DILIB[1]

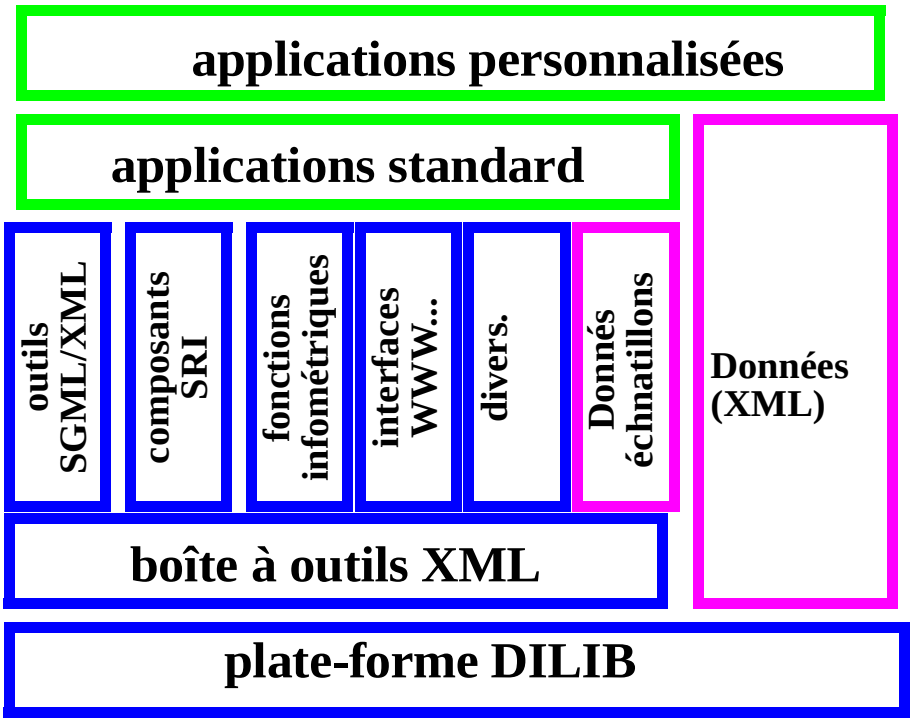
Une plate-forme pour l'Ingénierie du Document et de l'Information Scientifiques et Techniques

Formation recherche en ingénierie de l'IST

Investigation documentaire

Construction de Systèmes de Recherche d'Information

Construction de plate-formes d'exploitation de l'Information



1. Dilib est un prototype réalisé par le LORIA & l'INIST.

DILIB et NORMALISATION SGML

Notices bibliographiques et informations normalisées :

- formats simples

```
titre : Annuaire
aut : La Poste

<doc><titre>Annuaire</titre><aut>La Poste</aut>
     </doc>
```

- formats professionnels (CCF, USMARC, Unimarc...)

```
210 $a Paris $c Dunod $d 1988

<unimarc>...<f210><sa>Paris</sa><sc>Dunod</sc>
     <sd>1988</sd><f210>...</unimarc>
```

- fichiers inverses, index...

```
<idx><loc>Paris</loc><f>2</f>
     <l><e>0023</e><e>4123</e></l></idx>
```

Commandes Unix

- orientées SGML

```
SgmlSelect -g unimarc/f210/sa#=Paris
```

- Système de Recherche d'informations en KIT

```
IndexSelect -h base.ville.index -k Paris
```

