



Approches flexibles et incrémentales pour les humanités numériques

*Exploration de corpus
Avec des boites à outils XML*

<https://wicri-demo.istex.fr/wicri-chanson-roland.fr>

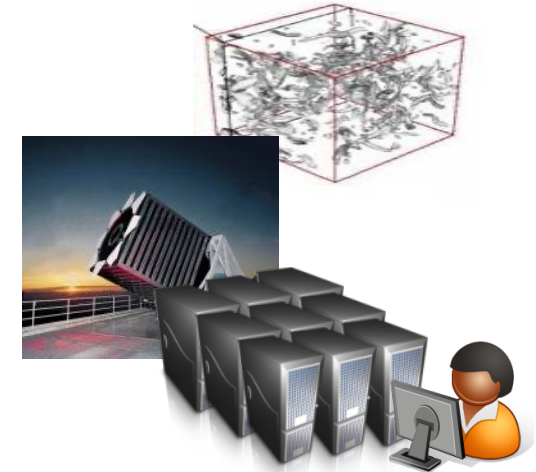


2000 : vision E-Science : quatrième paradigme sciences physiques et expérimentales

- ▶ Thousand years ago – **Experimental Science**
 - Description of natural phenomena
 - ▶ Last few hundred years – **Theoretical Science**
 - Newton's Laws, Maxwell's Equations...
 - ▶ Last few decades – **Computational Science**
 - Simulation of complex phenomena
 - ▶ Today – **eScience or Data-centric Science**
 - Unify theory, experiment, and simulation
 - Using data exploration and data mining
 - Data captured by instruments
 - Data generated by simulations
 - Data generated by sensor networks
 - Scientists over-whelmed with data
 - Computer Science and IT companies have technologies that will help
- (With thanks to Jim Gray)



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



... Merci à Tony Hey (Microsoft Research)

Dilib, une boîte à outils Sxml

- ▶ SXML : XML lite (mais JSON+)
 - Compatible avec les outils Unix
 - Un document = Une ligne Unix
- ▶ Origine INIST puis LORIA
 - 1990 : Ilib : ISO 2709 (MARC, Pascal...)
 - Un LEGO pour les corpus
 - 2000 : Dilib : métadonnées hétérogènes
- ▶ 2018 : LorExplor / ISTEEX
 - traiter du corpus volumineux,
 - Textuel, multi-dtd
 - Réseau MediaWiki
 - Générations de modèles wiki
 - Robots



```
<index>
  <kw>Requiem</kw>
  <list>
    <item>004321</item>
    <item>012345</item>
  </list>
  <f>2</f>
</index>
```

Programme / plan

- ▶ Ce matin : wikis et hypertextes pour la science et la culture
- ▶ Après-midi
 - Document structuré, Unix, XML,
 - Serveurs d'exploration
 - Curation de données

Le document structuré

- ▶ Dans les années 80, on applique aux documents les mécanismes de structuration des programmes.

Application au Web (html)

Acte II, Scène 2

DON RODRIGUE À moi, Comte, deux mots.

LE COMTE Parle.

DON RODRIGUE Ôte-moi d'un doute.
Connais-tu bien Don Diègue ?

LE COMTE Oui.

DON RODRIGUE Parlons bas, écoute.
Sais-tu que ce vieillard fut la même vertu,
La vaillance et l'honneur de son temps ? Le sais-tu ?

```
<h1>Acte II, Scène 2</h1>  
<br/> <b>DON RODRIGUE</b> À moi Comte, deux mots.  
<br/> <b>LE COMTE</b>&nbsp;&nbsp;&nbsp;... &nbsp;&nbsp;&nbsp;Parle
```

La Text Encoding Initiative (TEI)

Corpus pour l'ordinateur, lisible par l'homme

```
<div type="Act" n="I"><head>Acte II</head>
  <div type="Scene" n="1"><head>Scène 2</head>
    <sp><speaker>Rodrigue</speaker>
      <l part="i">À moi, comte, deux mots.</l></sp>
    <sp><speaker>Comte</speaker><l part="m">Parle</l></sp>
    <sp><speaker>Rodrigue</speaker>
      <l part="f">Ôte-moi d'un doute</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="i">Connais-tu bien Don Diègue ?</l></sp>
    <sp><speaker>Comte</speaker><l part="m">Oui</l></sp>
    <sp><speaker>Rodrigue</speaker>
      <l part="f">Parlons bas, écoute.</l>
      <l>Sais-tu que ce vieillard fut la même vertu,</l>
      <l>La vaillance et l'honneur de son temps ? Le sais-tu ?</l></sp>
    ...
  </div>
  ...
</div>
```

Acte II, Scène 2

DON RODRIGUE À moi, Comte, deux mots.

LE COMTE Parle.

DON RODRIGUE Ôte-moi d'un doute.

Connais-tu bien Don Diègue ?

LE COMTE Oui.

DON RODRIGUE Parlons bas, écoute.

Sais-tu que ce vieillard fut la même vertu,
La vaillance et l'honneur de son temps ? Le sais-tu ?

Dans une bibliothèque numérique, on peut trouver tous les passages où parle Rodrigue...

On peut aussi représenter des [signalements bibliographiques](#)

Codification des notices bibliographiques

A09 01 1 FRE @1 L'Ecrivain et le grand homme
A12 01 1 @1 DUFIEF (Pierre-Jean) @9 ed.
A15 01 @1 Université de Bretagne Occidentale @3 FRA @Z 1 aut.
A15 02 @1 UMR 6563 du CNRS @3 FRA @Z 1 aut.
A15 03 @1 Centre d'Etude des Correspondances de Brest @3 FRA @Z 1 aut.

```
<fA12 i1="01" i2="1">  
  <s1>DUFIEF (Pierre-Jean)</s1>  
  <s9>ed.</s9>  
</fA12>  
<fA15 i1="01">  
  <s1>Université de Bretagne Occidentale</s1>  
  <s3>FRA</s3>  
  <sZ>1 aut.</sZ>  
</fA15>  
<fA15 i1="02">  
  <s1>UMR 6563 du CNRS</s1>  
  <s3>FRA</s3>  
  <sZ>1 aut.</sZ>  
</fA15>  
<fA15 i1="03">  
  <s1>Centre d'Etude des Correspondances de Brest</s1>  
  <s3>FRA</s3>  
  <sZ>1 aut.</sZ>  
</fA15>
```

Unix -> Linux

- ▶ Un système conçu par des informaticiens en 1971
- ▶ En langage C (relativement évolué)
- ▶ Pour leur faciliter la vie...

Or les informaticiens manipulent des ensembles textuels
Les programmes et leur documentation

Unix est donc un système d'exploitation pour manipuler
des corpus. !

Unix – redirections

- ▶ Une commande Unix
 - Capte un flux d'entrée (stdin)
 - Pour produire un flux de sortie (stdout)

Exemples

```
grep Roland < ManuscritOxford
```

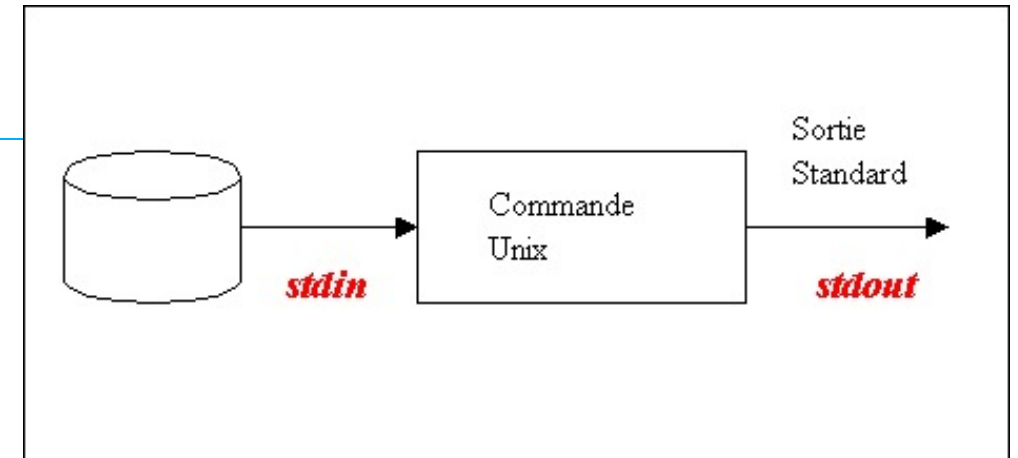
Édite les vers contenant Roland dans le manuscrit d'Oxford

```
grep Roland < ManuscritOxford > versRoland
```

Range dans un fichier les vers en question

```
grep Olivier < versRoland
```

Édite les vers contenant Olivier (ici : et Roland)



