pages.

ECBD et Web sémantique (3)

Fouiller le Web pour :

- Mettre en œuvre et enrichir des structures sémantiques pour organiser les documents avec des *points de vue* particuliers : classification de documents par points de vue.
- Peupler ces structures sémantiques pour l'annotation de documents et la manipulation par le contenu.
- Concevoir des ontologies finalisées pour des traitements particuliers et locaux : « Web sémantiques locaux ».

L’expérience ESCRIRE



Conclusion

- C’est fini pour aujourd’hui, mais il reste encore beaucoup à faire, de tous les côtés ... ainsi :

  $\longrightarrow$ La préparation et la mise en forme des données.

  $\longrightarrow$ L’analyse et le traitement des données complexes.

  $\longrightarrow$ L’interprétation et la mise en œuvre des systèmes de connaissances pour l’ECBD, et plus spécifiquement dans le cadre du Web sémantique.

  $\longrightarrow$ Des enjeux importants liés aux domaines des données : biologie, chimie, médecine, espace, météo, finances, ...

- **Et pour finir** : le petit mot de l’orpailleur, fouiller les données est comparable à rechercher de l’or ou faire de l’archéologie : il faut s’habituer au terrain, et puis c’est un travail long, difficile (il faut recommencer souvent ...) et même quelquefois salissant, et bien sûr, cela ne peut donner aucun résultat ... mais il faut tenir bon et être optimiste !

INFORSID'2003

organisé avec le concours de :