Bibliothèque de fonctions en langage C

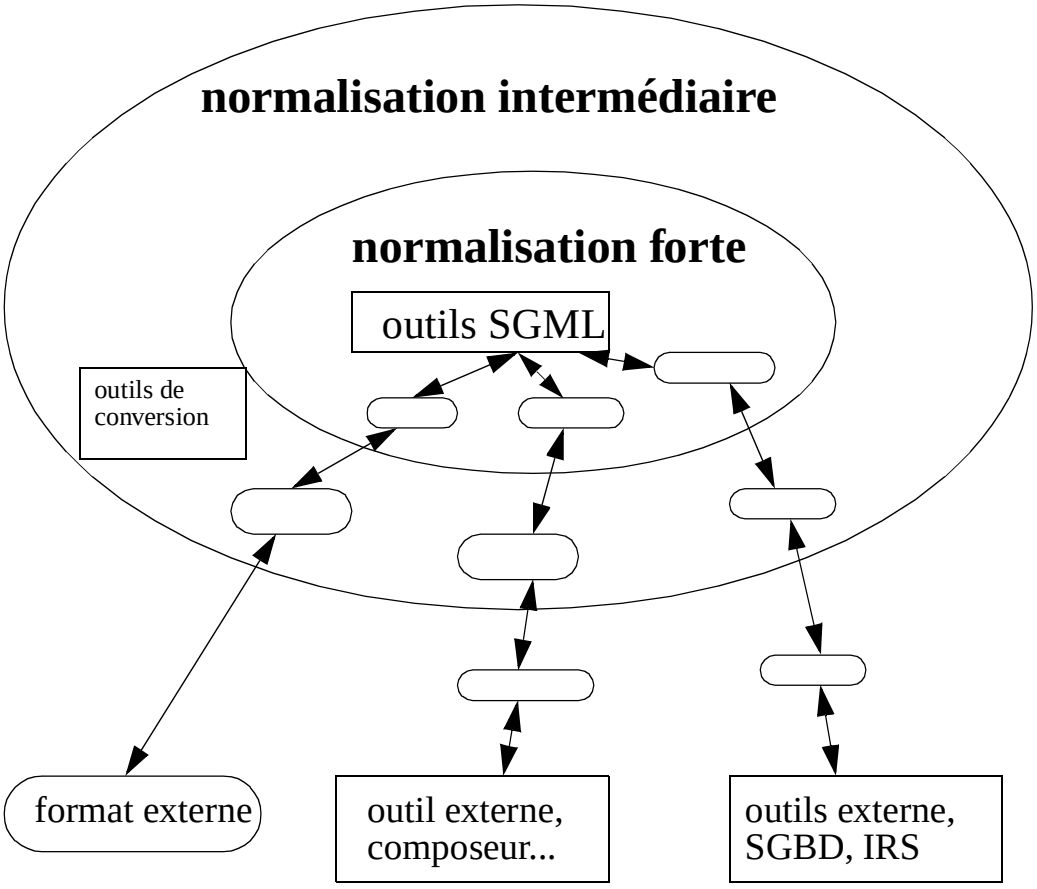
```
SgmlAddSon(zone210, SgmlCreateLeaf("sd","1990"));
```

Interfaces et outils spécifiques

- avec des progiciels, exemple LaTeX, WWW, Texto...
- extensions linguistiques, infométriques (clusterisation...)

Dilib - normalisation - introduction

monde hétérogène



donnée

outils

- normalisation forte : outils principaux
- normalisation intermédiaire : destinée à faciliter l'utilisation des outils de conversion

Manipulation des collections d'objets SGML

chemin de balises DILIB

philosophie proche de Xpath

Idée générale

s'inspirer des path d'unix pour désigner un élément dans une structure SGML

```
<doc><tit>a</tit><kw><e>m1</e><e>m2</e></kw></doc>
```

`doc/tit` désigne de façon unique l'élément `<tit>a</tit>`

`doc/kw/e` désigne les éléments `<e>m1</e>` et `<e>m2</e>`

Chemins élémentaires sur enregistrements simples

- chemin de balise = suite de spécifieurs de balises séparés par des /
- spécifieur commençant par une lettre : ensemble des fils du noeud courant ayant un identificateur identique au spécifieur

```
doc/kw -> <kw><e>m1</e><e>m2</e></kw>
doc/kw/e -> <e>m1</e>
            <e>m2</e>
```

Utilisation des métadonnées «anciennes» dans le monde SGML

Avantages :

