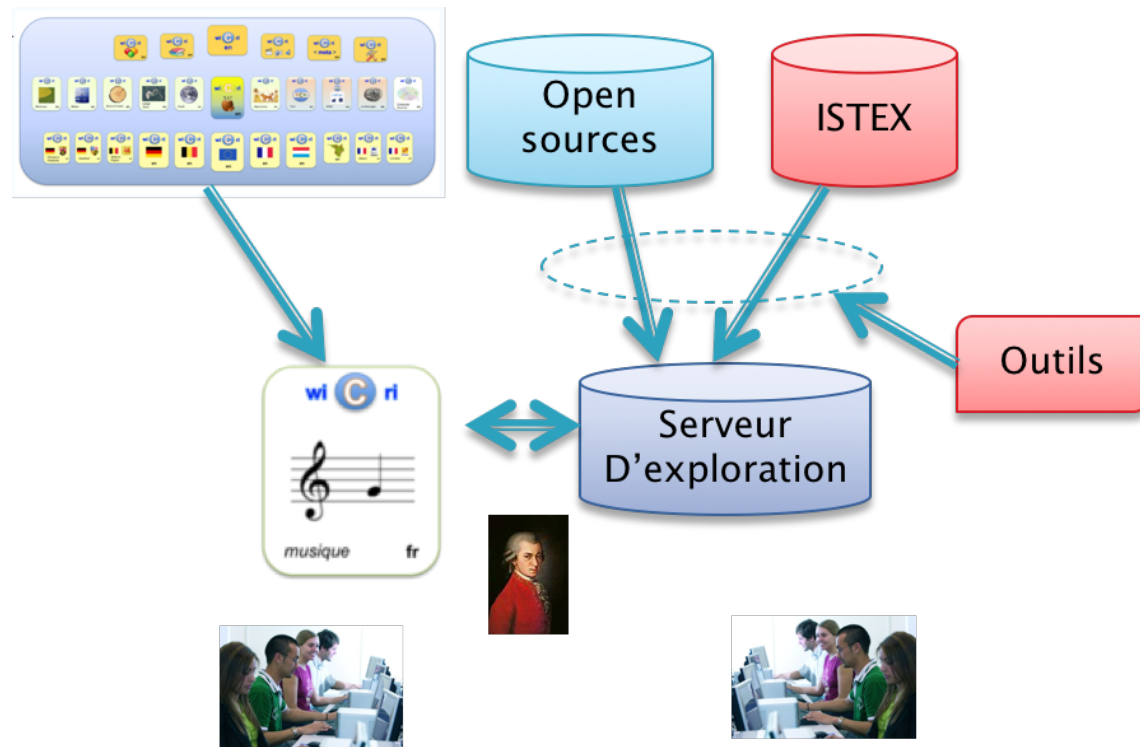


TP Université de Lorraine  
M2 Documentation numérique  
Exploration, curation de corpus

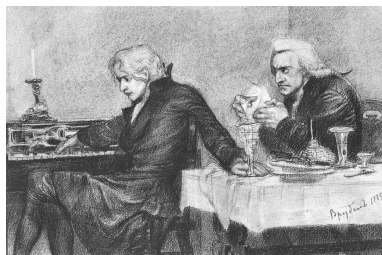
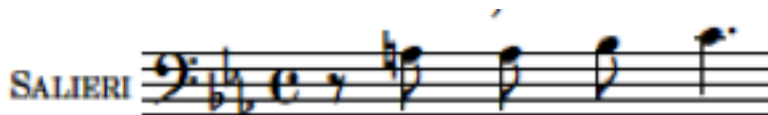
# Introduction : objectif LorExplor

- ▶ Construire de la connaissance structurée, sémantisée, en mode collaboratif et en explorant des corpus



# Construire de la connaissance

- ▶ Scientifique ou culturelle,
- ▶ Editorialisée en mode collaboratif.
- ▶ MediaWiki : moteur de Wikipédia
  - L'excellence du texte scientifique hypertexte
  - Une infrastructure de coopération (ex Wikipédia)



Ot - verg ya pa  
Ot - verg ya ra  
In ear - ly years  
Je re - pous - sai

Mozart  
15.000

CIDE



actes

OCR  
8.000

TEI  
1.000

# Structurer la connaissance

- ▶ Outils de base : indexation, polysémie, modèles.
  - Wikipédia est le noyau terminologique du Web sémantique
- ▶ Semantic MediaWiki et les liens sémantiques



sur un livret de [[A pour auteur de livret::Lorenzo da Ponte]].