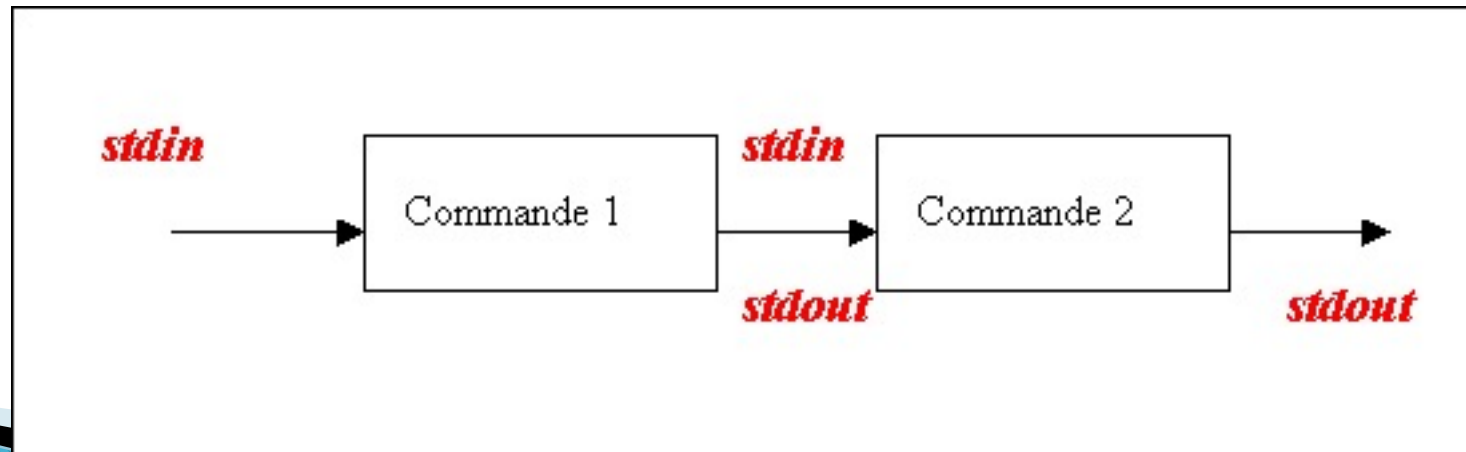
Les pipes dans Unix

- ▶ Permettent d'enchaîner des commandes
- ▶ Exemples

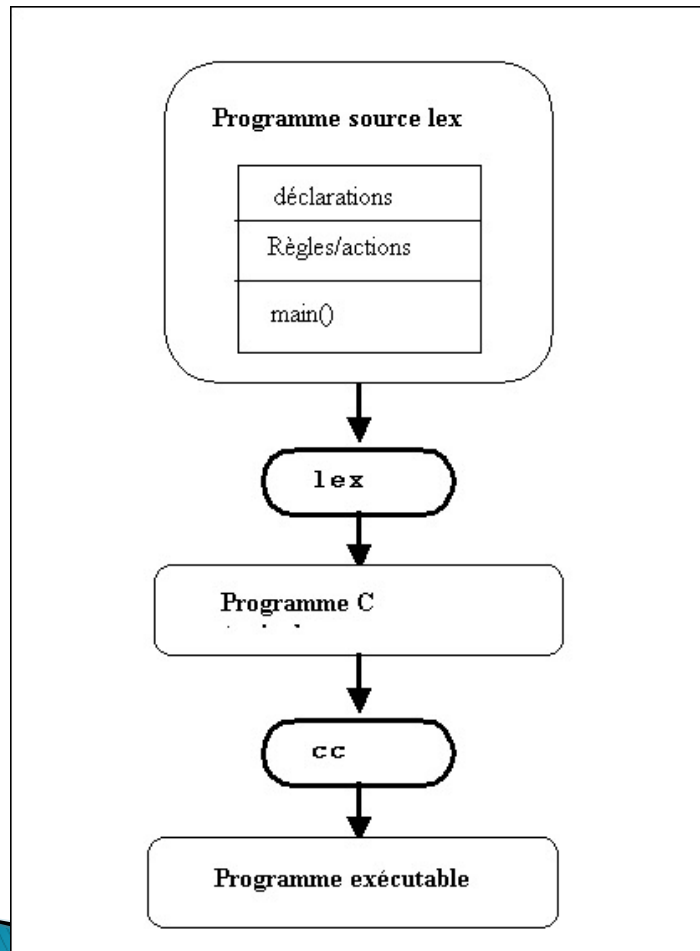
```
grep Roland < ManuscritOxford
```

```
grep Roland < ManuscritOxford | grep Olivier
```

```
grep Roland < ManuscritOxford | grep Olivier | wc
```



Unix – langage C – lex (analyseur lexical)



```
%%  
chevals      printf("chevaux");  
hibous      printf("hiboux");  
%%  
main() {  
    yylex();  
}
```

Exemple : génération de tableau MediaWiki

```
%%  
^[A-Za-z«]    {printf ("|\n|"); ECHO;}  
^[1-9][0-9]+  printf ("%s&nbsp;&nbsp;&nbsp;\n|", yytext);  
[ ]*\n         printf ("\n|-\n");  
%%  
main()  
{  
printf ("{| \n|-\n");  
yylex();  
printf ("|}");  
}
```

```
1Carles li Reis, nostre emperere magnes,  
Set anz tuz pleins ad estet en Espagne :
```

```
{|  
-  
|1&nbsp; &nbsp;&nbsp;  
|Carles li Reis, nostre emperere magnes,  
-  
|  
|Set anz tuz pleins ad estet en Espagne :  
-  
|}
```

- 1 Carles li Reis, nostre emperere magnes,
Set anz tuz pleins ad estet en Espagne :

Rappel, écrire la musique



<score>

```
{  
  \time 2/4  
  \clef bass  
  c4 c g g a a g2  
}
```

Les commandes sont précédées d'un \

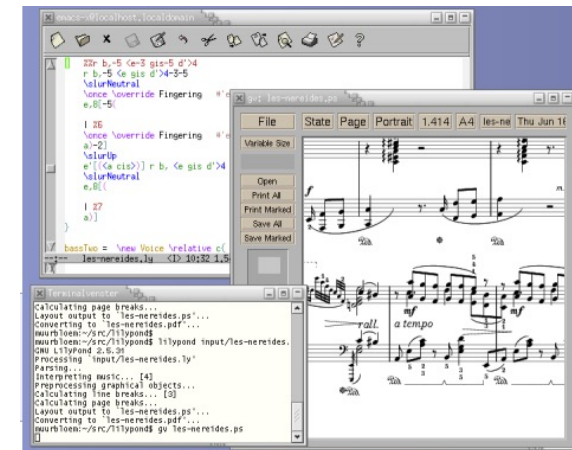
Les lettres représentent les notes

Les chiffres annoncent la durée

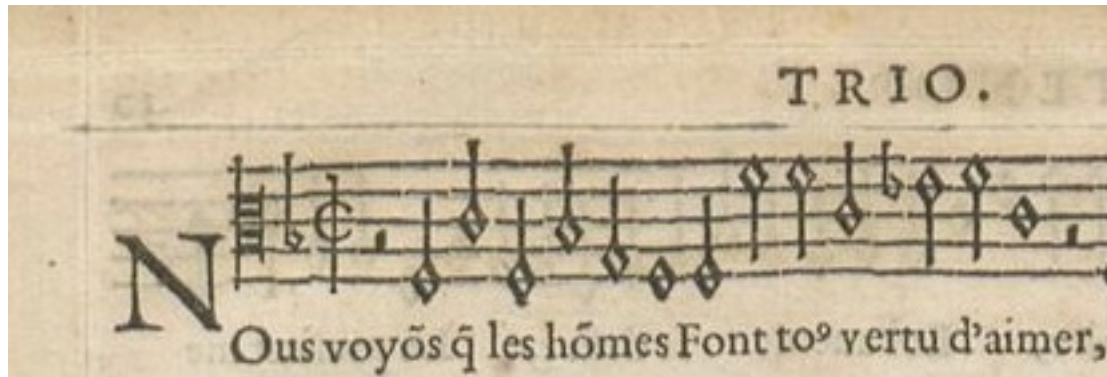
</score>



a = la
b = si
c = do
d = ré
...



Manipulation du wikitexte et de Lilypond avec des analyseurs syntaxiques (Lex)



```
%%  
a      printf ("fis");  
bes    printf ("g");  
c      printf ("a");  
d      printf ("b");  
e      printf ("cis");  
ees    printf ("c");  
f      printf ("d");  
g      printf ("e");  
%%  
main()  
{  
    yylex();  
}
```



Manipulation de documents XML

- ▶ Un programme Unix gère des lignes
- ▶ Une ligne est terminée par un « retour chariot » (\n)
- ▶ Un document XML sera manipulé sur une ligne

```
<doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>  
<doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

```
cat < monCorpusBD \
| grep "<date>194" \
| wc
```

Imprime le nombre de documents édités entre 1940 et 1949.

Boîte à outil XML



- ▶ Ilib 1990 (INIST) – Dilib 1992 (Loria)
- ▶ Premières expérience en lex (préfiguration XSLT...)

```
<doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>  
<doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

```
%%  
"<doc>"      ;  
"</doc>"     printf("\n\n");  
"<tit>"      printf("<b>");  
"</tit>"     printf("</b>\n");  
"<aut>"      printf("::");  
"</aut>"     ;  
"<date>"     printf (" (");  
"</date>"    printf (")");  
%%
```

```
<b>Tintin au Congo</b>  
::Hergé (1948)  
  
<b>Tintin en Amérique</b>  
::Hergé (1949)
```

Tintin au Congo

Hergé (1948)

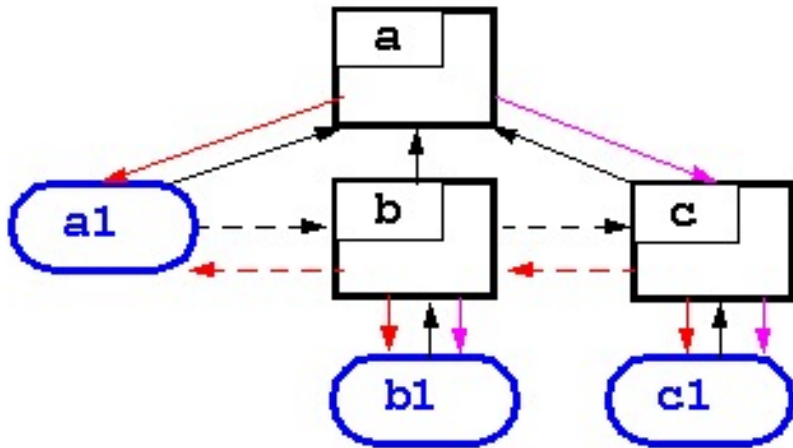
Tintin en Amérique

Hergé (1949)

Programmation DOM Document Object Model



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



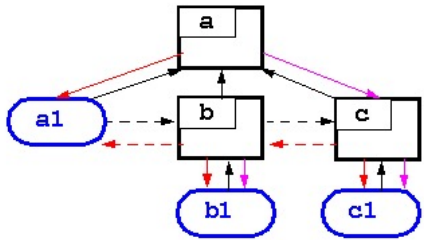
```
#include "SxmlNode.h"
main()
{
    SxmlNode *root, *tit;
    root =SxmlElementCreate("doc");
    tit= SxmlLeafCreate("tit", "Tintin au Congo");
    SxmlAppendChild (root, tit);
    SxmlAppendChild (root,
        SxmlLeafCreate("aut", "Hergé"));
    SxmlAppendChild (root,
        SxmlLeafCreate("date", "1949"));
    SxmlPrint(root);
    putchar('\n');
    exit(0);
}
```

```
<doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>
```