Pas de modification des pratiques de catalogage

Utilisation de l'Ingénierie SGML

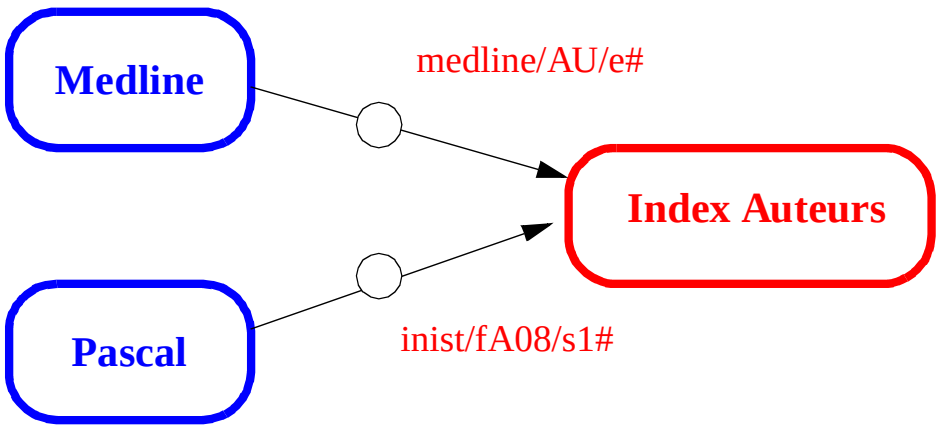
Pas de reformatage lourd

(outils indépendants d'une DTD)

Exemples, commandes DILIB avec chemin de balises

```
SgmlSelect -g usmarc/f130/sa#?Paris?
SgmlSelect -g medline/TI#?Paris?
```

MedExplore, index communs à plusieurs sources hétérogènes



Normalisation Niveau Structure : SGML

chemin de balises

Exemple de commande utilisant des chemins de balises

```
<doc><tit>a</tit><kw><e>m1</e><e>m2</e></kw></doc>

SgmlSelect -s doc/kw/e -p @s1

     <e>m1</e>
     <e>m2</e>
```

Compléments sur les chemins de balises

- spécifieur réduit à une étoile : tous les fils

```
doc/* -> <tit>a</tit>
         <kw><e>m1</e><e>m2</e></kw>
```

- spécifieur numérique : rang du fils (1= premier, 0=dernier)

```
doc/kw/2   -> <e>m2</e>
doc/kw/2/1 -> m2
```

- Accés aux chaines contenues dans un élément terminal : #

```
doc/tit   -> <tit>a</tit>
doc/tit#  ->a
```

Normalisation, exemple manipulation d'enregistrements

Commandes Dilib

- SgmlCut

```
fra   camus   <aut><n>Camus</n><f>Albert</f><c>France</c></aut>
bel   herge   <aut><n>Herg&eacute;</n><c>Belgium</c><aut>
```

```
SgmlCut 2  aut/f
```

```
fra   <aut><n>Camus</n><c>France</c></aut>
bel   <aut><n>Herg&eacute;</n><c>Belgium</c><aut>
```

- SgmlSelect -g (analogue à grep)

```
fra  camus   <aut><n>Camus</n><f>Albert</f><c>France</c></aut>
bel  herge   <aut><n>Herg&eacute;</n><c>Belgium</c><aut>
```

```
SgmlSelect -g aut/c#?[Ff]rance? -g aut/n -p @g2
<n>Camus</n>
```

- SgmlSelect -s (split : éclatement)

```
01  <doc><t>SGML</t><k>ISO</k><k>Document</k><k>SGML</k></doc>
02  <doc><t>UNIMARC</t><k>IFLA</k><k>ISO 2709</k></doc>
```

```
SgmlSelect -s doc/k# -p @s1 -p @1
```

```
ISO         01
Document    01
SGML        01
IFLA        02
ISO 2709    02
```

Utilisation des commandes de base Unix

```
SgmlSelect -g aut/c#?[Ff]rance? -p @2 | wc
```

Ensemble de données DILIB, organisation HFD

(Hierarchic File organization for Documentation)

