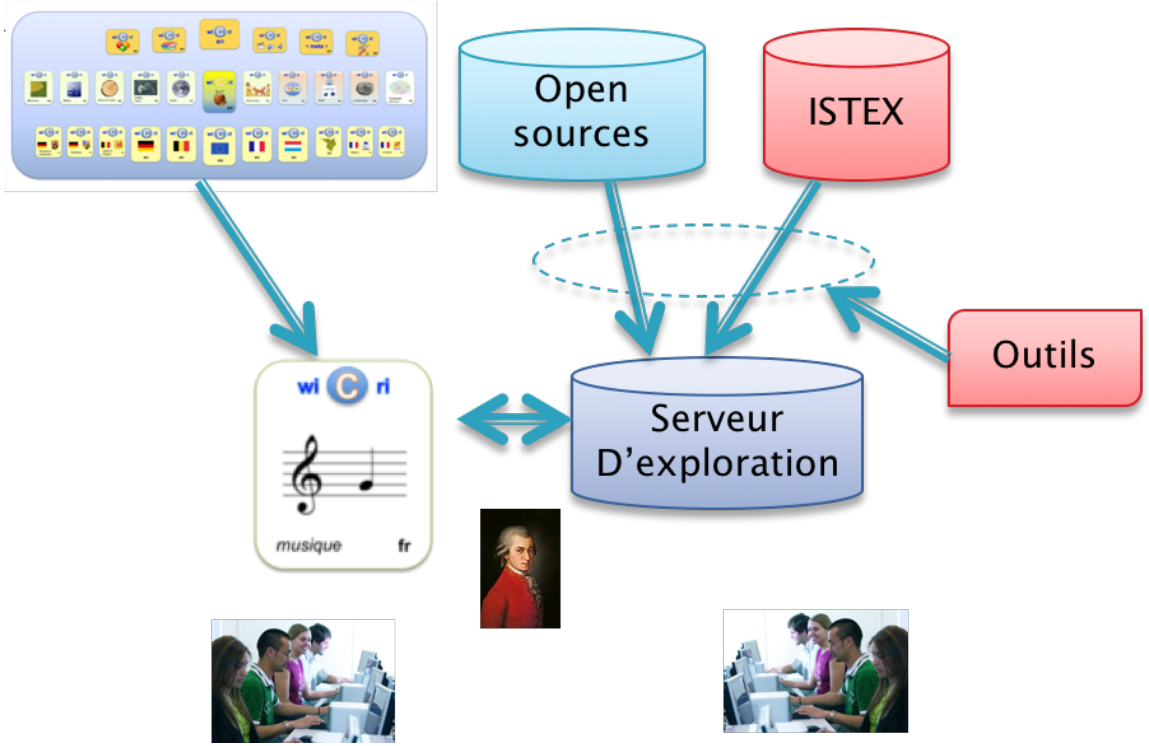
TP Université de Lorraine M2 Documentation numérique

Exploration, curation de corpus

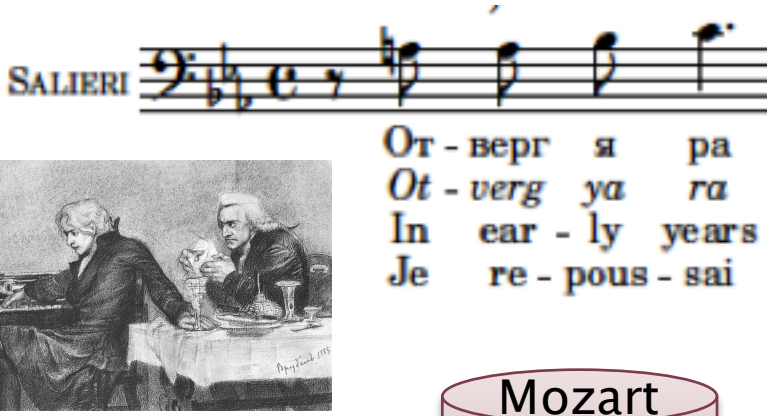
Introduction : objectif LorExplor

- 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)



