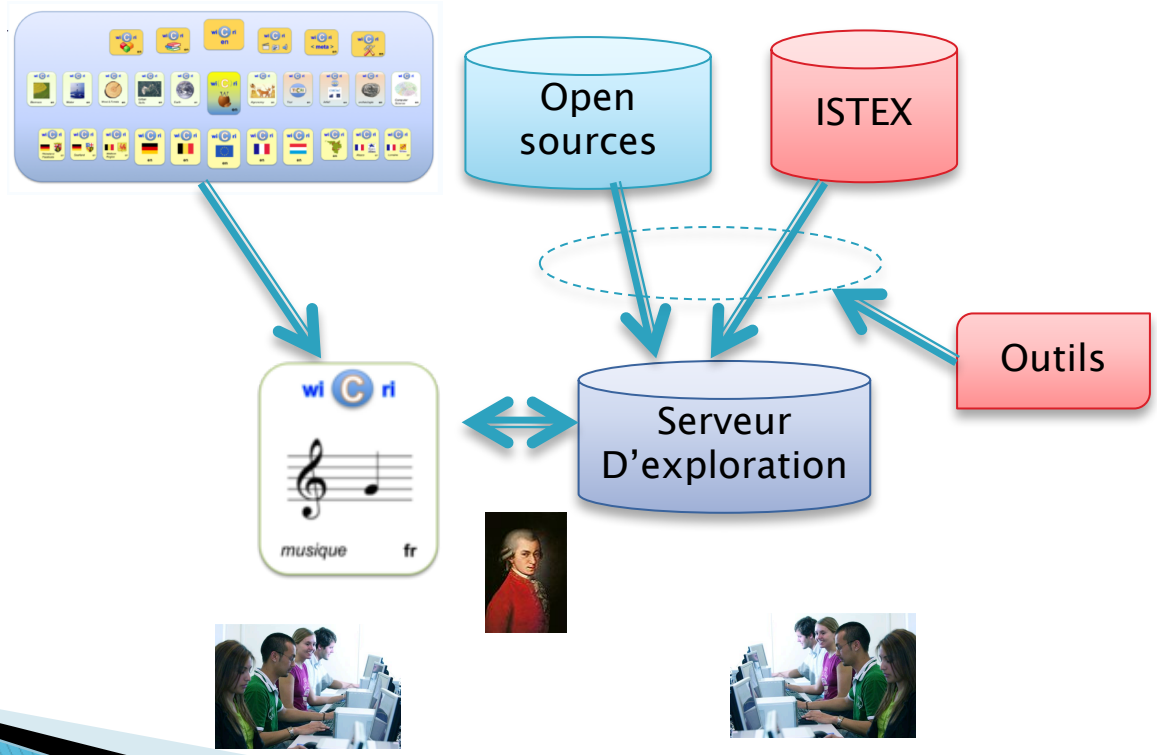
L'excellence documentaire pour tous ? chiche !

Wicri / LorExplor
CARIST – Atelier TDM, Mars 2017
Jacques Ducloy

Usages ciblés par LorExplor

- Co-construction de portail scientifique ou culturel
- Recherches exploratoires avec contraintes de temps
  - Biblio thèse, réponse à appel à projets…
- Dans une perspective de coopération mondialisée

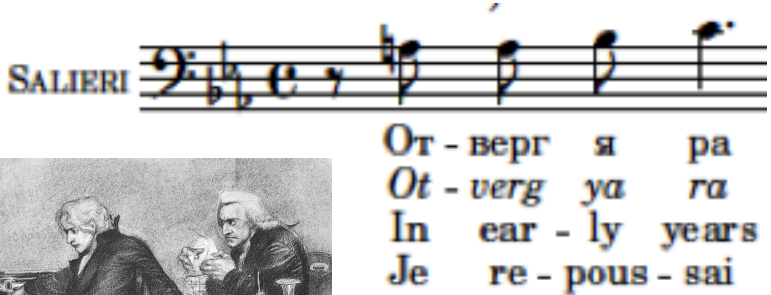


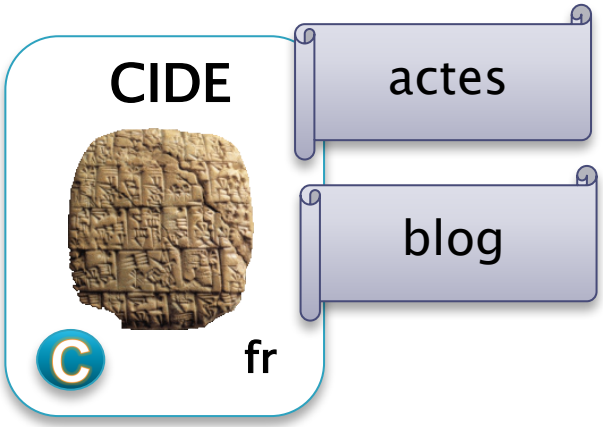
Recherche d'information – vs – Exploration des connaissances

- Un article de Tim Berners-Lee sur le web sémantique ?
- Quelle est la plus ancienne référence à l'hypertexte ?
- Quelle est l'œuvre de Mozart la plus citée ?
- Concernant le Kazakhstan :
  - Quelles sont les principales coopérations entre la Lorraine et le Kazakhstan ?
  - Avec quels laboratoires internationaux l'Université de Lorraine peut-elle s'allier pour coopérer avec ce pays ?
- Quels sont les poissons encore sauvages qui peuvent être domestiqués ?