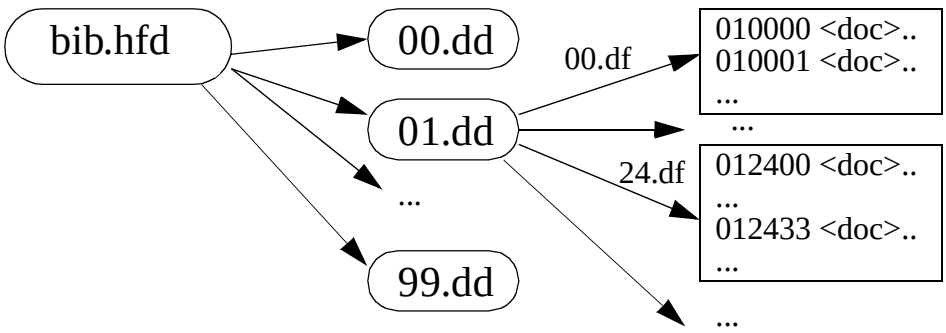
Idée générale

1000000 records

= 100 répertoires de 100 fichiers de 100 enregistrements

- record = key <tabulation> document
- clé = 6 chiffres exemple: 012433
- 2 premiers -> répertoire 01.dd
- 2 suivants -> fichier 24.df
- 2 derniers -> numéro d'enregistrement 33
- adresse Unix du fichier contenant 012433

-> `bib.hfd/01.dd/24.fd`



Extensions

- clé de longueur quelconque..
- Fichiers inverses, d'index, d'associations bâtis sur cette structure

Utilisation - Fichiers Inverses

Commande DamHfdSelect

- Syntaxe :
```
DamHfdSelect -h hfd <myListOfKey
```
- le fichier d'entrée contient une clé par ligne
- attention le nom du hfd ne doit pas être suffixé
- exemple :
```
DamHfdSelect -h $DILIB/Data/InRocCdd/BibTexto <<...
001100
002345
...
```

Fichiers Inverses

```
<idx><k>DTD</k><f>2</f>
     <l><e>001020</e><e>012354</e></l></idx>
```

Commande HfdBrowser

- Syntaxe :
```
HfdBrower -i myIndex -k myKey -a after -b before
```

Commande BaseQuery

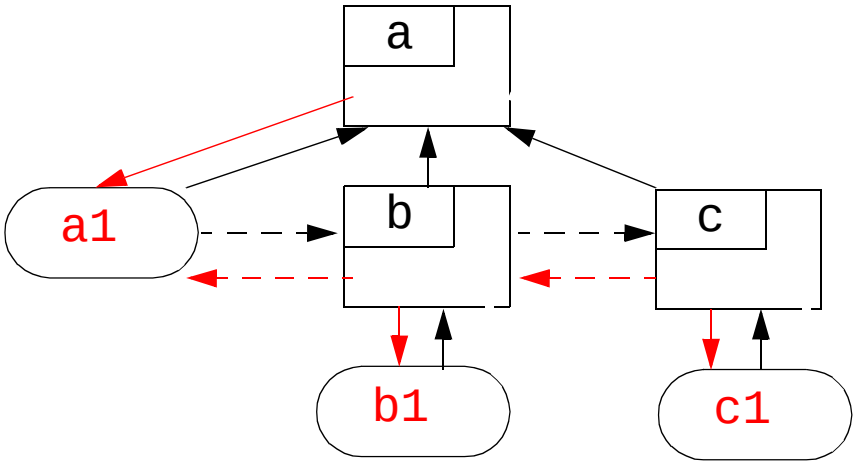
- Syntaxe :
```
BaseQuery -b myBase -e expression [-s session]
```
- expression (en Sgml) opérateur and or andNot
- exemple :

```
<and><term index="aut">Herge</term>
     <or><andNot><term>Tintin</term>
               <term>pastiche</term>
               <andNot></or></and>
```

Parser, arbres SGML - notions de base (exemples dans DILIB -> DOM W3C)

