



ISTEX
L'excellence documentaire pour tous

Wicri / LorExplor Une cyberinfrastructure analyser, diffuser la science

Séminaire ISTEX 2017, Nancy
Jacques Ducloy, Thierry Daunois, Ali Tebbakh

Plan

- ▶ Rappel : Objectifs Wicri et LorExplor
- ▶ Résultats obtenus
- ▶ Quelques observations
- ▶ Conclusion
- ▶ Visite éventuelle

- ▶ Diapositives complétées sur :

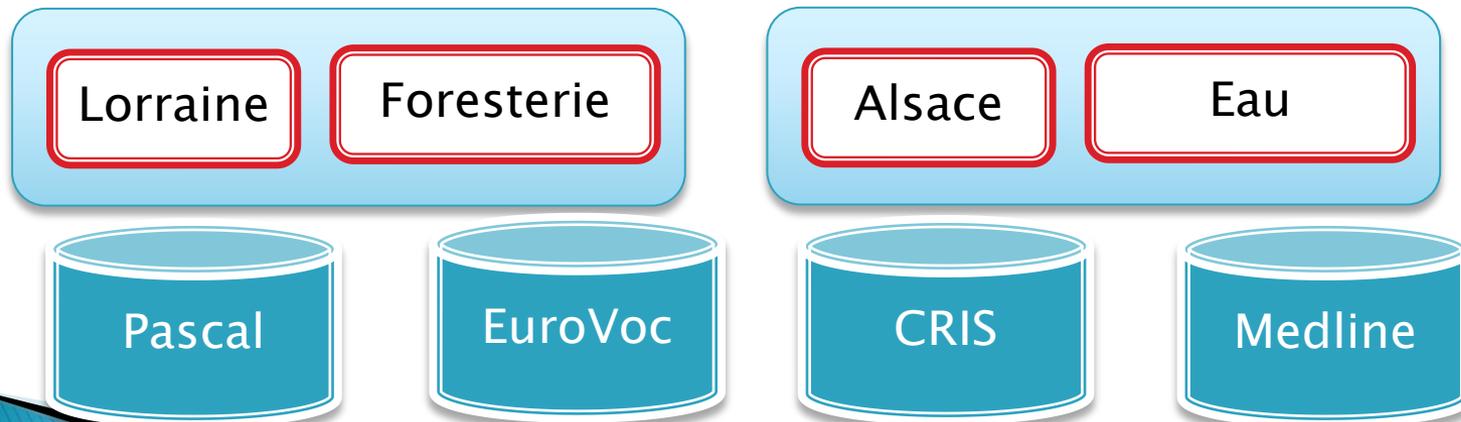
- ▶ https://lorexplor.istex.fr/Wicri/Source/fr/index.php/Utilisateur:Jacques_Ducloy/Blog/S%C3%A9minaire_ISTEX_2017

Contexte politique

- ▶ Pôles international IST/connaissance scientifique sur Nancy :
 - (ATILF + INIST + ...) : 1000 personnes
- ▶ Opportunités : Fin de la Bulle Internet FDP
 - Retour au fondamental : la connaissance scientifique
 - Architecture système d'information sémantique'2000
 - ISTEEX : du carburant pour explorer la science
- ▶ INIST : 150 postes perdus dans les domaines scientifiques :
 - Il faut penser réseau de coopérations
 - ISTEEX : une occasion de déployer des services de proximité