Objectif des recherches sur corpus Construire de la connaissance

- Résultat d'une démarche
  - ou phase intermédiaire
- Scientifique ou culturelle,
- Editorialisée, en mode collaboratif et hypertexte
- Conception incrémentale (modèles)
- MediaWiki : moteur de Wikipédia
  - L'excellence du texte scientifique hypertexte
  - Une infrastructure de coopération (ex Wikipédia)



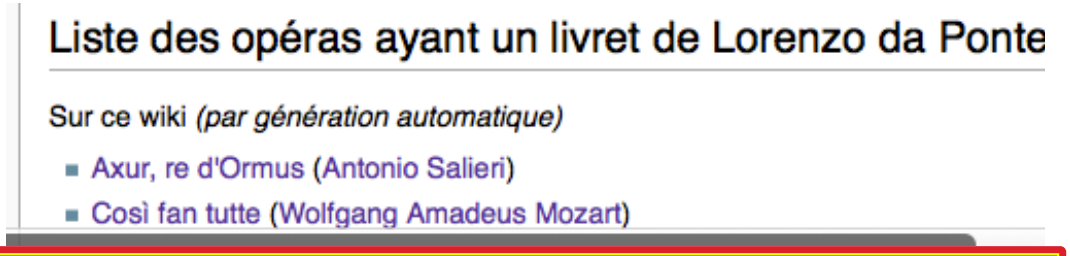


Structurer la connaissance et les règles de curation…

A pour auteur de livret

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

Semantic MediaWiki :
L'excellence dans la construction sémantique
Dans une approche RDF




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

Lire, écrire, rédiger, explorer dans l'interdisciplinarité

Wikis communs

de service

thématiques

régionaux

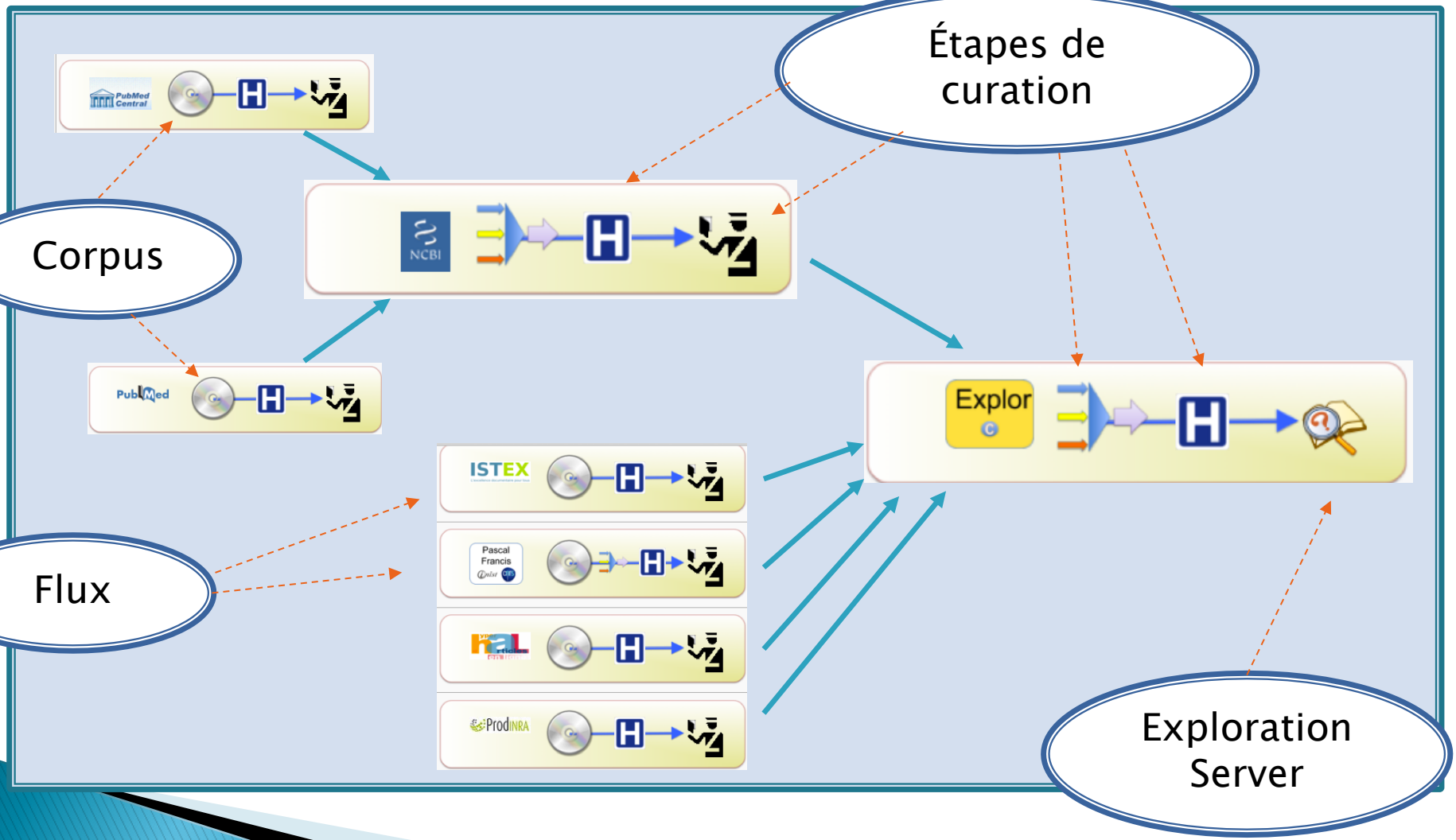
Wikis institutionnels plus ou moins associés