Objet fondamental : SgmlNode

```
<a>a1<b>b1</b><c>c1</c></a>
```



Type de base : SgmlNode

- méthodes :

`SgmlFather (*SgmlNode)`	*parentNode*
`SgmlNext (*SgmlNode)`	*nextSibling*
`SgmlPrevious (*SgmlNode)`	*previousSibling*

Types dérivés

- SgmlMark (correspondant à <a>, <b>,<c>)

`SgmlFirst (*SgmlNode)`	*firstChild*
`SgmlLast(*SgmlNode)`	*lastChild*

- SgmlData (a1, b1, c1)

Construction d'arbres SGML

Constructeur de base :

```
SgmlNode *SgmlCreateNode(type);
          char type;
```

Constructeurs effectivement utiles :

```
SgmlNode *SgmlCreateMark(tag);
          char *tag;
SgmlNode *SgmlCreateLeaf(tag, string);
          char *tag;
          char *string;

SgmlCreateMark("a");         /* <a></a> */
SgmlCreateLeaf("a","text"); /* <a>text</a> */
```

Méthodes de construction de base

```
SgmlNode *SgmlAddFirst(pere, fils);
                       SgmlNode *pere, *fils;
SgmlNode *SgmlAddLast(pere,fils);
```

Exemple :

```
/* création de <a><b>b1</b><c>c1</c></a> */
#include "Sgml.h"
SgmlNode *root;
     root =SgmlCreateMark("a");
     SgmlAddLast(root, SgmlCreateLeaf("b","b1"));
     SgmlAddLast(root, SgmlCreateLeaf("c","c1"));
```

accès à l'environnement d'un noeud

Les voisins ou parents

toutes les fonctions retournent un pointeur NULL en cas d'échec.

```
SgmlNode *SgmlNext(node);
SgmlNode *SgmlPrevious(node);
SgmlNode *SgmlFirst(node);
SgmlNode *SgmlLast(node);
SgmlNode *SgmlFather(node);
```

Les caractéristiques d'un noeud

- pour un noeud de type Mark

```
char *SgmlTag(node);
```

- pour un noeud de type Data

```
char *SgmlDataString(noeud);
```

- test de type

```
int SgmlIsData(node);
int SgmlIsMark(node);
```

Iterations sur les composants d'un noeud

Principe

Toutes les fonctions renvoient la valeur NULL en cas d'échec ou d'absence d'un élément.

Squelette d'une itération sur les fils d'un noeud

```
if ((son = SgmlFirst (node))
  {
     do
        { traitement sur son }
     while ((son=SgmlNext(son));
  }
else
  { traitement de l'exception }
```

Exemple

édition de toutes les étiquettes des fils d'un noeud

```
editSonTag(node)
          SgmlNode *node;
{
  SgmlNode *son;
  if ((son=SgmlFirst(node)))
  do{ if(SgmlIsMark(son))
            printf("%s\n",SgmlTag(son));
      } while ((son=SgmlNext(son));
}
```

Dilib - Import Export de structures SGML

Si la chaîne d'entrée est en forme normale, la construction d'un arbre SGML peut se faire sans DTD.

Conversion SGML <-> string

```
char *SgmlToString(node);
                   SgmlNode *node;
SgmlNode* SgmlFromString(string);
                   char *string;
```

Entrées-sorties (Dam = Dilib Access Method)

```
SgmlPrint(node);
SgmlFilePrint(file, node);
              FILE *file;
SgmlNode *SgmlInputNextDocument();
```

Exemple

Impression de tous les premiers fils ayant «a» pour tag.

```
#include «Sgml.h»
main()
{
  SgmlNode *docu ,*son;
  while(docu=SgmlInputNextDocument())
    {
       if (son=SgmlFirst(docu))
         {if (strcmp(SgmlTag(son),"a")==0)
                 {SgmlPrint(son);putchar('\n');
    }}};
}
```