Parser DOM

Pour manipuler des flots de données

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



```
#include "XmlNode.h"
#include <stdio.h>
main()
{
    XmlNode *myNodeTest;
    ...
    myNodeTest=XmlNodeFromString("<a>a1<b>b1</b><c>c1</c></a>");
    ...
}
```

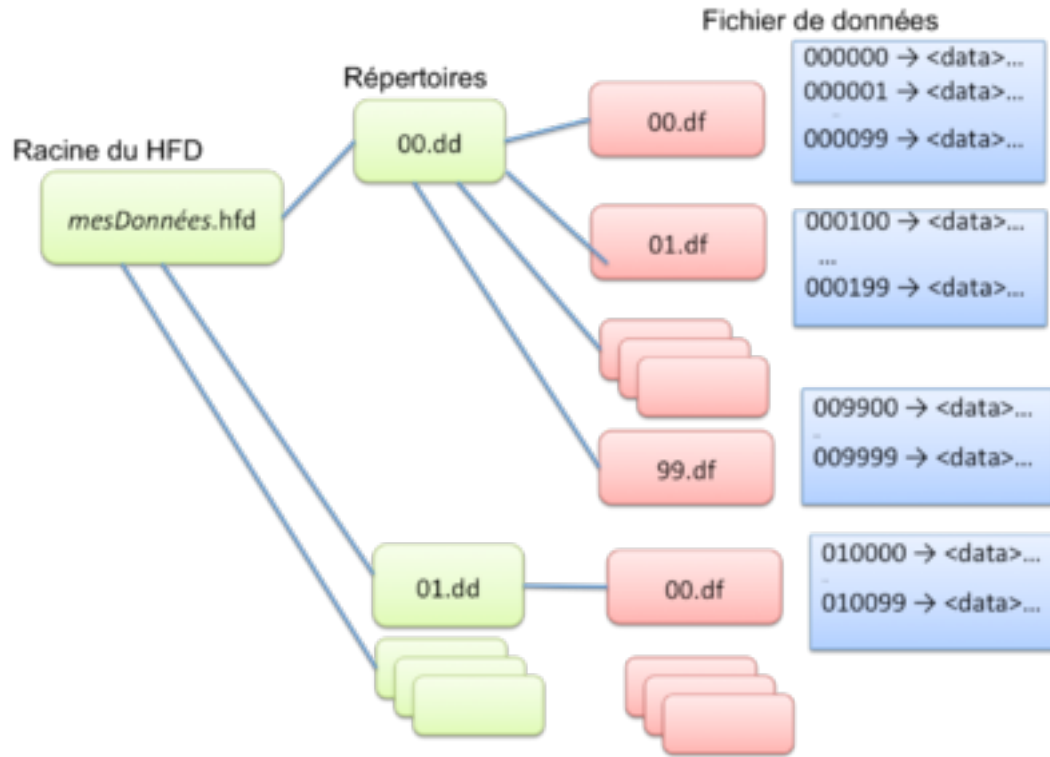
```
<doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>
<doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

```
#include "XmlNode.h"
main()
{
    XmlNode *docu ,*titre;
    while(docu=XmlNodeGetDocumentElement())
    {
        if (titre=XmlNodeFirstChild(docu))
        {if (strcmp(XmlNodeName(titre), "tit")==0)
            {XmlNodePrint(titre);putchar('\n');
        }
    }
}
```

```
<tit>Tintin au Congo</tit>
<tit>Tintin en Amérique</tit>
```

Dilib / Organisation HFD

Pour ranger les corpus



Clé HFD

Chaque document est précédé

- d'une clé (6 caractères)
- Avec une tabulation

Commandes

HfdCat pour créer un flot de données

```
cat < monCorpusBD \
| grep "<date>194" \
| wc
```

```
HfdCat < monCorpusHfd \
| grep Hergé
```

```
001230 <doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>
002345 <doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

Dilib / Commandes d'extraction basées sur des Xpath (chemins de balise)

- ▶ Commande SxmlSelect
 - Options -g (grep) -p (print)

```
001230 <doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>  
002345 <doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

```
HfdCat < monCorpusHfd \  
| grep Hergé \  
| SxmlSelect -g doc/tit/1 -g doc/date/1 -p @g2 -p @g1 \  
| sort -r
```

```
1949 Tintin en Amérique  
1948 Tintin au Congo
```

Dilib / création de listes d'index

► Commande IndexBuildRec

```
001230 <doc><tit>Tintin au Congo</tit><aut>Hergé</aut><date>1948</date><doc>  
002345 <doc><tit>Tintin en Amérique</tit><aut>Hergé</aut><date>1949</date><doc>
```

```
HfdCat < monCorpusHfd \  
| SxmlSelect -g doc/aut/1 -p @g1 -p @1 \  
| sort \  
| IndexBuildRec \  
| grep Hergé \  
| SxmlIndent
```

```
Goscinny 000000  
...  
Hergé 001230
```

```
...  
Hergé 001230  
Hergé 002345
```

```
<idx>  
  <k>Hergé</k>  
  <f>2</f>  
  <l>  
    <e>001230</e>  
    <e>002345</e>  
  </l>  
</idx>
```

Exemples récapitulatifs



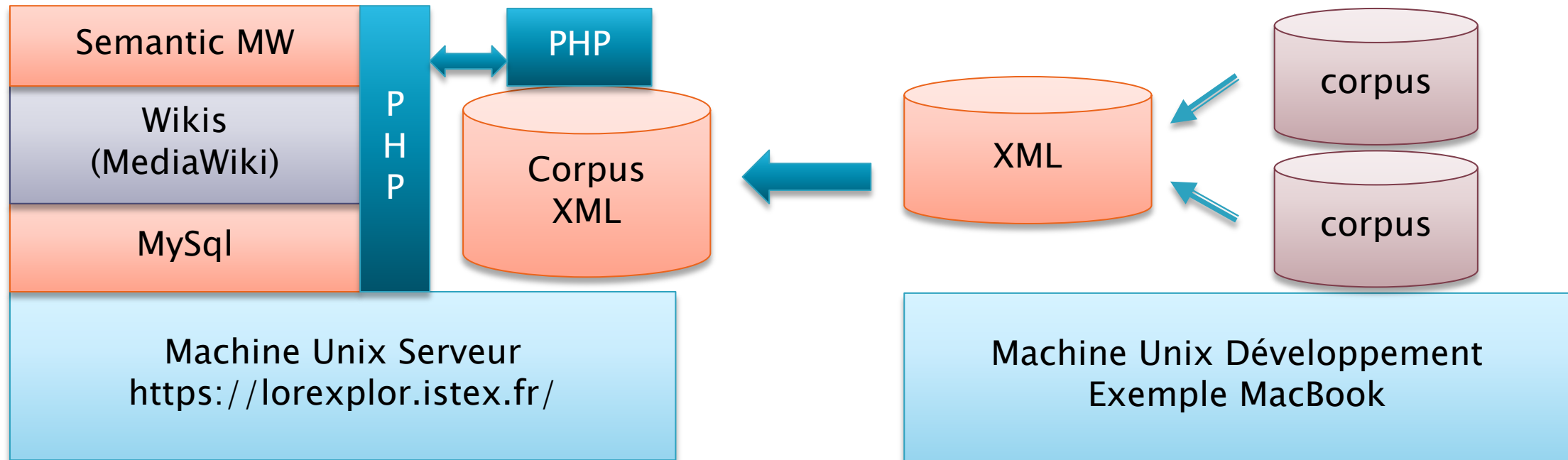
- ▶ Quelles sont les œuvres de Mozart les plus citées dans un corpus ?
 - Idée générale : utiliser le catalogue Köchel
 - Résultat : Sonate KV. 448