Sur les wikis communs, on peut travailler dans plusieurs thématiques

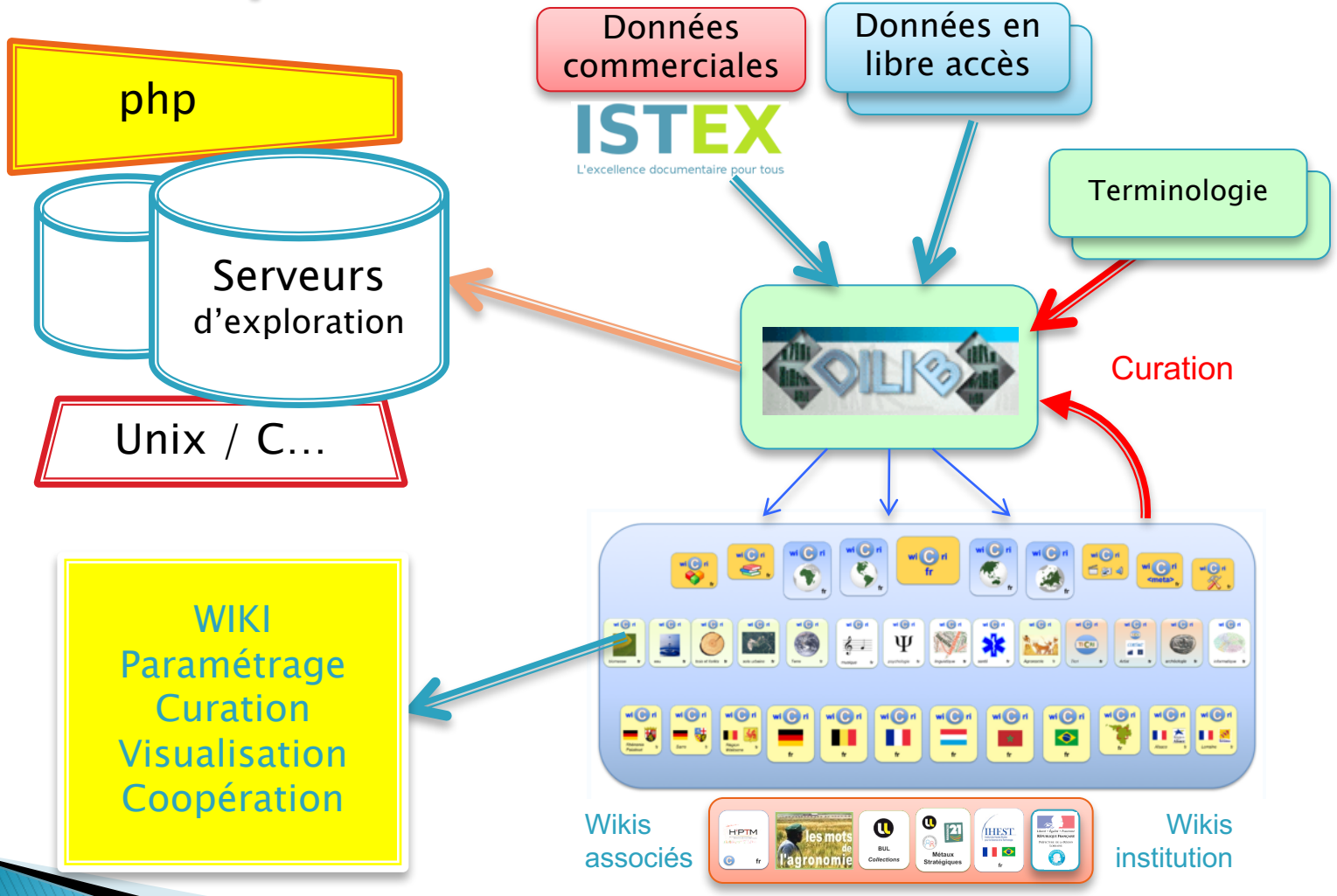
Le réseau Wicri : mutualisation des applications et thématiques

- Wicri : des centaines d'expérimentations
  - dans 14 thématiques,
  - sous 16 points de vue géographiques,
  - 10.000 pages de contenu
  - 80.000 pages de métadonnées
  - 300.000 interventions, 30.000.000 visites…
- Exemples d'interactions sur le wiki CIDE :
  - Les actes des conférences en texte intégral
    - => mise en perspective de cette communauté par rapport à l'activité internationale
  - Les comités, auteurs en relations sémantiques
    - => règles de curation pour les serveurs d'exploration

Plateformes de curation et d'exploration
ISTEX, Pascal/Francis, Hal, PubMed, PubMed Central