Outils pour l'analyse de l'Information

Les fichiers Inverses

Exemple sur une base «jouet», les bandes dessinées, une notice

```
ref : BD02
tit : Tintin en Amérique
aut : Herge
mcl :   Tintin, Milou, Amérique du Nord,
     indien, Chicago, gangster, montagne
```

Le fichier inverse des mots clés

```
[7] montagne
[7] Amérique du Nord
[7] école
[6] Tintin
[6] Milou
[5] mouton
[5] berger
[4] indien
[4] adolescence
[4] Blueberry
```

Outils pour l'analyse de l'Information

Les fichiers Inverses

Exemple sur les mots clés (thème Wallerian degeneration)

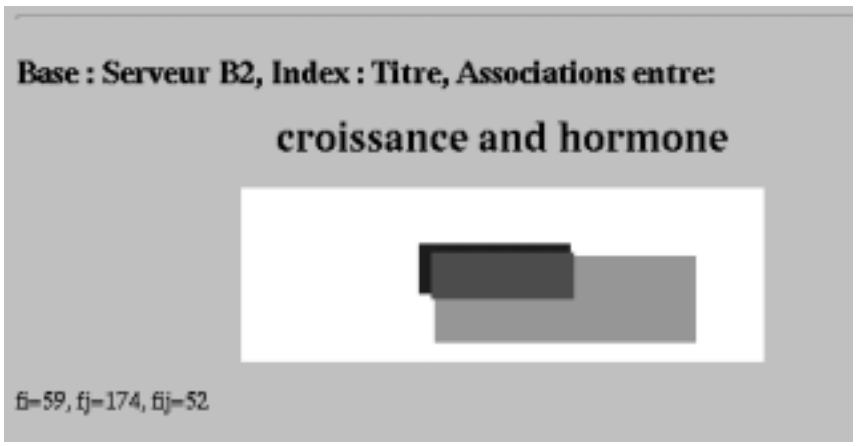
```
[393] Rats
[268] *Nerve Degeneration
[247] Wallerian Degeneration
[241] Microscopy, Electron
[215] *Wallerian Degeneration
[184] Time Factors
[128] Middle Age
[121] Mice
[120] Rats, Inbred Strains
[115] Adult
[76] Nerve Degeneration
```

Exemple sur les auteurs

```
[19] Griffin JW
[16] Bignami A
[15] Myers RR
[14] Friede RL
[14] Csillik B
[13] Powell HC
[12] Said G
[12] Dahl D
[11] Trapp BD
[11] Perry VH
[11] Knyihar-Csillik E
```

Outils pour l'analyse de l'Information

Les fichiers d'Associations



Exemple sur les auteurs

```
[10] Perry VH - Brown MC
[10] Knyihar-Csillik E - Csillik B
[10] Dahl D - Bignami A
[9] Powell HC - Myers RR
[8] Friede RL - Bruck W
[7] Privat A - Fulcrand J
[7] Marty R - Fuentes C
[7] Gueuning C - Graff GL
[6] Mackinnon SE - Hunter DA
```

Le fichier des associations des mots-clés de la base BD

```
[6] Milou - Tintin
[5] montagne - mouton
[5] berger - mouton
[5] berger - montagne
```

Outils pour l'analyse de l'Information

Les aggrégats d'associations (clusters)