```
HfdCat Data/Main/Exploration/biblio.hfd \
| SxmlFindText -r "[K][Vv]*[ \.]*[0-9][0-9]* » \
| SxmlSelect -p @5 -p @1 | sort | IndexBuildRec
```

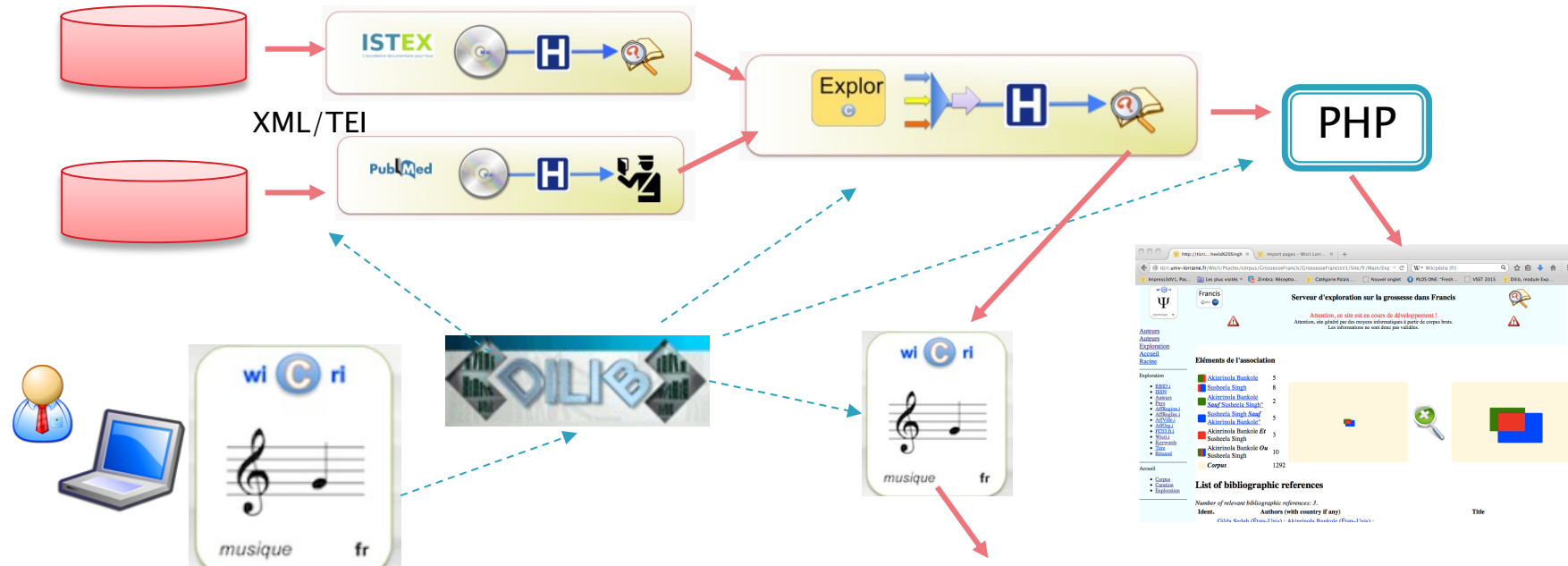
- ▶ Quelles sont les applications de « *dance therapy* » avec une dimension artistique ?
 - Recherche de présence de chorégraphes (nom-prénom) en utilisant un filtre créé pour les noms binomiaux

Machine serveur – machine développement

Unix, langage C, PHP, XML, JSON, etc...



ISTEX – Serveur – génération



[[Explor plateforme MozartV1 /Carte France|taille=400]]



Pays	Région	Villes
1. France (67) ↗	1. Californie (11) ↗	1. Paris (9) ↗
2. États-Unis (31) ↗	2. Île-de-France (9) ↗	2. Marseille (5) ↗
3. Royaume-Uni (14) ↗	3. Occitanie (région administrative) (7) ↗	3. Montpellier (4) ↗
4. Allemagne (14) ↗	4. Massachusetts (6) ↗	4. Londres (4) ↗
5. Canada (11) ↗	5. Angleterre (6) ↗	5. Grenoble (4) ↗
6. Italie (10) ↗	6. État de New York (5) ↗	6. Berlin (4) ↗
7. Espagne (8) ↗	7. Maryland (5) ↗	7. Toulouse (3) ↗
8. Suisse (6) ↗	8. Caroline du Nord (5) ↗	8. Prague (3) ↗
9. Australie (6) ↗	9. Arizona (5) ↗	9. Montréal (3) ↗
10. Pays-Bas (5) ↗	10. Washington (État) (4) ↗	10. Zurich (2) ↗
Mots-clés anglais	Mots des titres	ISSN/revue
1. Astrophysics (3) ↗	1. data (10) ↗	1. SPIE proceedings series (6) ↗
2. State of the art (2) ↗	2. analysis (7) ↗	2. 1932-6203 (5) ↗
3. Software package (2) ↗	3. software (6) ↗	3. Lecture Notes in Computer Science (4) ↗
4. Real time (2) ↗	4. microbial (6) ↗	4. Eos Trans. AGU (3) ↗
5. Quebec (2) ↗	5. marine (5) ↗	5. 2034-9250 (3) ↗
6. Perspective (2) ↗	6. genome (5) ↗	6. 1091-6490 (3) ↗
7. Open source software (2) ↗	7. distributed (5) ↗	7. 0096-9941 (3) ↗
8. Measurement sensor (2) ↗	8. genomic (4) ↗	8. 0027-8424 (3) ↗
9. Library network (2) ↗	9. control (4) ↗	9. 2047-217X (2) ↗
10. Information policy (2) ↗	10. web (3) ↗	10. 1545-7885 (2) ↗



Combinaison d'index : AutAff

- ▶ Auteurs réduits à
 - Nom initiale prénom
 - + Affiliations
- ▶ Destiné initialement à la curation
- ▶ A l'expérience : détection des acteurs

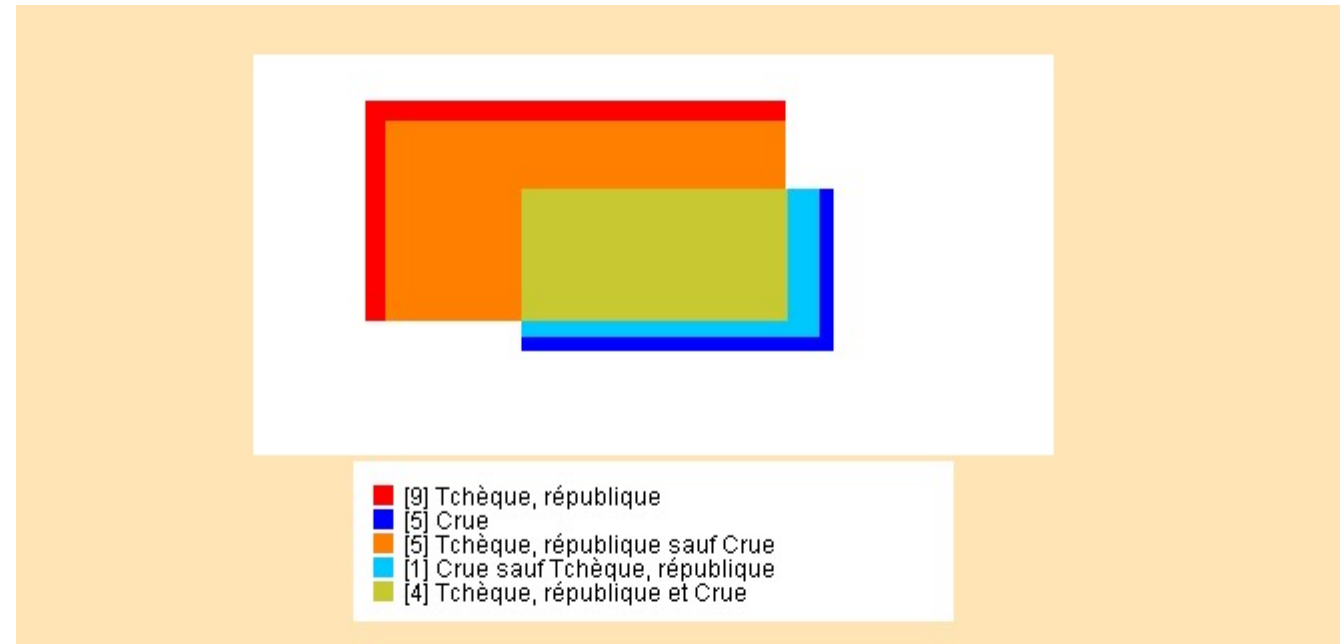
365 [Lees A](#)
 325 [Lang A](#)
 303 [Louis E](#)
 279 [Poewe W](#)
 251 [Bhatia K](#)
 248 [Quinn N](#)
 218 [Goetz C](#)
 216 [Jankovic J](#)

Department of Neurology, Juntendo University School of Medicine, Tokyo	004177
Department of Neurology, Juntendo University School of Medicine, Urayasu Hospital, Tokyo, Japan	002388
Department of Neurology, Juntendo University, School of Medicine, Tokyo, Bunkyo-ku, Japan	004111
Department of Neurology, Jutendo University, School of Medicine, Tokyo, Japan	003517
Department of Neurology, Research Institute for Diseases of Old Age, Juntendo University School of Medicine, Tokyo, Japan	000C30
Department of Neurology, Research Institute for Diseases of Old Ages,	000393

Y. Mizuno	Department of Neurology, Juntendo University School of Medicine, Tokyo, Japan	002384
	INSERM U 289 & Fédération de Neurologie, Hôpital de la Salpêtrière-47, Bd de l'Hôpital-75651 Paris, Cedex 13, France	003B80
	NONE	002384 003B80
Yoshi Mizuno	Department of Neurology, School of Medicine, Jutendo University School of Medicine, Bunkyo-Ku, Tokyo, Japan	000610
	Juntendo University Tokyo, Japan	003891
	NONE	000610 003891
Yoshikino Mizuno	Department of Neurology, Juntendo University School of Medicine, Bunkyo-ku, Tokyo, Japan	000622
	NONE	000622
	Research Institute for Diseases of Old Ages, Juntendo University School of Medicine, Bunkyo-ku, Tokyo, Japan	000622

Associations

- ▶ Permettent de visualiser les relations liant 2 concepts

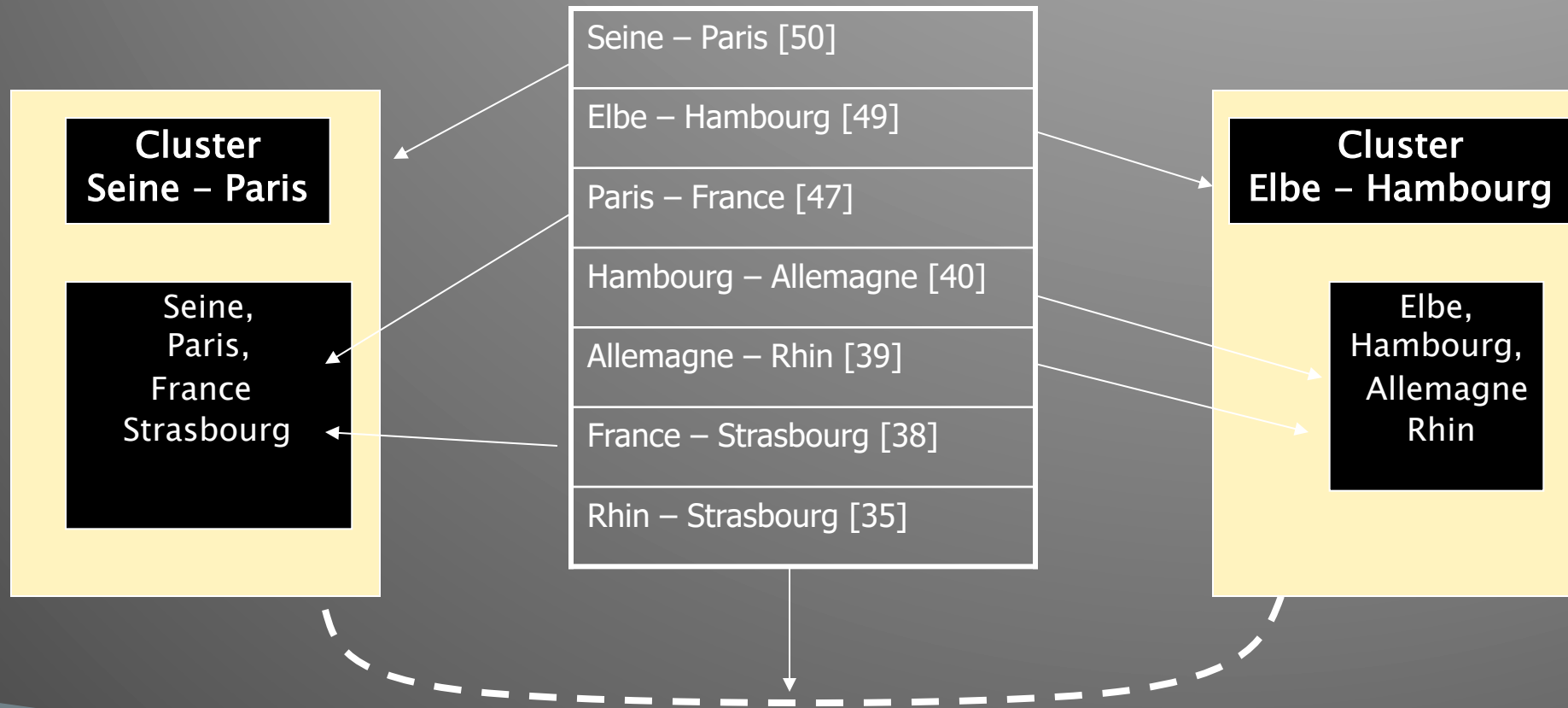


Liste d'associations

- ▶ Intéressant mais difficilement utilisable
- ▶ Exemple : base sur l'hydrographie en Allemagne

Nom des associations	Fij
Bayern - Allemagne	34
Précipitation - Allemagne	30
Bassin-versant - Allemagne	30
Pollution - Allemagne	26
Erosion des sols - Allemagne	25
Pollution de l'eau - Allemagne	24
Hydrologie - Allemagne	24
Cours d'eau - Allemagne	24
Sol - Allemagne	22
Modèle - Allemagne	22
Eau - Allemagne	21
Pollution de l'eau - Pollution	19
Allemagne - Action anthropique	19
Protection de la nature - Allemagne	17
Utilisation du sol - Allemagne	16
Fluviale - Allemagne	16
Ecoulement - Allemagne	15
Baden-Württemberg - Allemagne	15
Hessen - Allemagne	14
Végétation - Allemagne	13

Clusterisation



Trésor de la langue française (dictionnaire)

- ▶ 1960 création de l'ARTLF (Recteur Paul Imbs)
 - Un Trésor (= corpus) numérique pour la recherche en linguistique
 - Valorisable par un pari : un dictionnaire
- ▶ 1963 : Acquisition du Gamma 60
- ▶ 1969 : 600 ouvrages saisis
- ▶ 1971 : Tome 1 du dictionnaire
- ▶ 1972 : Passage au Cii 10070
- ▶ 1982 : ARTLF avec Chicago
- ▶ 1994 : dernier tome

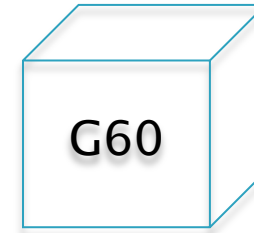


Outils de production, fiches, groupes binaires

- ▶ Le Gamma 60 produit des concordances
 - parfait pour les mots de faible fréquence.
 - Exemple : *dodécaphonique* (4)
 - Les concordances sont éditées sur « listing »
 - Dans les autres cas : les groupes binaires
 - Co occurrences sur 5 termes sémantiques avant ou après un exemple
 - *Le corbeau fait son nid en haut des arbres.*