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]].
```

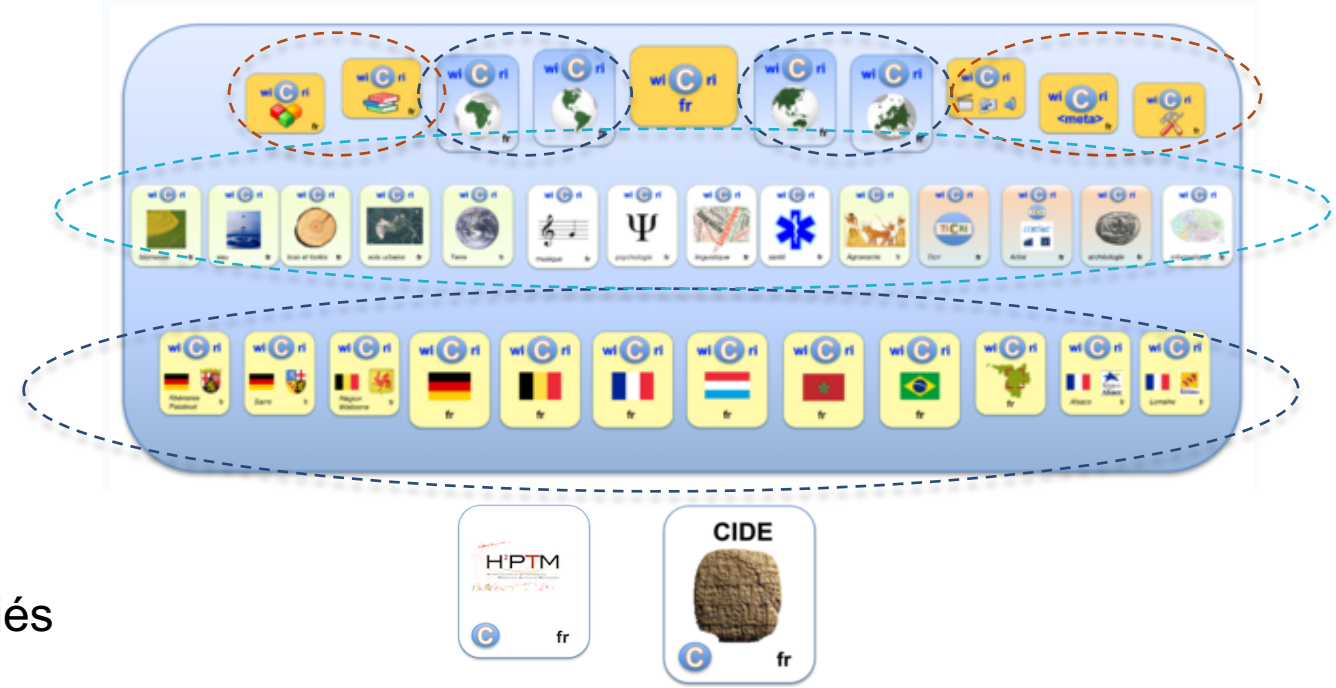