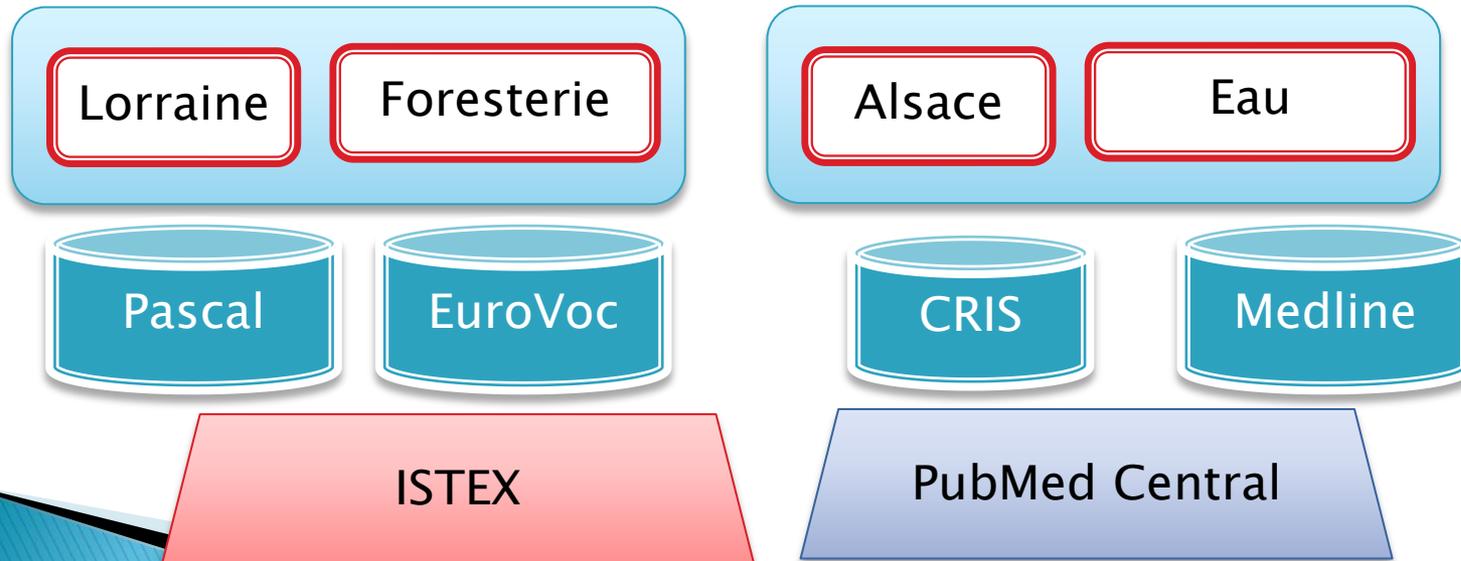
Objectifs WICRI

- ▶ Une infrastructure en réseau, alimentée par les acteurs de la recherche, pour diffuser l'essentiel de la recherche...
- ▶ Retombée technique : un réseau encyclopédique, terminologique et documentaire (type Pascal)