Exemple sur les mots clés, liste des aggrégats

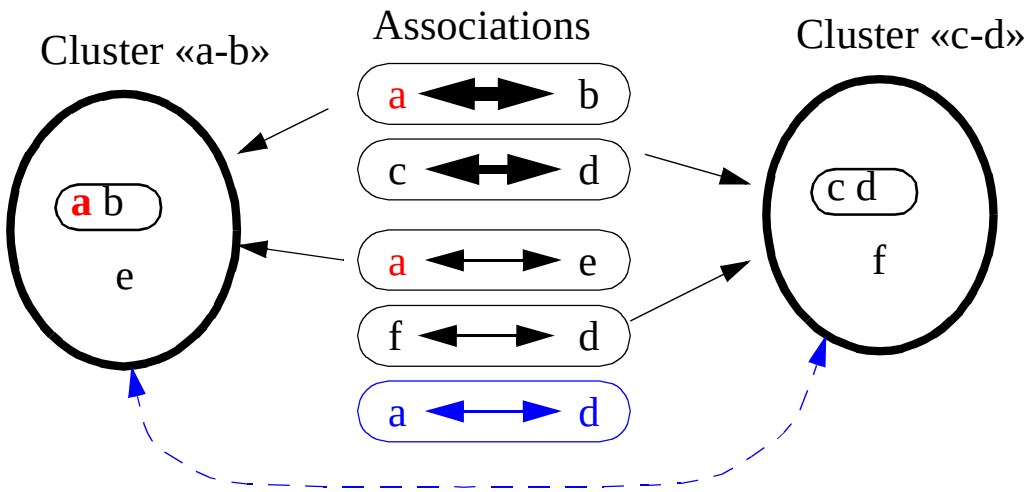
Milou - Tintin
Amérique du Nord - indien
école - adolescence
les Dupondt - lion
Jules César - Obelix
Haddock - drogue

Un aggrégat

List of Key-Words
[6] Milou
[6] Tintin
[3] Afrique
[7] montagne
[5] mouton
[5] berger
[3] touriste
[3] gangster
Internal Relationships
[6] Milou - Tintin
[3] Afrique - Tintin
[3] Afrique - Milou
[2] Tintin - montagne
[5] montagne - mouton
...
External Relationships
Amérique du Nord - indien
les Dupondt - lion
Haddock - drogue

Outils pour l'analyse de l'Information

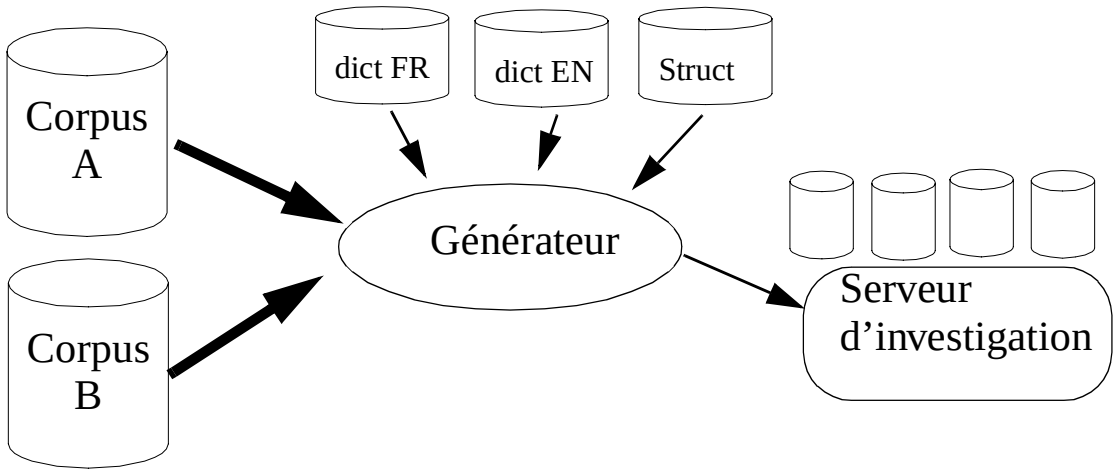
Les agrégats d'associations (clusters)



Exemple sur les mots clés (BIBAN - images Art Nouveau) liste des agrégats

Association la plus forte	F max	N docs
Ecole de Nancy - Verre	43	81
Intérieur - Architecture	7	63
Exposition - Nancy (FRA), Exposition (1909)	6	11
Bois - Ebénisterie	5	9
Papillon - Coupe	4	12

Générateur de Serveur d'Investigation



Les dictionnaires

```
Base/name          BD          Comics
yes                Oui         Yes
```

La structure

```
<server code=Demo>
    <base code=BD>
        <index code=aut>
            <path>doc/aut/e#
            <cross>kw
        </index>
        <index code=kw>
            <path>doc/mcl/e#
            <cross>aut
        <index>
```