```
{{#ask:[[a pour auteur de livret::{{PAGENAME}}]]
| format=ul| ?A pour compositeur=compositeur :
```

Le Réseau Wicri pour gérer l'interdisciplinarité

Wikis de service

Wikis thématiques

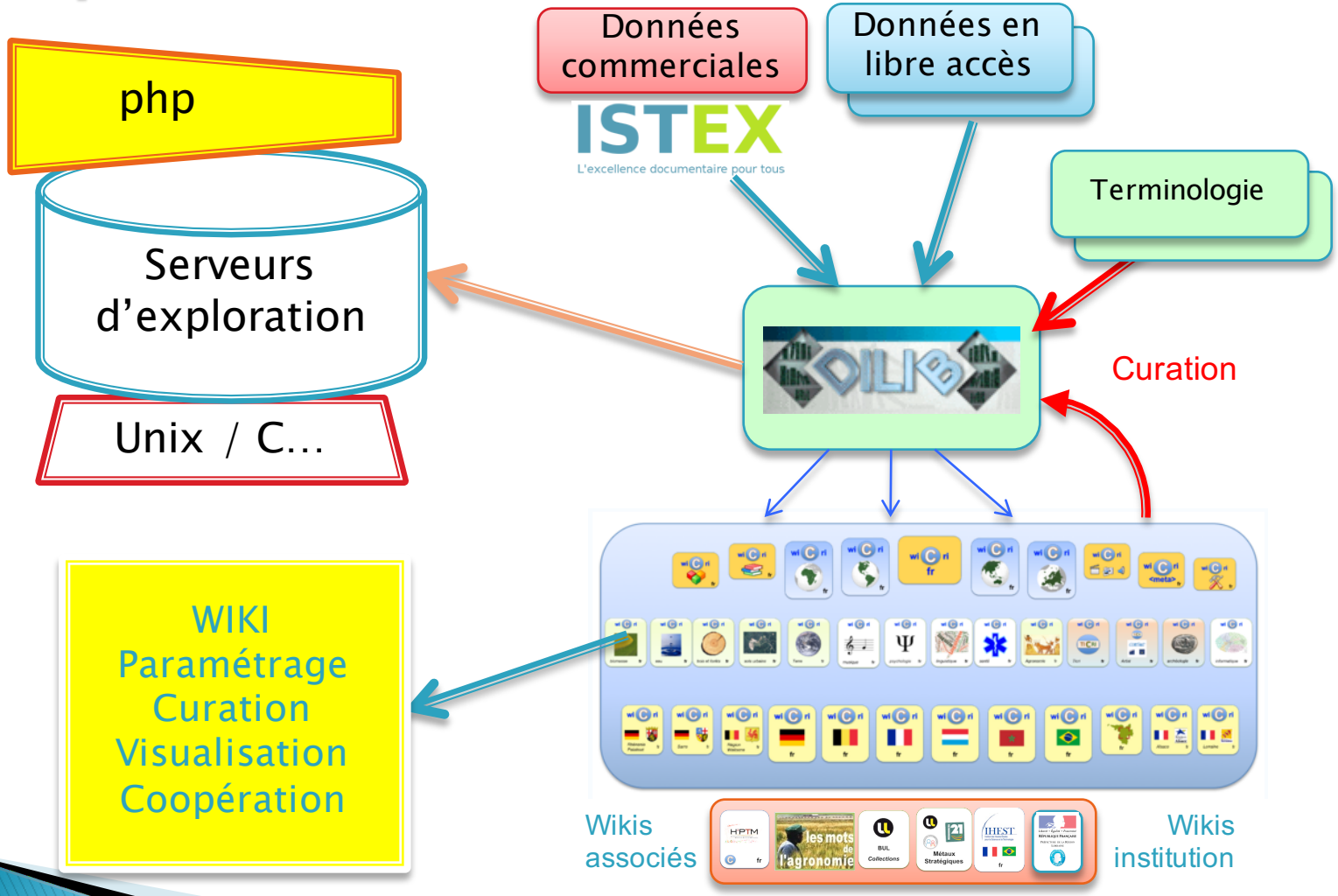
Wikis régionaux



Wikis associés

Un espace (wiki) par thématique ou région
On peut travailler dans plusieurs thématiques

Atelier flexible pour explorer un corpus

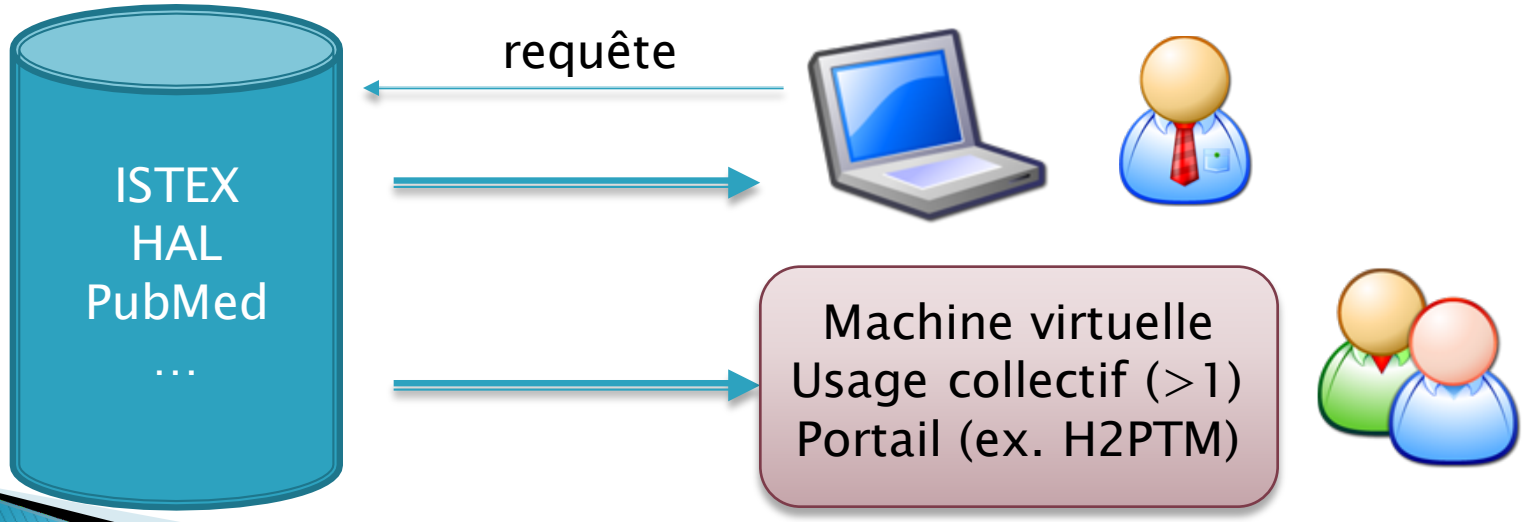


Information retrieval – vs – Exploration des connaissances

- 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
    - ISTEX : (author.affiliations:*France) : 513.000
  - Un poisson (Lota lota) : 217 PubMed / 3.400 ISTEX
  - Mozart : 442 PubMed / 14.000 ISTEX
  - Hypertexte : 800 PubMed/ 5.500 Pascal / 22.000 ISTEX
  - 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

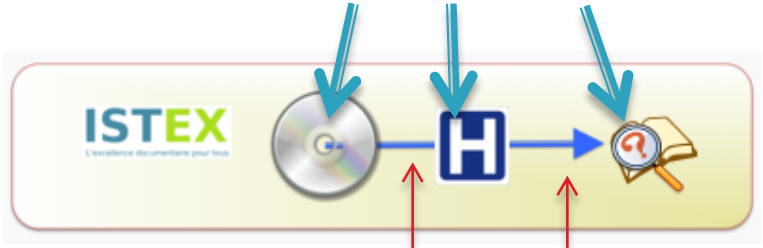
	Department of Neurology, Juntendo University School of Medicine, Tokyo	004177