 - Le système propose (sur un listing)
 - un ensemble de mots associés avec des listes d'exemples
 - Multiples algorithmes de classement sélection
 - en fonction des fréquences des termes

[...] *pour chaque mot*
[...] *pour chaque occurrence*
un contexte de trois lignes,
le mot étudié figurant
obligatoirement
dans la ligne du milieu

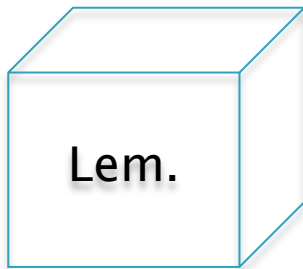
Groupes binaires (pour la mémoire)



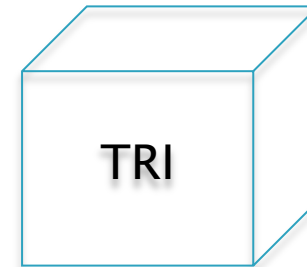
Texte corrigé



texte

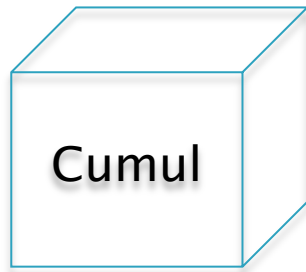


mots



Mots triés

Groupes binaires (suite)



Mots triés

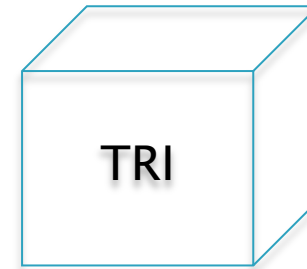
Abaca
Abaca
Abat

....

Couples
Fréquence - mot

2 abaca
5 abat

...



Couples triés

...

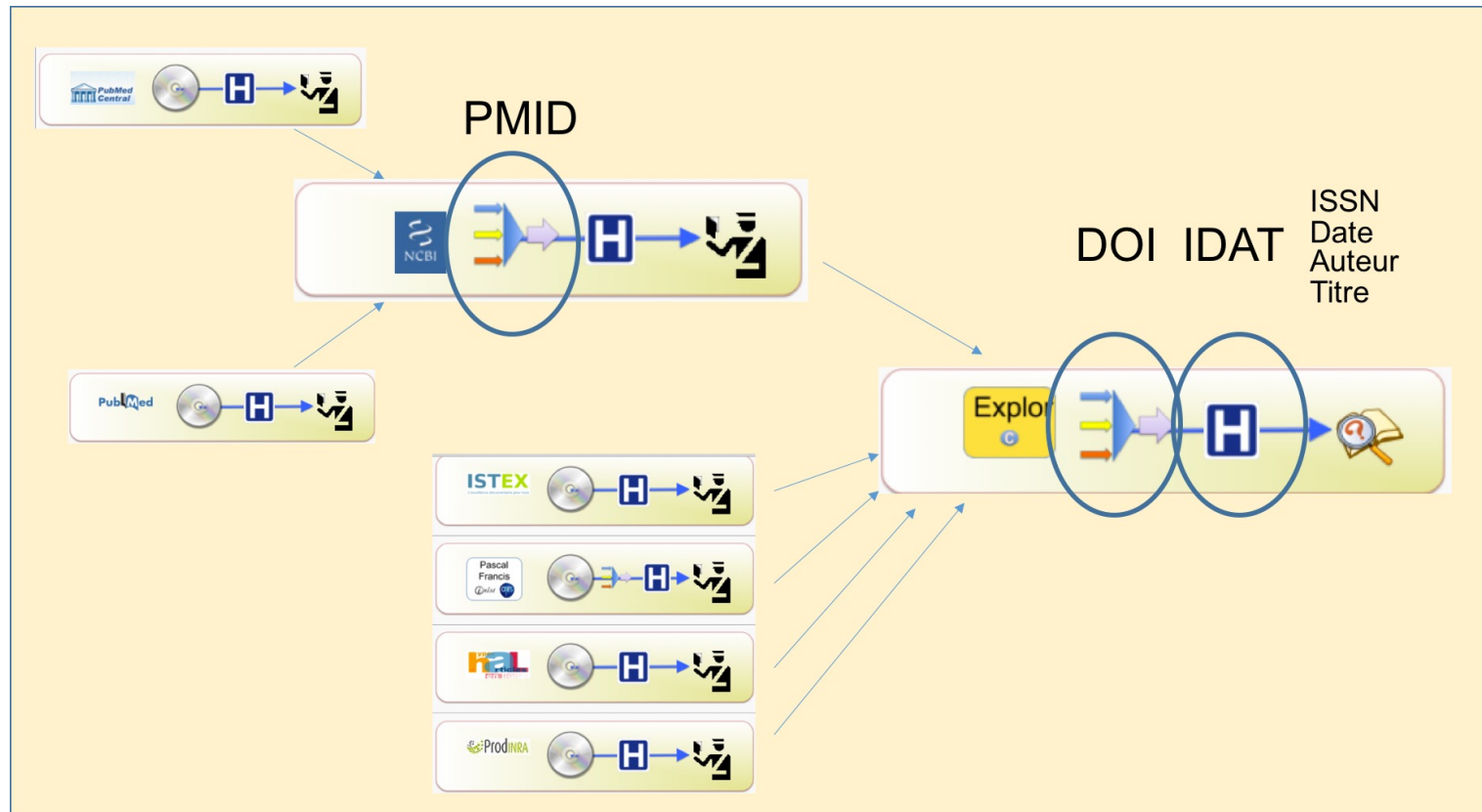
5 abat

...

2 abaca

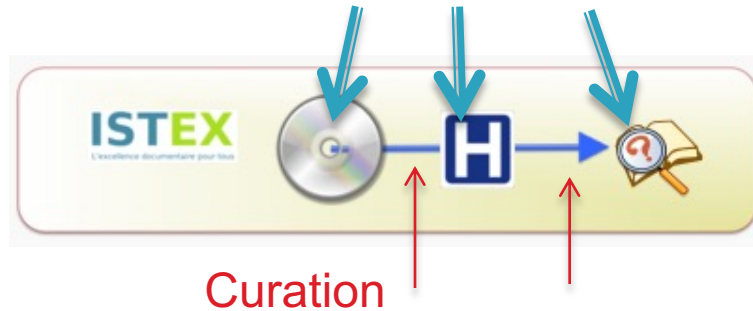
ETC durée 1 mois

Enrichissement : dédoublonnage ISTEK / Pascal / Hal / MEDLINE...



Serveur d'exploration

Systeme d'information orienté exploration



http://ticri...heela%20Singh Import pages - Wicri Lorr... +

ticri.univ-lorraine.fr/Wicri/Psycho/corpus/GrossesseFrancis/GrossesseFrancisV1/Site/fr/Main/Exp W - Wikipédia (fr)

Francis **Serveur d'exploration sur la grossesse dans Francis**

Attention, ce site est en cours de développement !
 Attention, site généré par des moyens informatiques à partir de corpus bruts.
 Les informations ne sont donc pas validées.

Eléments de l'association

	Akinrinola Bankole	5
	Susheela Singh	8
	Akinrinola Bankole	2
	Sauf Susheela Singh	
	Susheela Singh Sauf	5
	Akinrinola Bankole	5
	Akinrinola Bankole Et	3
	Susheela Singh	
	Akinrinola Bankole Ou	10
	Susheela Singh	
	Corpus	1292

List of bibliographic references

Number of relevant bibliographic references: 3.

Ident.	Authors (with country if any)	Title
	Gilda Sedeh (Frans Unis) ; Akinrinola Bankole (Frans Unis)	



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	H																		He
2	Li	Be											B	C	N	O	F	Ne	
3	Na	Mg											Al	Si	P	S	Cl	Ar	
4	K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr	
5	Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe	
6	Cs	Ba	Lu	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn	
7	Fr	Ra	Lr	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Uut	Uuq	Uup	Uuh	Uus	Uuo	

Tableau périodique des éléments chimiques

Serveur d'exploration

Parcourir les index

Pays

1. France (67) [↗](#)
2. États-Unis (31) [↗](#)
3. Royaume-Uni (14) [↗](#)
4. Allemagne (14) [↗](#)
5. Canada (11) [↗](#)
6. Italie (10) [↗](#)
7. Espagne (8) [↗](#)
8. Suisse (6) [↗](#)
9. Australie (6) [↗](#)
10. Pays-Bas (5) [↗](#)

Région

1. Californie (11) [↗](#)
2. Île-de-France (9) [↗](#)
3. Occitanie (région administrative) (7) [↗](#)
4. Massachusetts (6) [↗](#)
5. Angleterre (6) [↗](#)
6. État de New York (5) [↗](#)
7. Maryland (5) [↗](#)
8. Caroline du Nord (5) [↗](#)
9. Arizona (5) [↗](#)
10. Washington (État) (4) [↗](#)

Villes

1. Paris (9) [↗](#)
2. Marseille (5) [↗](#)
3. Montpellier (4) [↗](#)
4. Londres (4) [↗](#)
5. Grenoble (4) [↗](#)
6. Berlin (4) [↗](#)
7. Toulouse (3) [↗](#)
8. Prague (3) [↗](#)
9. Montréal (3) [↗](#)
10. Zurich (2) [↗](#)

Mots-clés anglais

:

1. Astrophysics (3) [↗](#)
2. State of the art (2) [↗](#)
3. Software package (2) [↗](#)
4. Real time (2) [↗](#)
5. Quebec (2) [↗](#)
6. Perspective (2) [↗](#)
7. Open source software (2) [↗](#)
8. Measurement sensor (2) [↗](#)
9. Library network (2) [↗](#)
10. Information policy (2) [↗](#)

Mots des titres

1. data (10) [↗](#)
2. analysis (7) [↗](#)
3. software (6) [↗](#)
4. microbial (6) [↗](#)
5. marine (5) [↗](#)
6. genome (5) [↗](#)
7. distributed (5) [↗](#)
8. genomic (4) [↗](#)
9. control (4) [↗](#)
10. web (3) [↗](#)

ISSN/revue

1. SPIE proceedings series (6) [↗](#)
2. 1932-6203 (5) [↗](#)
3. Lecture Notes in Computer Science (4) [↗](#)
4. Eos Trans. AGU (3) [↗](#)
5. 2324-9250 (3) [↗](#)
6. 1091-6490 (3) [↗](#)
7. 0096-3941 (3) [↗](#)
8. 0027-8424 (3) [↗](#)
9. 2047-217X (2) [↗](#)
10. 1545-7885 (2) [↗](#)

Corpus : méfiance / curation

▶ Exemple : Mozart

- 15.000 documents (Musique + médecine)
- Quelques problèmes de type « avenue Mozart »
- Plus sérieux :
 - Musique : peu de signalement d'affiliations
 - Médecine : forte politique d'affiliations
- Les statistiques se focalisent sur la médecine...

▶ Exemple : Parkinson en France

- Parkinson : 90.000 documents
- Extrait de 4000 documents :
 - peu de bruit
- Parkinson en France :
 - beaucoup de bruit.

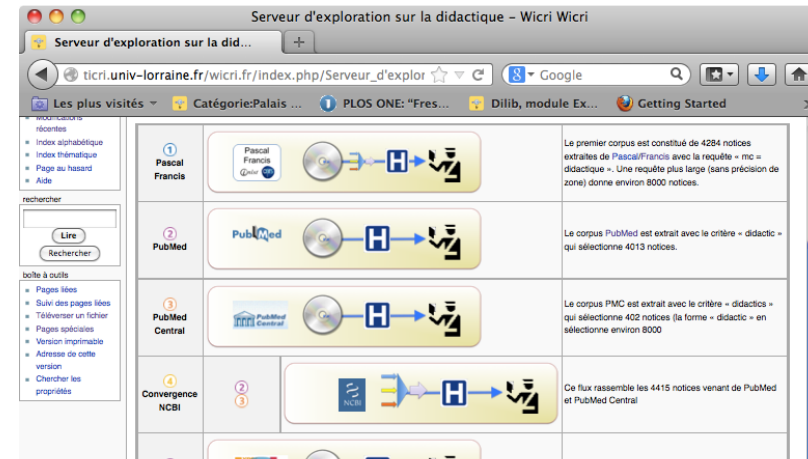


▶ Quelle formation donner à un bibliothécaire pour accompagner un chercheur dans une démarche de curation?

Curation des données



- ▶ Exemple : identifier les pays dans un contexte hétérogène



numérique	alpha -3	alpha -2	Nom français usuel	Nom ISO du pays ou territoire
004	AFG	AF	Afghanistan	AFGHANISTAN
710	ZAF	ZA	Afrique du Sud	AFRIQUE DU SUD
248	ALA	AX	Åland	Modèle:Tri1ÅLAND, ÎLES
008	ALB	AL	Albanie	ALBANIE
012	DZA	DZ	Algérie	Modèle:Tri1ALGÉRIE
276	DEU	DE	Allemagne	ALLEMAGNE
020	AND	AD	Andorre	ANDORRE
024	AGO	AO	Angola	ANGOLA
660	AIA	AI	Anguilla	ANGUILLA

Curation des données – pays

- ▶ Codes ISO (exemple Pascal)
 - Vers le web sémantique (via Wikipédia/WikiData)