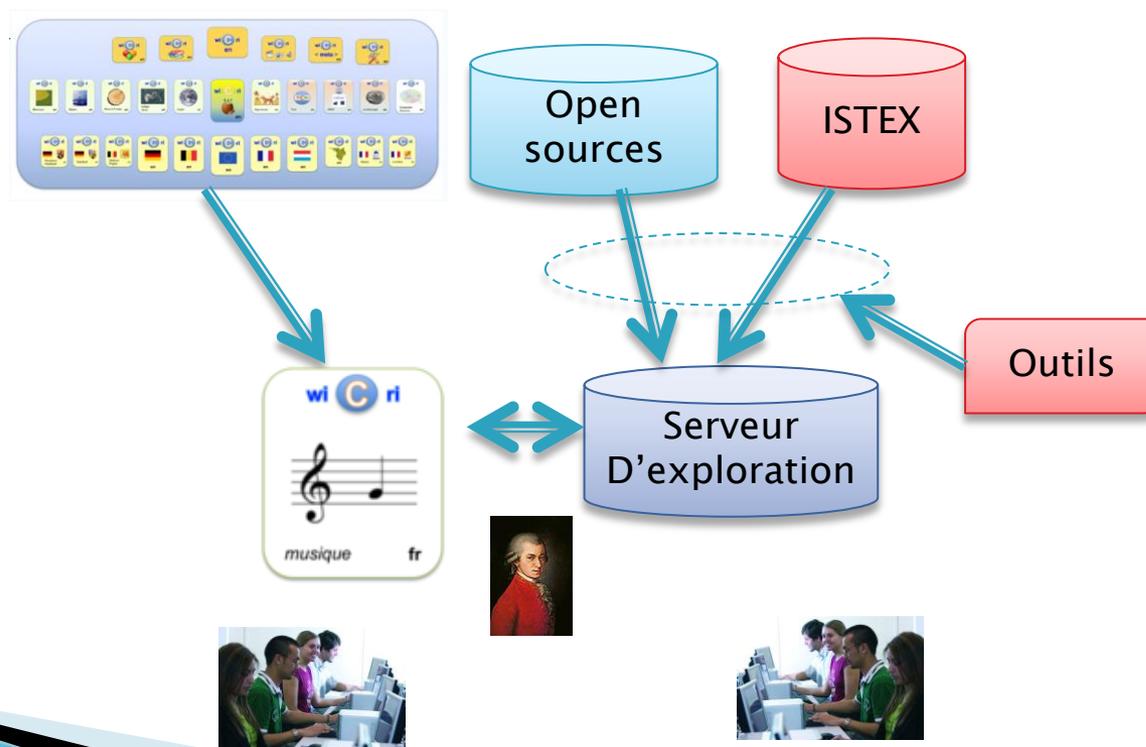
Objectifs LorExplor

- ▶ Objectif initial
 - Coopération INIST – Université de Lorraine pour :
 - analyser les usages des acteurs de la recherche
 - Expérimenter de nouvelles technologies et pratiques
- ▶ Objectif reformulé
 - Démonstrateur en contexte pédagogique



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



Résultat : Semantic MediaWiki pour décrire les systèmes de recherche

The screenshot shows a Semantic MediaWiki page for 'CIDE 2009 Montréal'. The page title is 'CIDE 2009 Montréal'. Below the title, there is a navigation menu with 'page', 'discussion', 'modifier', 'historique', 'supprimer', and 'ren'. The main content area is titled 'Présentation générale' and includes a 'modifier' link. Below this, there is a table with the following data:

CIDE 2009 Montréal	
Cycle :	CIDE
Début :	Oct 21 2009
Fin :	Oct 23 2009
Site :	europa.org/CIDE12
Ville :	Montréal
Pays :	Canada

The page also features a sidebar with a navigation menu and a search bar.

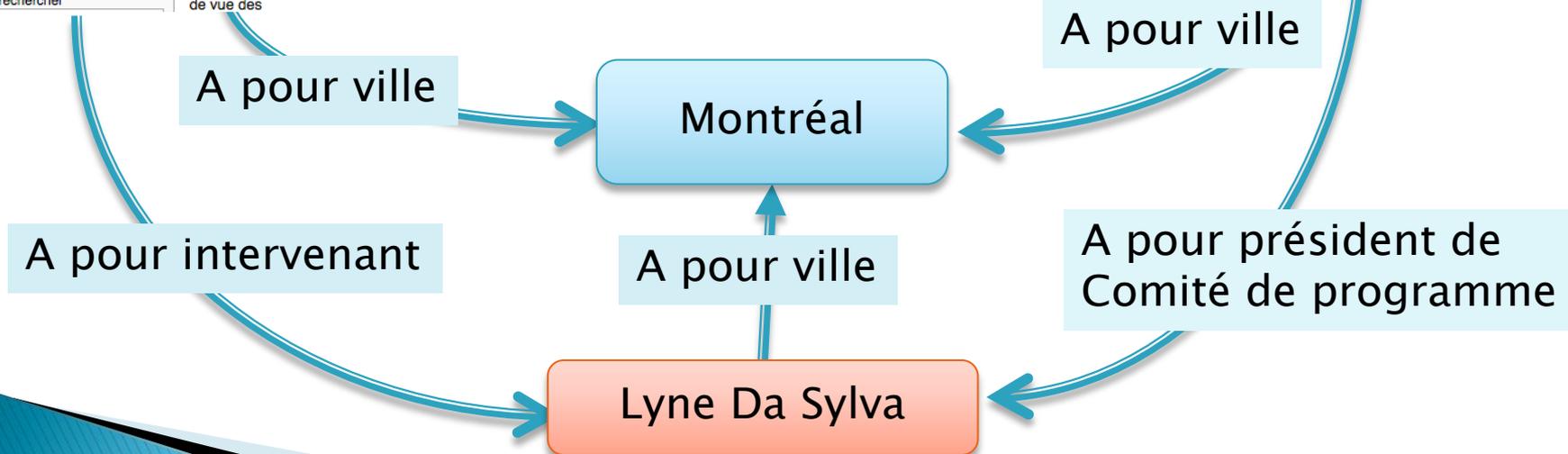
The screenshot shows a Semantic MediaWiki page for 'Acfas 2017 Montréal - Analyser la science'. The page title is 'Acfas 2017 Montréal - Analyser la science'. Below the title, there is a navigation menu with 'page', 'discussion', 'modifier', 'historique', 'supprimer', 'renommer', and 'pr'. The main content area is titled 'Sommaire [masquer]' and includes a table of contents with the following items:

- 1 Introduction
- 2 Organisation
- 3 Programme
 - 3.1 Conférences invitées
 - 3.2 Pratiques communicationnelles
 - 3.3 Usages et usagers

Below the table of contents, there is a table with the following data:

Acfas 2017 Montréal - Analyser la science	
Début :	8 mai 2017
Fin :	9 mai 2017
Ville :	Montréal
Pays :	Canada

The page also features a sidebar with a navigation menu and a search bar.



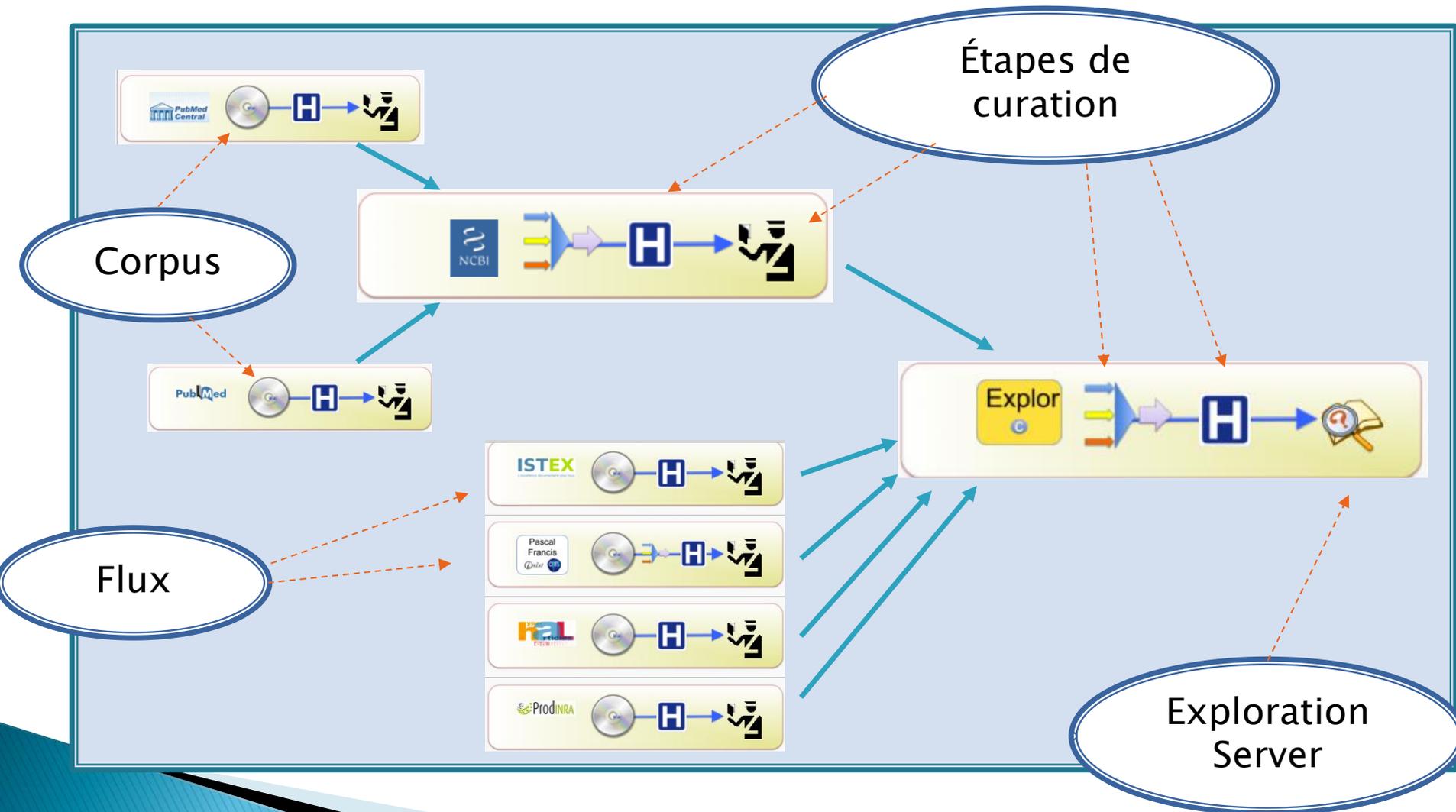
Résultats : Le réseau Wicri

- ▶ 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...
- ▶ Avec la machine virtuelle ISTE_X
 - Faisabilité d'un réseau physique (UL – INIST)
- ▶ Exemples d'expérimentations :
 - Actes des conférences CIDE, H2PTM, VSST, en texte intégral dans une approche sémantique en lien avec ISTE_X
 - Réédition numérique de texte anciens
 - Ecriture collective avec partition musicale
 - Restauration de la revue AMETIST

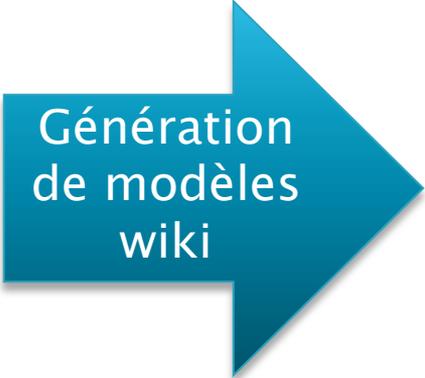
Résultats : les serveurs d'exploration

- ▶ Plus de 100 expérimentations
 - dont Master Science information UL
 - Visibilité de la ville du Havre, (5303 au total/ dont 3003 ISTEEX)
 - 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)

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



Résultat : Serveur d'exploration parcourir des index après curation



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) [↗](#)

Résultat : curation des noms de pays

- ▶ Exemple : identifier les pays dans un contexte hétérogène



Serveur d'exploration sur la didactique - Wicri Wicri

Server d'exploration sur la didactique - Wicri Wicri