Liste des opéras ayant un livret de Lorenzo da Ponte [modifier]

Sur ce wiki (par génération automatique)

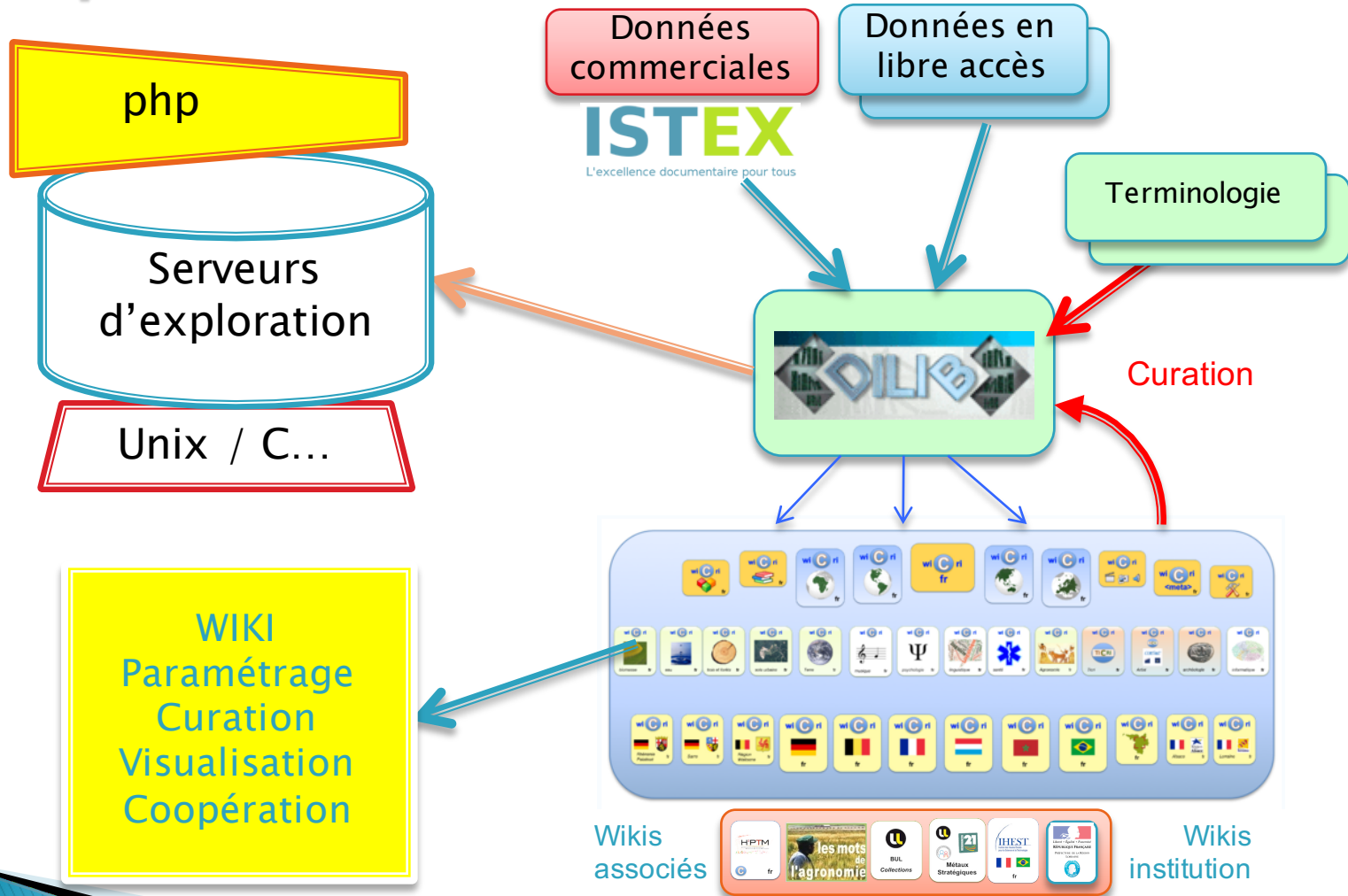
- Axur, re d'Ormus (Antonio Salieri)
- Così fan tutte (Wolfgang Amadeus Mozart)

```
{{#ask: [[a pour auteur de livret::{{PAGENAME}}]]  
| format=ul | ?A pour compositeur=compositeur :
```