Un atelier flexible pour explorer des corpus



Bilan et exemples de sujets

- Plus de 100 expérimentations
  - dont Master Science information UL
    - Visibilité de la ville du Havre, (5303 au total/ dont 3003 ISTEX)
    - Le cobalt au Maghreb, (4062 / 3027)
    - Un poisson : le scalaire, (1329 / 906)
    - Un arbre fruitier : l'oranger (8819 / 2923)
    - Le libre accès en Belgique, (3696 / 2961)
  - et Paris 8
    - La Maladie de Parkinson en France, (11473 / 3727)
    - La Paléopathologie (5459 / 2469)
    - Le nickel au Maghreb (3337/ 2500)
    - Université de Trèves (6789 / 2846)
    - Un poisson : l'esturgeon (4057/2398)
    - La thérapie familiale en francophonie (3463 /2817)
    - Le renard en Grande Région (3133 / 2219)
    - Le chêne en Belgique (3267/ 2739)
    - Système d'information stratégique et agriculture (3011 / 2042)

Corpus / méfiance / curation

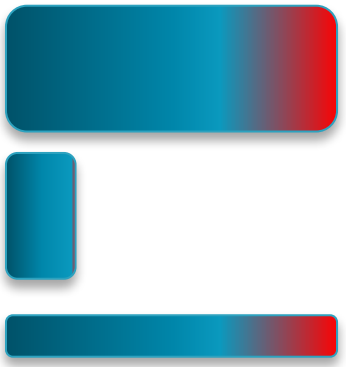
- Exemple : la méthode Scrum
  - Apparemment : 9.000 documents
  - En fait 90% de bruit du à l'OCR (sérum -> scrum)

- Exemple : le libre accès en Belgique
  - Apparemment : 4000 documents
  - En fait : 100 à 200 sont pertinents
    - *Title:The EADGENE Microarray Data Analysis Workshop (Open Access publication)*

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.



Conclusion : l'excellence documentaire pour… un retraité

- Le démonstrateur LorExplor
  - peu de moyens et peu de soutiens institutionnels
- Il démontre l'excellence
  - De l'API ISTEX (réalisée par l'INIST)
  - des solutions techniques Wicri/LorExplor
    - Synergie réseau SMW, Unix/xml, Corpus
  - d'une démanche basée sur
    - la confiance (modération a posteriori),
    - L'utilisateur omniprésent,
    - L'expertise mutualisée sans périmètre institutionnel
- Pour la suite : les machines virtuelles…
  - Volumes plus conséquents
  - Dernières versions : MediaWiki SMW
    - Outils de visualisation intégrés aux wikis…
- Aller le plus loin possible dans la réponse aux besoins

L'excellence documentaire pour quelques-uns

- De 2 à 5 personnes, multiples retombées
  - Peu de besoins en développement
    - On garde Wicri/LorExplor
  - Centre de formation
    - Faire comprendre les mécanismes avant de passer à l'utilisation d'outils clé en main.
  - Soutien logistique MediaWiki, Semantic MediaWiki

    Basé sur l'accompagnement
    - Edition numérique
    - Observatoires de la recherche, encyclopédies, terminologie
    - Base de donnée bibliographique spécialisée…
  - Avec un soutien logistique applicatif
    - Solutions opérationnelles (ElasticSearch, Mediawiki, Semantic MediaWiki)
    - Consolidation de logiciels issus de le recherche

    - Veille scientifique ou technologique
    - Infrastructure pour de la recherche sur les usages

L'excellence documentaire pour tous

- Dispositif d'accompagnement en réseau
  - Une cellule par université…
- Plan de formation à l'excellence documentaire pour tous

- Une stratégie « à la Wikipédia »
  - Passage à l'échelle
    - 40 wikis -> 1000 wikis
    - Consolidation de millions de règles
  - À la portée d'une université qui a une volonté politique
    - ou de l'ABES ou de l'INIST..

Après cette conclusion…

- Visite guidée..

Dilib, une boîte à outils Sxml

- SXML :
  - XML lite (mais JSON+)
  - Compatible avec les outils Unix
    - Un document = Une ligne Unix
- 1990 : Ilib
  - Pour : traiter l'ISO 2709 (MARC, Pascal…)
  - SGML dans une philosophie XML
- 2010 : Dilib 0.5…
  - Besoin : traiter du corpus
    - volumineux,
    - textuel multi-dtd
  - Unix + MediaWiki
- Un LEGO pour les corpus


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

Serveur d'exploration
Le plus simple : parcourir les index

Génération de modèles wiki

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

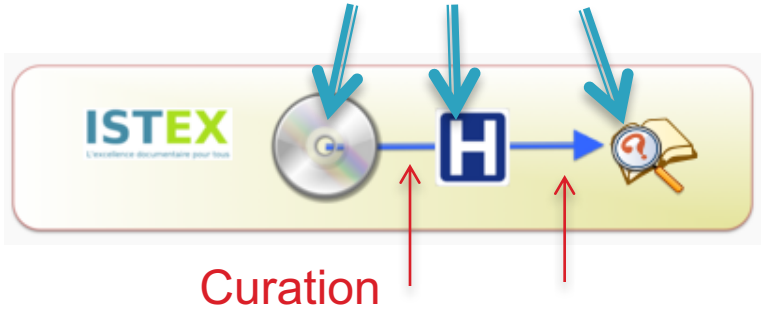
1. data (10)
2. analysis (7)
3. software (6)
4. microbial (6)
5. marine (5)
6. genome (5)
7. distributed (5)
8. genomic (4)
9. control (4)
10. web (3)