ticri.univ-lorraine.fr/wicri.fr/index.php/Serveur_d'explor

Les plus visités | Catégorie:Palais ... | PLOS ONE: "Fres... | Dilib, module Ex... | Getting Started

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 « didactique » qui sélectionne 4013 notices.
3	PubMed Central		Le corpus PMC est extrait avec le critère « didactique » qui sélectionne 402 notices (la forme « didactique » 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 régions



Ces types de carte sont visibles sur tous types de serveurs

Retombées ISTEEX

Changement de paradigme INIST

- ▶ L'API ISTEEX est utilisable pour des traitements avancés...
- ▶ Avec ISTEEX, l'INIST reprend la maîtrise :
 - Du traitement des données des éditeurs
 - De leur mise en ligne
- ▶ BRAVO
- ▶ Une inversion de tendance est-elle possible ?
- ▶ Oui mais :
 - Penser réseau
 - Mettre le centre de gravité sur l'aval...

Retours d'expériences

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)*

Retours d'expérience

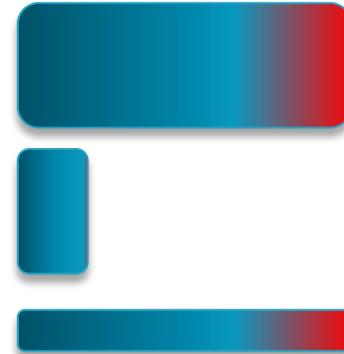
Corpus / méfiance / curation

▶ 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.



Retours d'expérience :

Le chantier Machine virtuelle

- ▶ Une machine virtuelle est disponible depuis mi avril 2017
- ▶ Elle est livrée nue sans aucun logiciel d'application...
- ▶ On doit y installer
 - environ 80 wikis avec un environnement technique très récent (1.28)
 - Environ 100 serveurs d'exploration
- ▶ Avec très peu d'assistance (pour l'instant)
- ▶ Expérience très formatrice mais... non reproductible
- ▶ Sans un minimum d'assistance les machines virtuelles ne seront pas utilisées...