	Department of Neurology, Juntendo University School of Medicine, Urayasu Hospital, Tokyo, Japan	002388
	Department of Neurology, Juntendo University, School of Medicine, Tokyo, Bunkyo-ku, Japan	004111
	Department of Neurology, 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

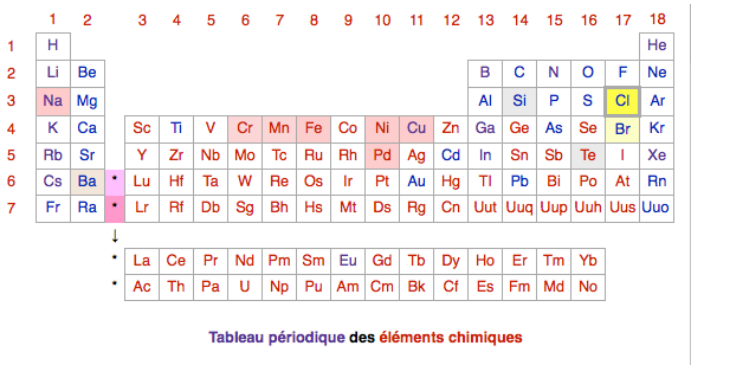
365 Lees A
325 Lang A
303 Louis E
279 Poewe W
251 Bhatia K
248 Quinn N
218 Goetz C
216 Jankovic J

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

Serveur d'exploration -> portail

Système d'information orienté exploration





Master UL 2016

Filtrage du texte

Heuristiques :

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

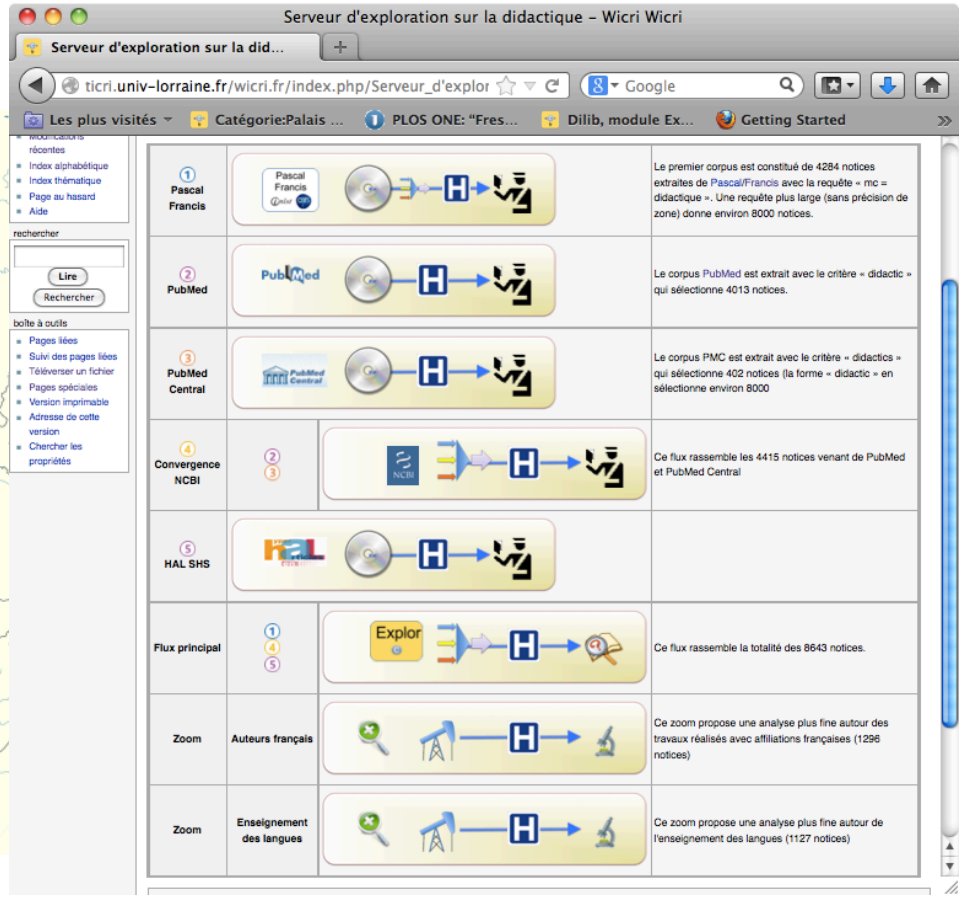
```
HfdCat Data/Main/Exploration/biblio.hfd                \
  | SxmlFindText -r "[K][Vv]*[ \.]*[0-9][0-9]* »  \
  | SxmlSelect -p @5 -p @1 | sort | IndexBuildRec
```

Autres exemples : noms binomiaux, formules

En prévision : Outils linguistiques

Curation des données

- Exemple : identifier les pays dans un contexte hétérogène



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, 1
                      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 W
                      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">Matthia
 <namePart type="family">Keller
 <role>
  <roleTerm type="text">author<
 </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	101		
Bonn	W-5300	531		

```
<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

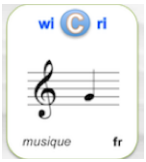
- Plusieurs sources (ISTEX, PubMed, Pascal…)
- Options de curation
- Portail en ligne personnalisé

ISTEX
HAL
PubMed

9 – Ajout de données de curation

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….

Wikis cibles



ou

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

Création serveur simple PubMed (ou ISTEX – 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 ISTEX, HAL, PubMed et PubMed Central
- Pour ISTEX
  - 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