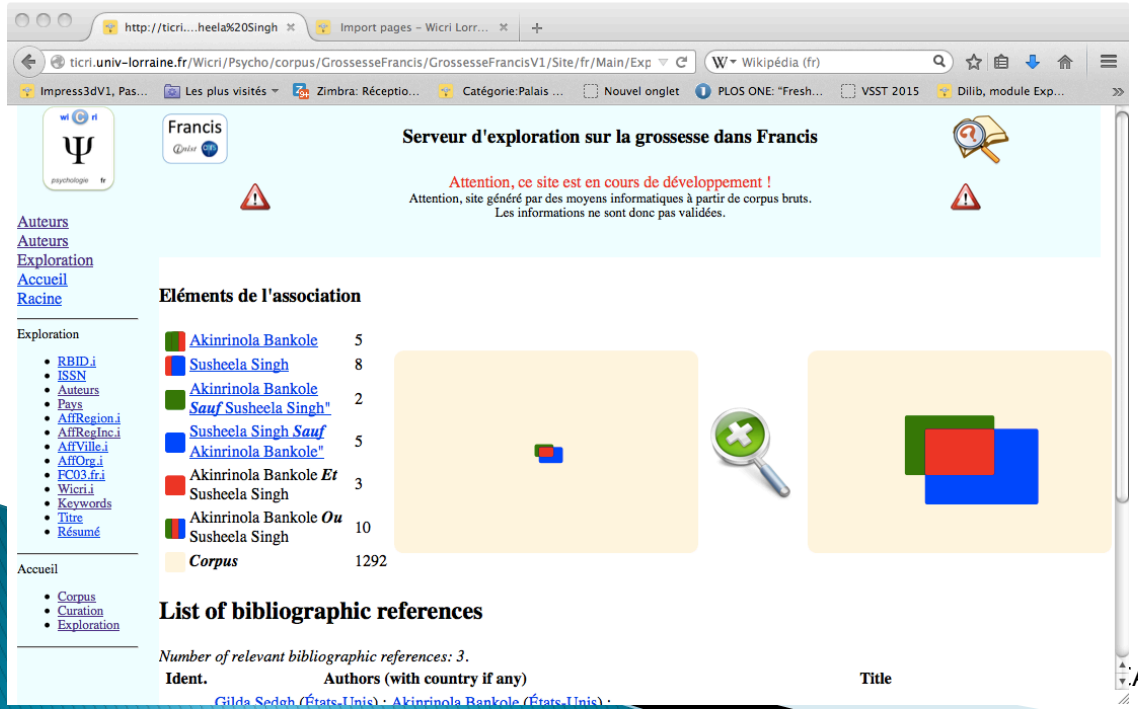
ISSN/revue

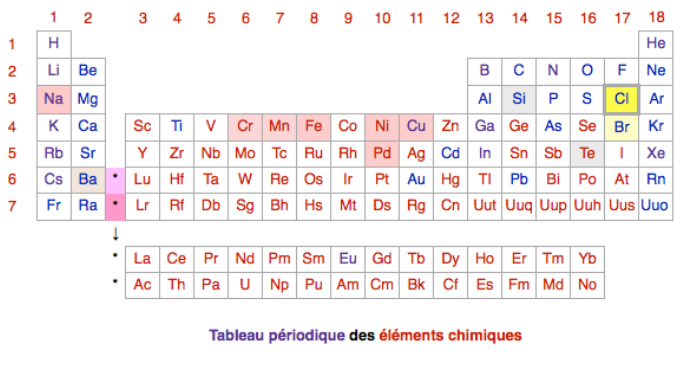
1. SPIE proceedings series (6)
2. 1932-6203 (5)
3. Lecture Notes in Computer Science (4)
4. Eos Trans. AGU (3)
5. 2324-9250 (3)
6. 1091-6490 (3)
7. 0096-3941 (3)
8. 0027-8424 (3)
9. 2047-217X (2)
10. 1545-7885 (2)