```

pA A01 01 1 @0 0302-9743
A05 @2 1375
A08 01 1 ENG @1 Hyperbook data modeling
A09 01 1 ENG @1 Electronic publishing, artistic
digital typography : Saint Malo, 1998
A11 01 1 @1 FRÖHLICH (P.)
A11 02 1 @1 HENZE (N.)
A11 03 1 @1 NEJDL (W.)
A12 01 1 @1 HERSCH (Roger D.) @9 ed.
A12 02 1 @1 ANDRE (Jacques) @9 ed.
A12 03 1 @1 BROWN (Heather) @9 ed.
A14 01 @1 Institut für Rechnergestützte V
Universität Hannover, Lange Laube
@3 DEU @Z 1 aut. @Z 2 aut. @Z 3 au
    
```

numé- rique	alpha -3	alpha -2	Nom français usuel	Nom ISO du pays ou territoire
004	AFG	AF	Afghanistan	AFGHANISTAN
710	ZAF	ZA	Afrique du Sud	AFRIQUE DU SUD
248	ALA	AX	Åland	Modèle:Tri1ÅLAND, ÎLES
008	ALB	AL	Albanie	ALBANIE
012	DZA	DZ	Algérie	Modèle:Tri1ALGÉRIE
276	DEU	DE	Allemagne	ALLEMAGNE
020	AND	AD	Andorre	ANDORRE
024	AGO	AO	Angola	ANGOLA
660	AIA	AI	Anguilla	ANGUILLA

Page récupérée de Wikipédia sur Wicri/Métadonnées

Curation des pays – Adresses

Adresses postales
(Springer, PubMed)