# Atelier flexible pour explorer un corpus

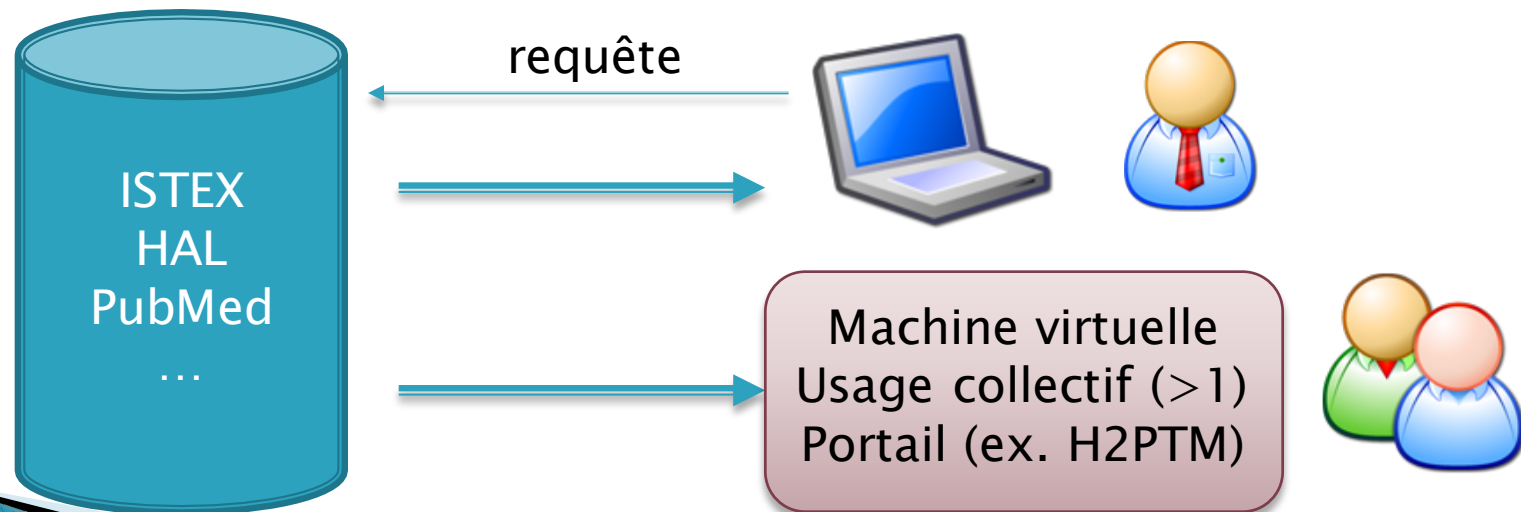


# Information retrieval – vs – Exploration des connaissances

- ▶ Un article de Martin Hatzinger sur la mort de Mozart ?
- ▶ Quelle est la plus ancienne référence à l'hypertexte ?
- ▶ Quelle est l'œuvre de Mozart la plus citée ?
- ▶ Quels sont les poissons les plus résistants aux eaux acides ?
- ▶ Quels sont les poissons encore sauvages qui peuvent être domestiqués ?
- ▶ Quels sont les laboratoires de Nancy qui travaillent avec la groupe Total ?
- ▶ Quelles sont les principales coopération entre la Lorraine et le Kazakstan ?
- ▶ Avec quels laboratoires l'Université de Lorraine peut s'allier pour coopérer avec le Kazakstan ?