Serveur d'exploration

Système d'information orienté exploration







Filtrage du texte

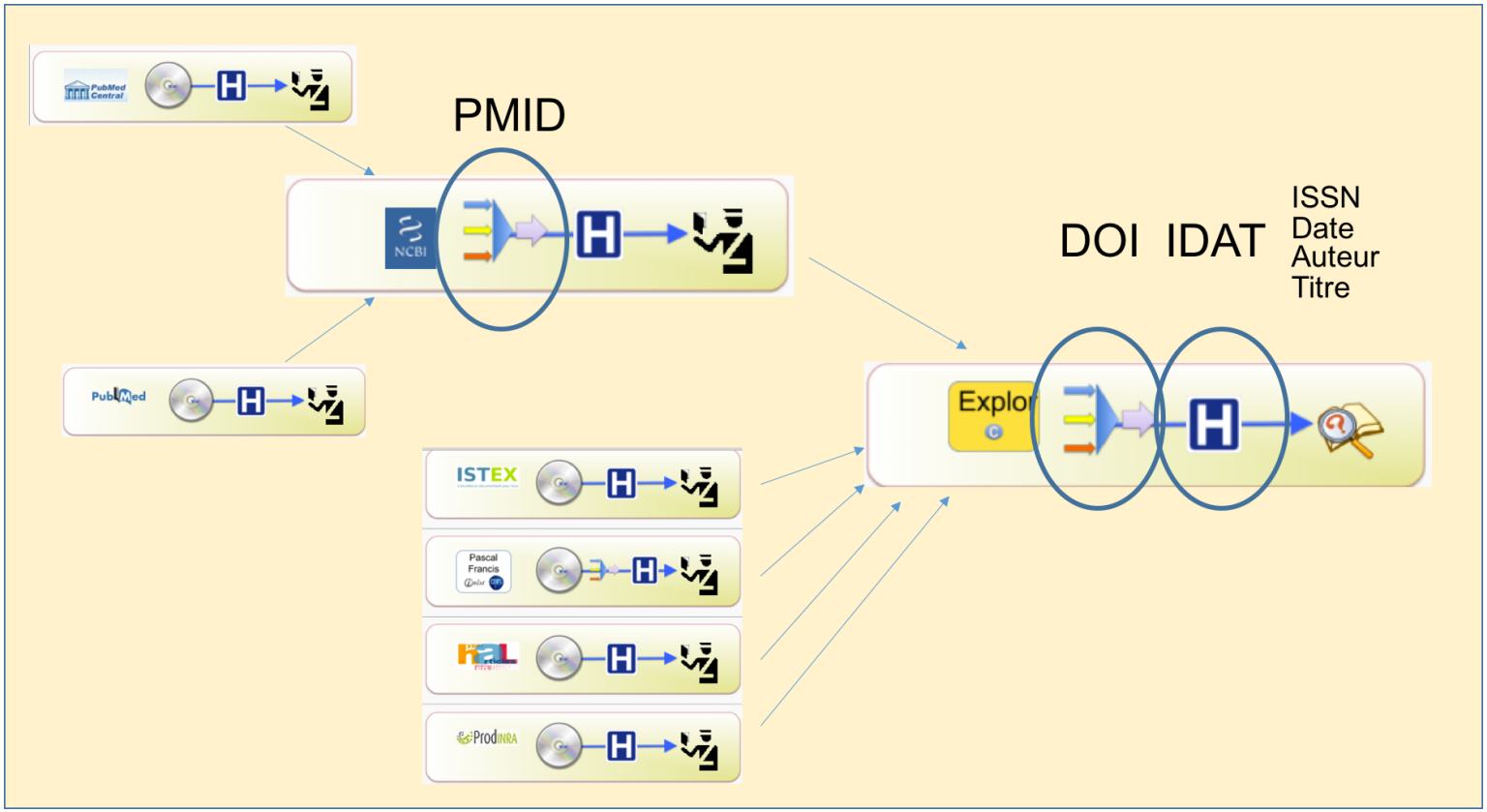
Heuristiques :