```
<titleInfo lang="eng">
  <title>Graph Access Pattern Diagrams (GAP-D): Towards a
  Unified Approach for Modeling Navigation over
  Hierarchical, Linear and Networked Structures</title>
</titleInfo>
<name type="personal">
  <namePart type="given">Matthias
  <namePart type="family">Keller
  <role>
    <roleTerm type="text">author</roleTerm>
  </role>
  <description>Matthias.keller@k
  <affiliation>Steinbuch Centre
  Karlsruhe Institute of Technol
  Karlsruhe, Germany</affiliatio
</name>
```

Forme française sur Wicri	Forme anglaise sur Wicri	Forme courantes
Afrique du Sud	South Africa	South Africa ; Republic of South Africa
Arabie saoudite	Saudi Arabia	Saudi Arabia
Allemagne	Germany	Germany ; Deutschland ; Federal Republic of Germany ; Bundesrepublik Deutschland ; FRG ; DDR ; West Germany ; W. Germany ; Fed. Rep. Germany ; GDR ; German Democratic Republic ; Deutsche Demokratische Republik
Argentine	Argentina	Argentina
Australie	Australia	Australia
Autriche	Austria	Austria ; Österreich

Page collective (mutualisée) sur Wicri/Métadonnées

Curation des régions



Ces types de carte sont visibles sur tous types de serveurs

Curation des régions

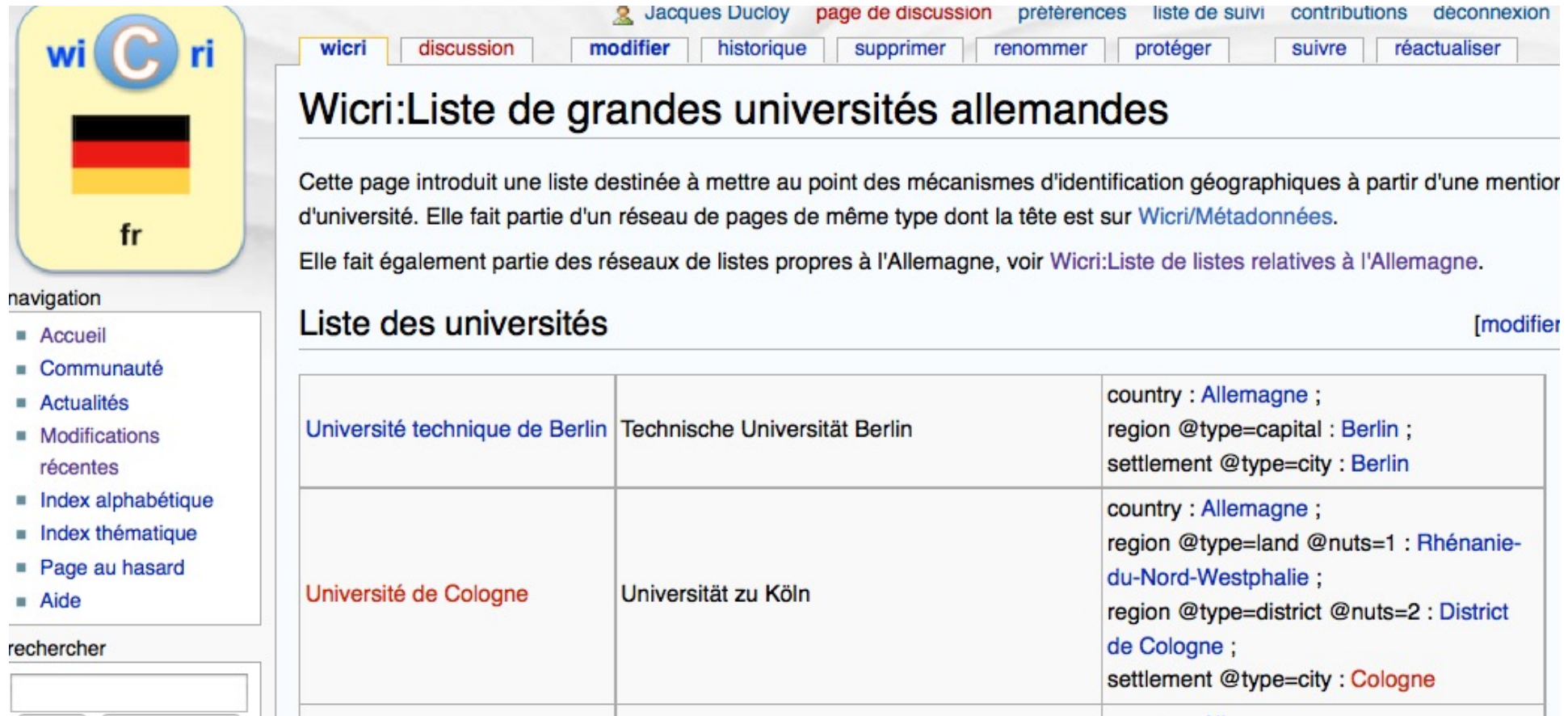
Sur Wicri/Allemagne

ville	code 4 chiffres	code 5 chiffres	formes courantes	district/land
Aix-la-Chapelle	W-5100	52056-52080	Aachen	region @type=land @nuts=1 : Rhénanie-du-Nord-Westphalie ; region @type=district @nuts=2 : District de Cologne
Augsbourg	W-8900	86000-86199	Augsburg	region @type=land @nuts=1 : Bavière ; region @type=district @nuts=2 : District de Souabe
Bayreuth	W-8580	95444-95448	Bayreuth	region @type=land @nuts=1 : Bavière ; region @type=district @nuts=2 : District de
Berlin	W-1000	10		
Bonn	W-5300	53		

Sur la machine
D'exploration