# Exploration des connaissances

- ▶ Navigation classique :
  - Portails : WOS, Wikipédia, ENT, RefDoc, Istex (site)
- ▶ LorExplor :
  - Dans un corpus spécialisé venant d'une source
  - Dans un ensemble de corpus
    - Filtrage de données
    - Outils statistiques et visualisation
    - Curation des données



# Exploitation d'un corpus

## ▶ Exemples de volumétrie

- Observatoire Lorrain => filtrage sur la France
  - PubMed : France[Affiliation] : 536.000
  - ISTEK : (author.affiliations:\*France) : 513.000
- Un poisson (Lota lota) : 217 PubMed / 3.400 ISTEK
- Mozart : 442 PubMed / 14.000 ISTEK
- Hypertexte : 800 PubMed / 5.500 Pascal / 22.000 ISTEK
- Revue Movements Disorders : 20.000 -> 10.000

## ▶ 2 modes complémentaires

- Navigation dans un serveur d'exploration
  - Axes simples : ISSN, auteurs, mots du titre,
  - Axes élaborés : pays, région, ville, organismes
  - Cartes
- Utilisation d'une boîte à outils logicielle (ici DILIB)
  - Parser XML et outils de filtrage / sélection / reformattages
  - Système de recherche d'information en kit

# Exploration, exemple index AutAff

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

Department of Neurology, Juntendo University School of Medicine, Tokyo	<a href="#">004177</a>
Department of Neurology, Juntendo University School of Medicine, Urayasu Hospital, Tokyo, Japan	<a href="#">002388</a>
Department of Neurology, Juntendo University, School of Medicine, Tokyo, Bunkyo-ku, Japan	<a href="#">004111</a>
Department of Neurology, Jutendo University, School of Medicine, Tokyo, Japan	<a href="#">003517</a>
Department of Neurology, Research Institute for Diseases of Old Age, Juntendo University School of Medicine, Tokyo, Japan	<a href="#">000C30</a>
Department of Neurology, Research Institute for Diseases of Old Ages, Juntendo University School of Medicine, Tokyo, Japan	<a href="#">000393</a>

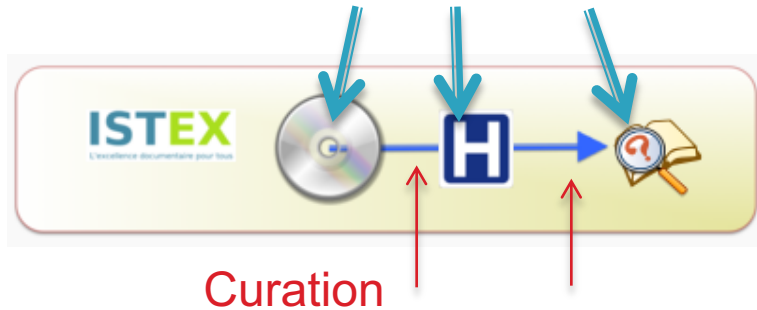
365 [Lees A](#)  
 325 [Lang A](#)  
 303 [Louis E](#)  
 279 [Poewe W](#)  
 251 [Bhatia K](#)  
 248 [Quinn N](#)  
 218 [Goetz C](#)  
 216 [Jankovic J](#)

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



# Serveur d'exploration -> portail

## Systeme d'information oriente exploration



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

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

Impress3dV1, Pas... Les plus visités Zimbra: Réceptio... Catégorie:Palais... Nouvel onglet PLOS ONE: "Fresh... VSST 2015 Dilib, module Exp... >>

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

**List of bibliographic references**

Number of relevant bibliographic references: 3.