- 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
```

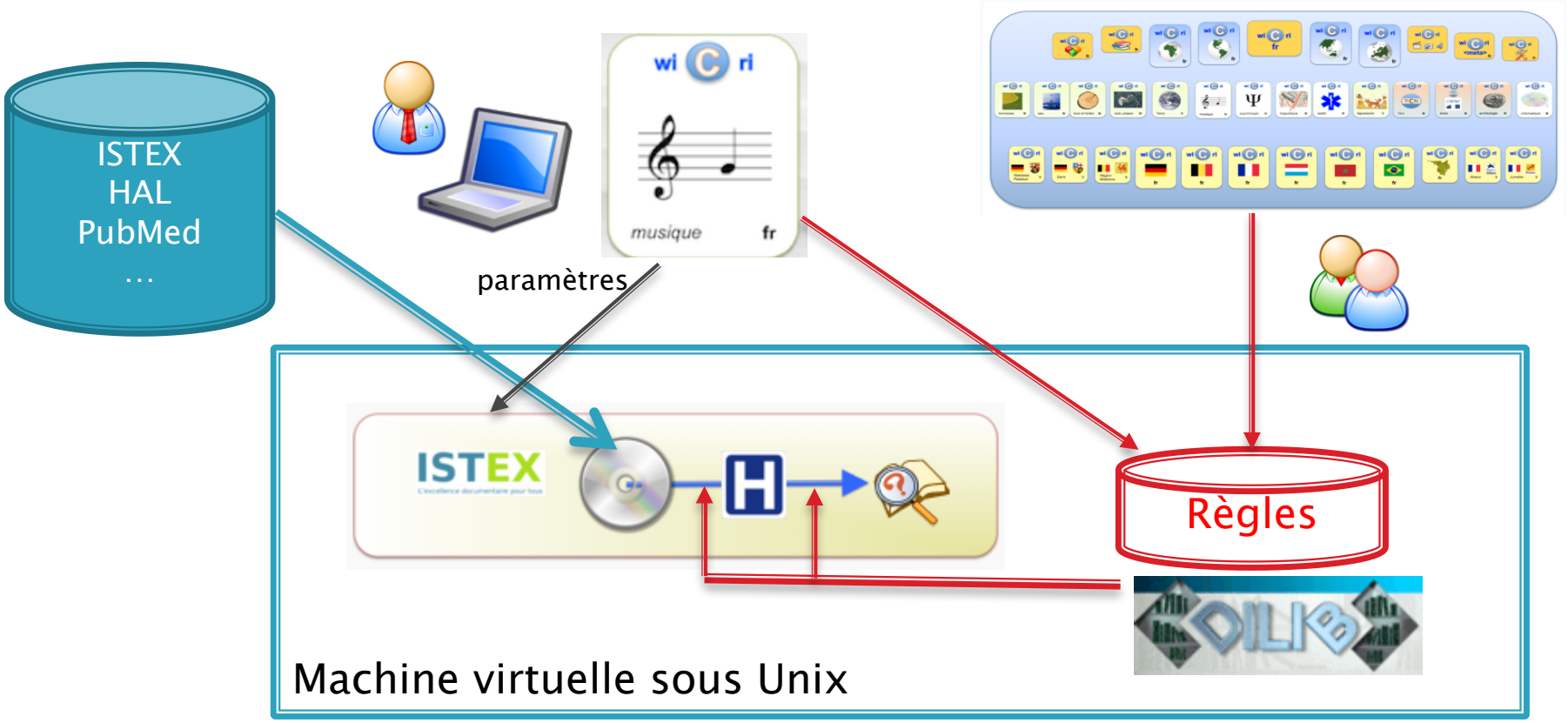
- Recherche de noms binomiaux
  - Recherche de chorégraphes (nom-prénom)
- Identification du cœur d'un corpus
  - (termes pondérés)

Enrichissement : dédoublonnage ISTEX / Pascal / Hal / MEDLINE…

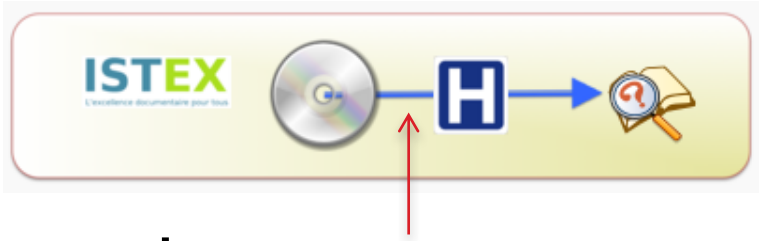


Génération de plateformes de curation et d'exploration

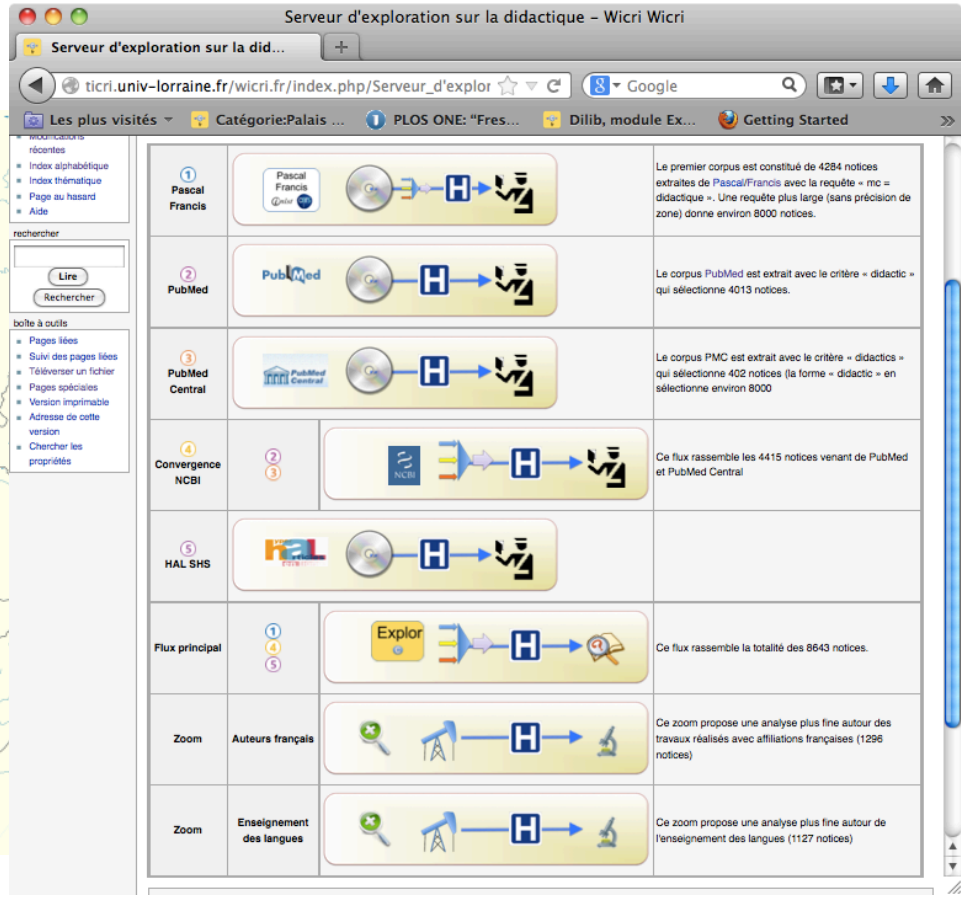
- Processus itératif mutualisé



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

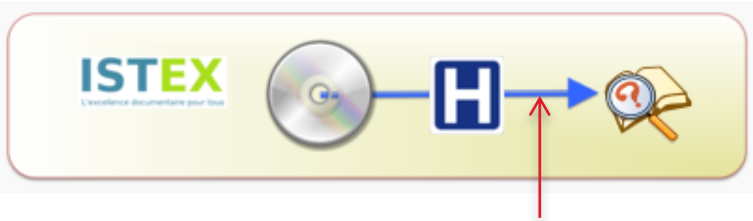
Adresses postales
(Springer, PubMed)