```
<r>
  <c1>
    <p>
      <k>Aix-la-Chapelle</k>
      <t>Aix-la-Chapelle</t>
    </p>
  </c1>
  <c2>
    <l>W-5100</l>
  </c2>
  <c3>
    <i>52056-52080</i>
  </c3>
  <c4>
    <l>Aachen</l>
  </c4>
  <c5>
    <region type="land" nuts="1">Rhénanie-du-Nord-Westphalie</region>
    <region type="district" nuts="2">District de Cologne</region>
  </c5>
  <c6>
    <l>
      </l>
    </c6>
  </r>
```

Curation des régions – suite



The screenshot shows a Wikiri page for 'Wicri:Liste de grandes universités allemandes'. At the top, there is a user profile for 'Jacques Ducloy' with links for 'page de discussion', 'préférences', 'liste de suivi', 'contributions', and 'déconnexion'. Below this is a navigation bar with buttons for 'wicri', 'discussion', 'modifier', 'historique', 'supprimer', 'renommer', 'protéger', 'suivre', and 'réactualiser'. The page title is 'Wicri:Liste de grandes universités allemandes'. The main text explains that the page is a list intended to identify geographical mechanisms from university mentions and is part of a network of similar pages, with a reference to 'Wicri/Métadonnées'. It also mentions a network of lists specific to Germany, with a reference to 'Wicri:Liste de listes relatives à l'Allemagne'. Below the text is a section titled 'Liste des universités' with a '[modifier]' link. A table lists two universities: 'Université technique de Berlin' (Technische Universität Berlin) and 'Université de Cologne' (Universität zu Köln). The table includes metadata for each entry, such as country, region, and settlement.

Navigation:

- Accueil
- Communauté
- Actualités
- Modifications récentes
- Index alphabétique
- Index thématique
- Page au hasard
- Aide

rechercher

Université technique de Berlin	Technische Universität Berlin	country : Allemagne ; region @type=capital : Berlin ; settlement @type=city : Berlin
Université de Cologne	Universität zu Köln	country : Allemagne ; region @type=land @nuts=1 : Rhénanie-du-Nord-Westphalie ; region @type=district @nuts=2 : District de Cologne ; settlement @type=city : Cologne

Règles de curation des données



Myriam Chimènes	Myriam Chimènes	affiliation : Institut de recherche en musicologie
Denis Herlin	Denis Herlin	affiliation : Institut de recherche en musicologie ; affiliation : Université de Tours
Paul Henry Lang	0027-4631:P. H. L.	affiliation : Université Columbia
Edward Lowinsky	Edward E. Lowinsky	affiliation @from=1961 : Université de Chicago



Université Columbia	Columbia University	country : États-Unis ; region @type=state : État de New York ; settlement @type=city : New York
Université Cornell	Cornell University	country : États-Unis ; region @type=state : État de New York ; settlement @type=city : Ithaca (New York)

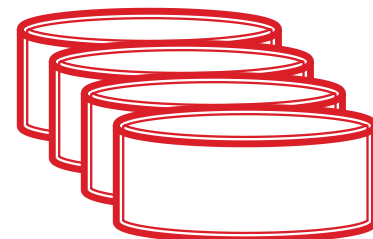
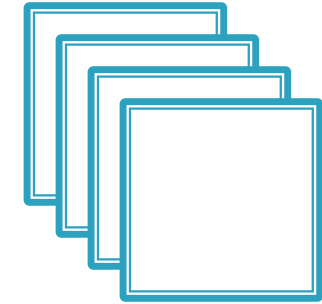
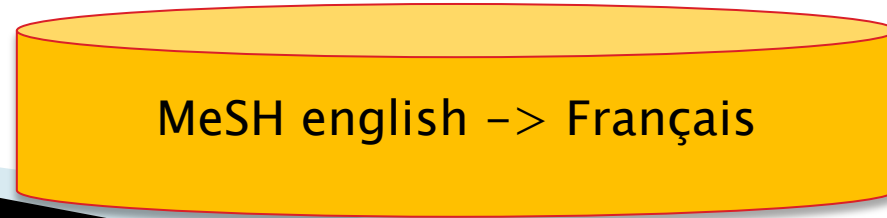
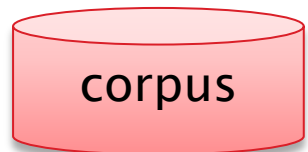
Santé : Serveurs à génération rapide (5')



NlmPubMedExplorCorpus -q influenza -s 1000



Part	Hyper	View
1. France (17) [p]	1. Coleridge (17) [p]	1. Paris (17) [p]
2. Blue-eyes (17) [p]	2. Noach-Peters (17) [p]	2. November (17) [p]
3. Proseman (17) [p]	3. Soudan: High administrative (17) [p]	3. November (17) [p]
4. Amoy (17) [p]	4. Resolutions (17) [p]	4. London (17) [p]
5. Canada (17) [p]	5. Agnes (17) [p]	5. Constantin (17) [p]
6. Italy (17) [p]	6. Soudan (17) [p]	6. Berlin (17) [p]
7. France (17) [p]	7. Soudan (17) [p]	7. November (17) [p]
8. Blue-eyes (17) [p]	8. Soudan (17) [p]	8. France (17) [p]
9. Soudan (17) [p]	9. Soudan (17) [p]	9. November (17) [p]
10. Proseman (17) [p]	10. November (17) [p]	10. June (17) [p]



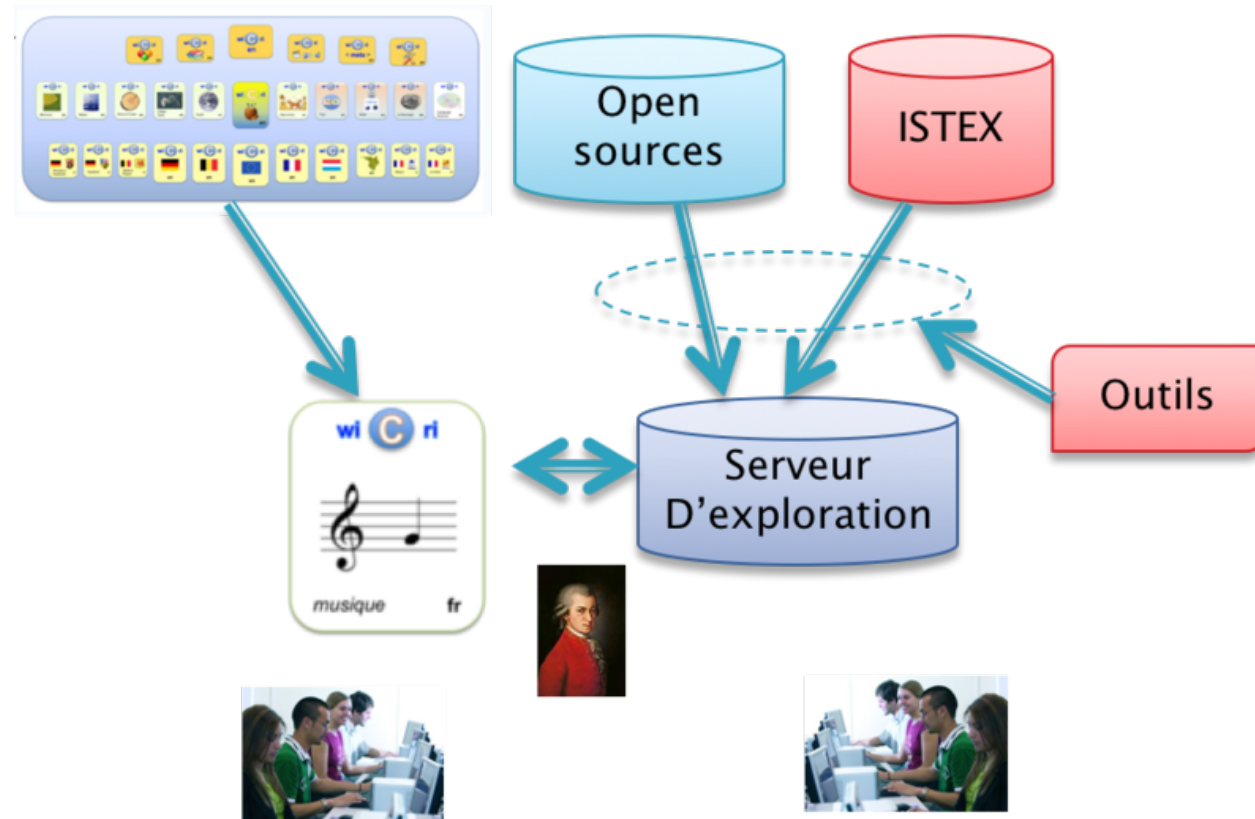
Pages paramètres

Base Xml

Santé : PubMed enrichi par ISTEKX / HAL

- ▶ PubMed
 - 30.000.000 articles (métadonnées)
 - Indexés par des spécialistes (chercheurs, médecins)
- ▶ ISTEKX
 - 22.000.000 articles en texte plein
 - Mais totalement hétérogène
- ▶ HAL
 - 2.000.000 articles 600.000 textes
 - Faible indexation, faible sélection
 - Très bon signalement institutionnel

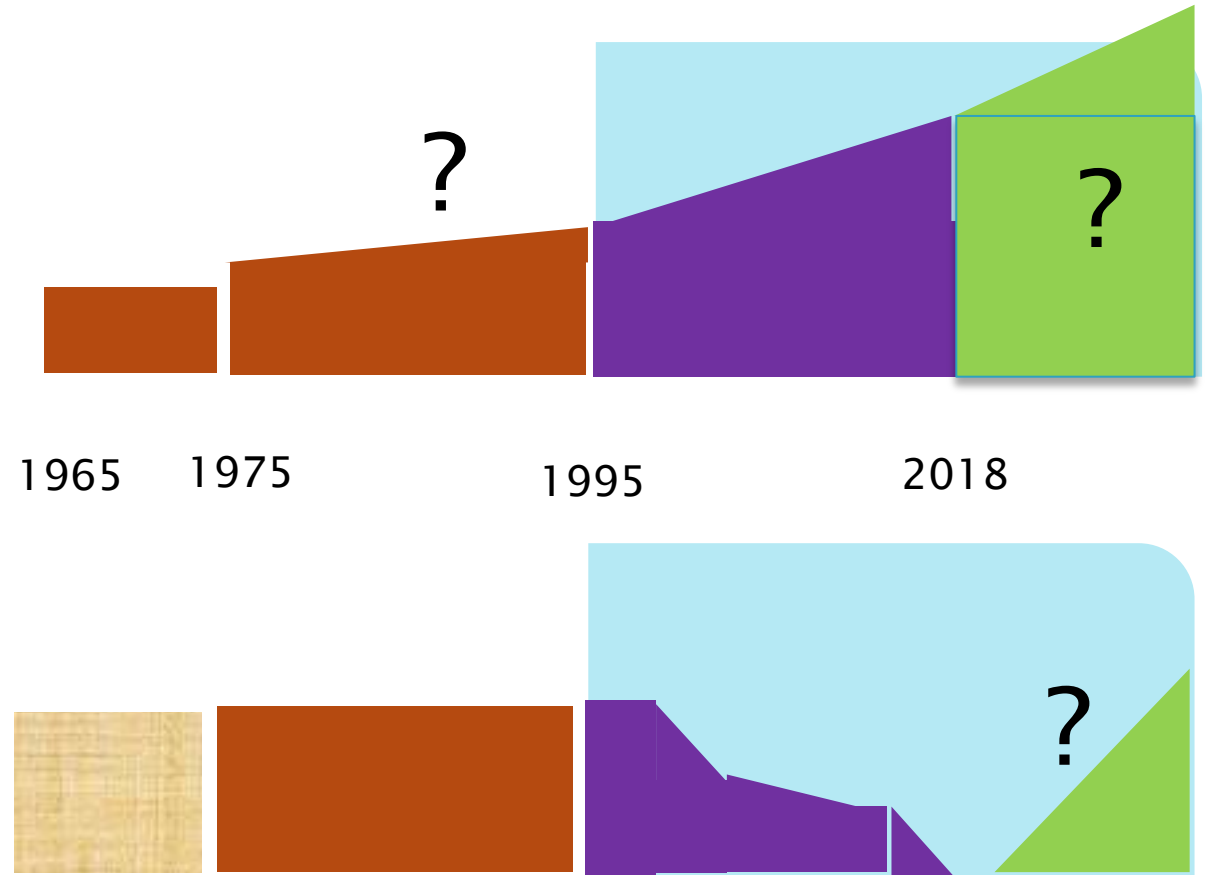
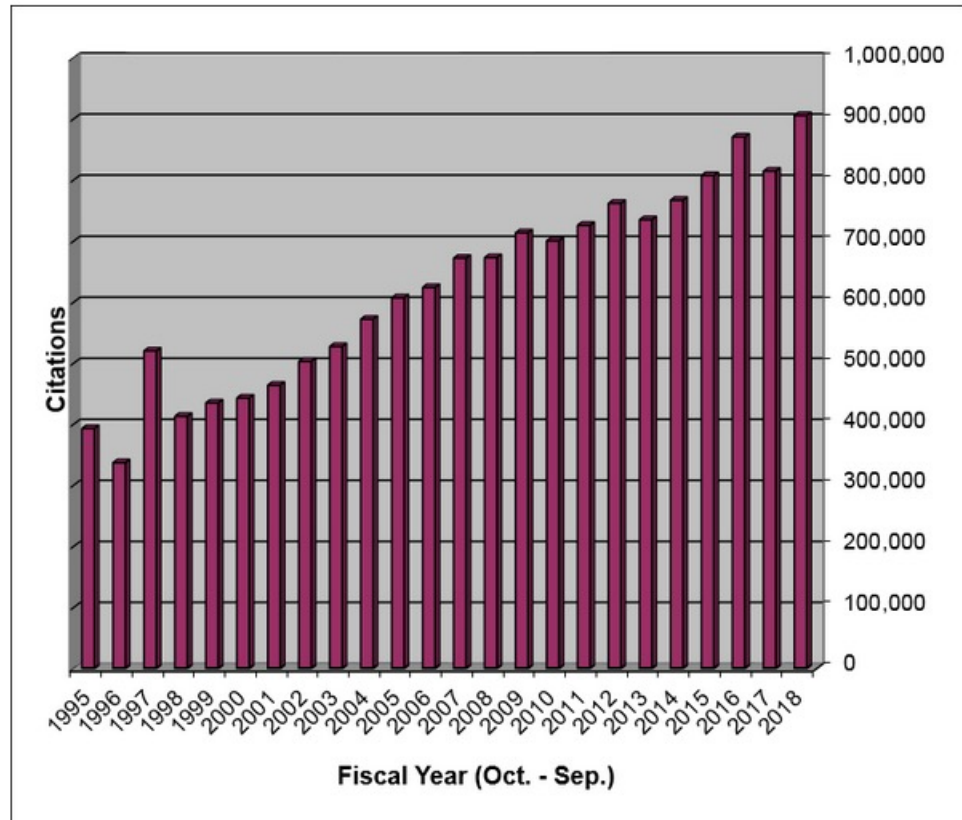
Ecrire et rechercher en mode collectif



Enjeux

Citations Added to MEDLINE® by Fiscal Year

The graph and chart below reflect the number of indexed¹ citations added to MEDLINE during each fiscal year since 1995.



Des voies pour l'avenir

▶ Comparaisons

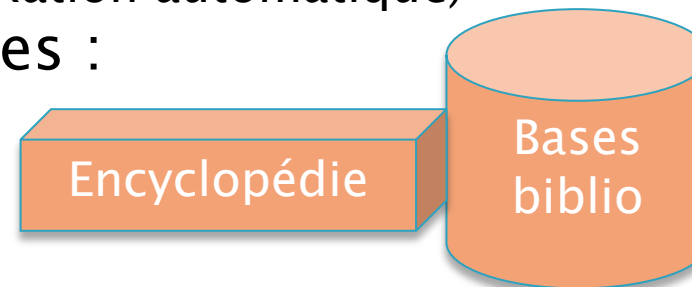
- USA : Les institutions
 - gèrent leurs bibliothèques et échangent des métadonnées
 - Conservent leur expertise en technologie numérique
 - Font confiance à l'expertise humaine (modération, indexation assistée)
- France : Les institutions
 - Délèguent vers des services centralisés (exemple HAL)
 - perdent leur maîtrise (et leur expertise)
 - Ne font pas confiance (contrôle a priori, indexation automatique)

▶ Quels moyens pour redémarrer les bases :

- les mêmes que l'encyclopédie
- pilote les sélections

▶ Quelle expertise

- édition numérique hypertexte de la connaissance
- En maîtrisant l'exploration de corpus



Conclusion

- ▶ Nouveau modèle de bibliothèque numérique
 - Des ouvrages hypertextes dans une bibliothèque hypertexte
 - Indépendance des praticiens par rapport aux informaticiens
 - Impossibilité de faire un cahier des charges pour explorer la connaissance
 - Exigence d'une culture numérique (équivalent du solfège ?)

- ▶ Merci pour votre attention et vos questions