Ident.	Authors (with country if any)	Title
Gilda Sedeh (Stats-Unis)	Akinrinola Bankole (Stats-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

# Filtrage du texte

Heuristiques :

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

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

Autres exemples : noms binomiaux, formules

En prévision : Outils linguistiques

# Curation des données

- ▶ Exemple : identifier les pays dans un contexte hétérogène



Server d'exploration sur la didactique - Wicri Wicri

Server d'exploration sur la did...

ticri.univ-lorraine.fr/wicri.fr/index.php/Server\_d\_explor

Google

Les plus visités

Catégorie:Palais ...

PLOS ONE: "Fres..."

Dilib, module Ex...

Getting Started

- Index alphabétique
- Index thématique
- Page au hasard
- Aide

rechercher

Lire

Rechercher

boîte à outils

- Pages liées
- Suivi des pages liées
- Téléverser un fichier
- Pages spéciales
- Version imprimable
- Adresse de cette version
- Chercher les propriétés

1	Pascal Francis		Le premier corpus est constitué de 4284 notices extraites de Pascal/Francis avec la requête « mc = didactique ». Une requête plus large (sans précision de zone) donne environ 8000 notices.
2	PubMed		Le corpus PubMed est extrait avec le critère « didactic » qui sélectionne 4013 notices.
3	PubMed Central		Le corpus PMC est extrait avec le critère « didactic » qui sélectionne 402 notices (la forme « didactic » en sélectionne environ 8000
4	Convergence NCBI		Ce flux rassemble les 4415 notices venant de PubMed et PubMed Central
5	HAL SHS		
Flux principal			Ce flux rassemble la totalité des 8643 notices.
Zoom	Auteurs français		Ce zoom propose une analyse plus fine autour des travaux réalisés avec affiliations françaises (1296 notices)
Zoom	Enseignement des langues		Ce zoom propose une analyse plus fine autour de l'enseignement des langues (1127 notices)

# 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, 3
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 cartes sont sur les wikis « portails thématiques »  
Wicri/Musique, Wicri/Santé, Wicri/Terre, Ticri/H2PTM



# 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		

```
<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>
```

Sur la machine  
D'exploration

# Création d'une plateforme paramétrable

- ▶ Plusieurs sources (ISTEX, PubMed, Pascal...)
- ▶ Options de curation
- ▶ Portail en ligne personnalisé



- 1 - Génération paramètres
- 3 - Chargement du (des) corpus
- 5 - Chargement paramètres
- 6 - génération du serveur
- 7 - envoi (FTP) su serveur sur machine Univ. Lorraine
- 10 - chargement règles curation
- 11 - Compilation DILIB
- 12 - Retour en 3, 4. 6....



9 - Ajout de données de curation

Wikis  
cibles



ou



2,4 - Adaptation paramètres  
8 - IHM (cartes..)

# Création serveur simple PubMed (ou ISTEK – métadonnées MODS)

## ► Configuration

- Unix + Serveur WWW (php), exemple MAMP
- DILIB (boite à outils + générateur)

## ► Commandes

```
NlmPubMedGetCorpusSize -q query
```

```
NlmPubMedFlashCorpus -q query
```

```
NlmPubMedExplorCorpus -q query -s size -d rootDir
```

## ► Exemple

```
NlmPubMedExplorCorpus -q mozart -s 200 -d testMoz
```



```
HfdCat testMoz/data/Main/Exploration/AffOrg.i.hfd \  
| grep Univ | wc
```

*Calcul du nombre d'universités dans l'index organisme*

# Passons à la pratique

- ▶ Par groupe :
  - Détermination d'un critère permettant d'extraire un corpus spécialisé sur ISTEEX, HAL, PubMed et PubMed Central
- ▶ Pour ISTEEX
  - IstexGetCorpus options -s 1
    - Estimation du volume d'une requête
  - IstexFlashCorpus
    - Vue sommaire du contenu d'un corpus
  - IstexExplorCorpus
    - Vue plus approfondie
- ▶ Pour PubMed, voir NLM Entrez