```
<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 types de carte sont visibles sur tous types de serveurs

Curation régions /codes postaux

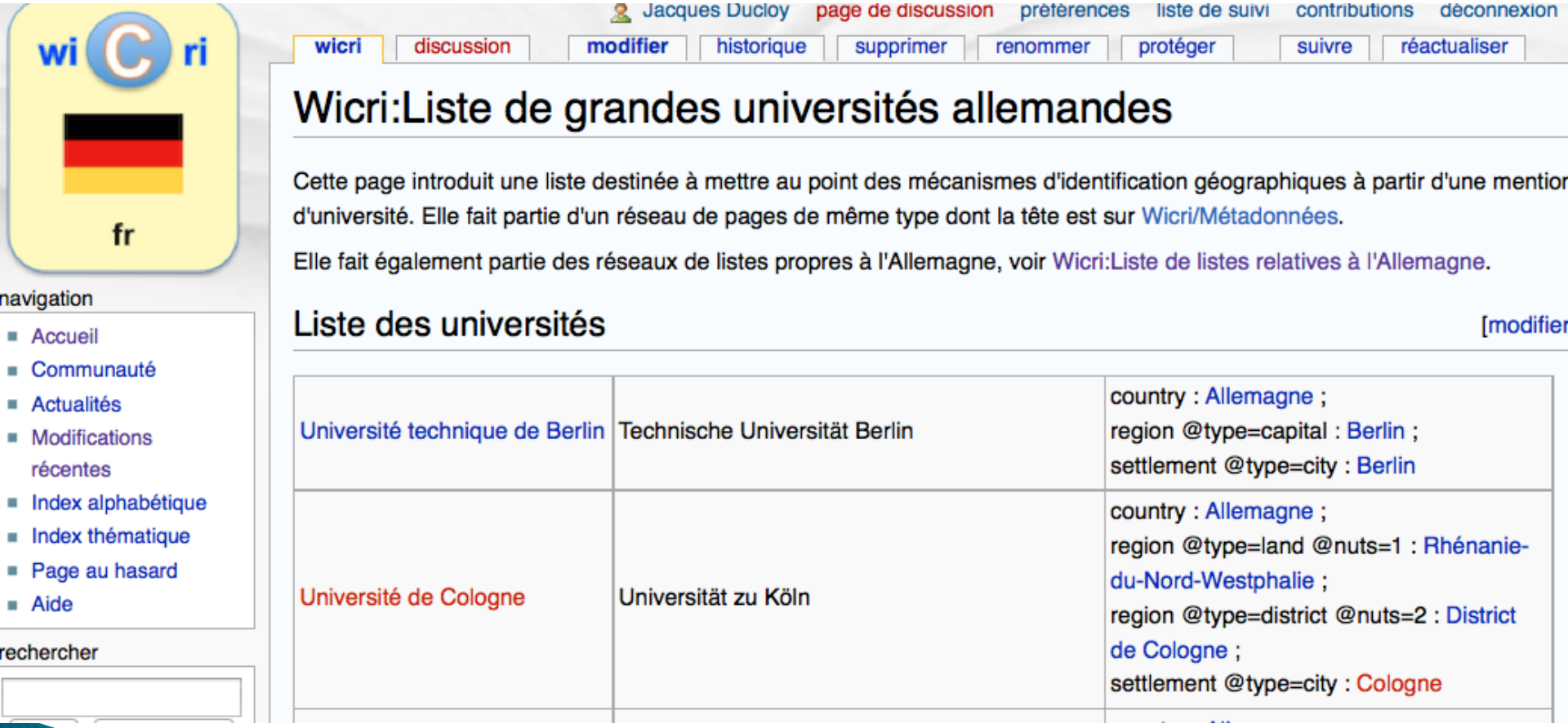
Sur Wicri/Allemagne

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

Curation des universités



wi C ri

fr

Jacques Ducloy page de discussion préférences liste de suivi contributions déconnexion

wicri | discussion | modifier | historique | supprimer | renommer | protéger | suivre | réactualiser

Wicri:Liste de grandes universités allemandes

Cette page introduit une liste destinée à mettre au point des mécanismes d'identification géographiques à partir d'une mention d'université. Elle fait partie d'un réseau de pages de même type dont la tête est sur Wicri/Métadonnées.

Elle fait également partie des réseaux de listes propres à l'Allemagne, voir Wicri:Liste de listes relatives à l'Allemagne.

Liste des universités [modifier]

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

navigation

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

rechercher

Merci

- Pour l'excellence de votre attention !
- Pour l'excellence de vos questions…