Résultat : la pertinence du slogan ISTEEX

- ▶ L'excellence documentaire pour tous...
 - Possible avec un traitement en aval
- ▶ Le démonstrateur LorExplor
 - peu de moyens et peu de soutiens institutionnels
 - montre l'intérêt des solutions techniques
 - Synergie réseau SMW, Unix/xml, Corpus
- ▶ Mais ce n'est qu'un démonstrateur...
 - Utilisable
 - pour expérimentations,
 - formations
 - Solutions ponctuelles
 - Evaluation d'un déploiement à grande échelle

Conclusion : pour aller plus loin

- ▶ Un grand projet de la taille « Pascal 1990 » est-il possible ?
- ▶ 100 universités ou établissements avec une cellule « valorisation numérique » de 4 personnes = 400 personnes !
- ▶ Les réseaux sémantiques induisent un changement de paradigme avec une démarche basée sur
 - la confiance (modération a posteriori),
 - L'utilisateur omniprésent,
 - L'expertise mutualisée sans périmètre institutionnel

Conclusion

- ▶ ISTEK est un accélérateur potentiel
 - Mais pas une fin en soi...
- ▶ Pour devenir un succès, il doit associer
 - Traitement en amont et en aval
 - Exploration des connaissances anciennes et construction de connaissances nouvelles..
- ▶ LorExplor veut continuer à explorer ce nouveau paradigme
 - Si possible en s'intégrant à un grand projet...

Après cette conclusion...

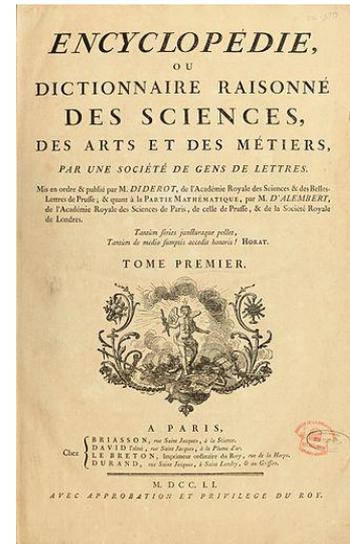
- ▶ Merci pour vos questions
- ▶ Visite guidée..
- ▶ <https://lorexplor.istex.fr/>
- ▶ Les diapos qui suivent sont en réserve pour questions éventuelles.

Recherche d'information – vs – Exploration des connaissances

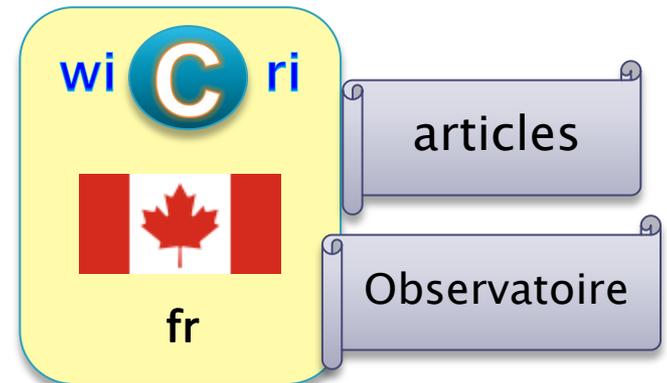
- ▶ Un article de Tim Berners-Lee sur le web sémantique ?
- ▶ Quelle est la plus ancienne référence à l'hypertexte ?
- ▶ Quelle est l'œuvre de Mozart la plus citée ?
- ▶ Concernant le Kazakhstan :
 - Quelles sont les principales coopérations entre la Lorraine et le Kazakhstan ?
 - Avec quels laboratoires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

Cyberinfrastructure humaniste: Construire de la connaissance

- ▶ Publier le résultat d'une démarche
 - scientifique ou culturelle,
 - en mode collaboratif et hypertexte
- ▶ MediaWiki : moteur de Wikipédia
 - L'excellence du texte scientifique hypertexte
 - Conception incrémentale (modèles)
- ▶ Compatibilité MediaWiki (version 1.28)



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



Structurer la connaissance et les règles de curation...



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

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

A pour auteur de livret



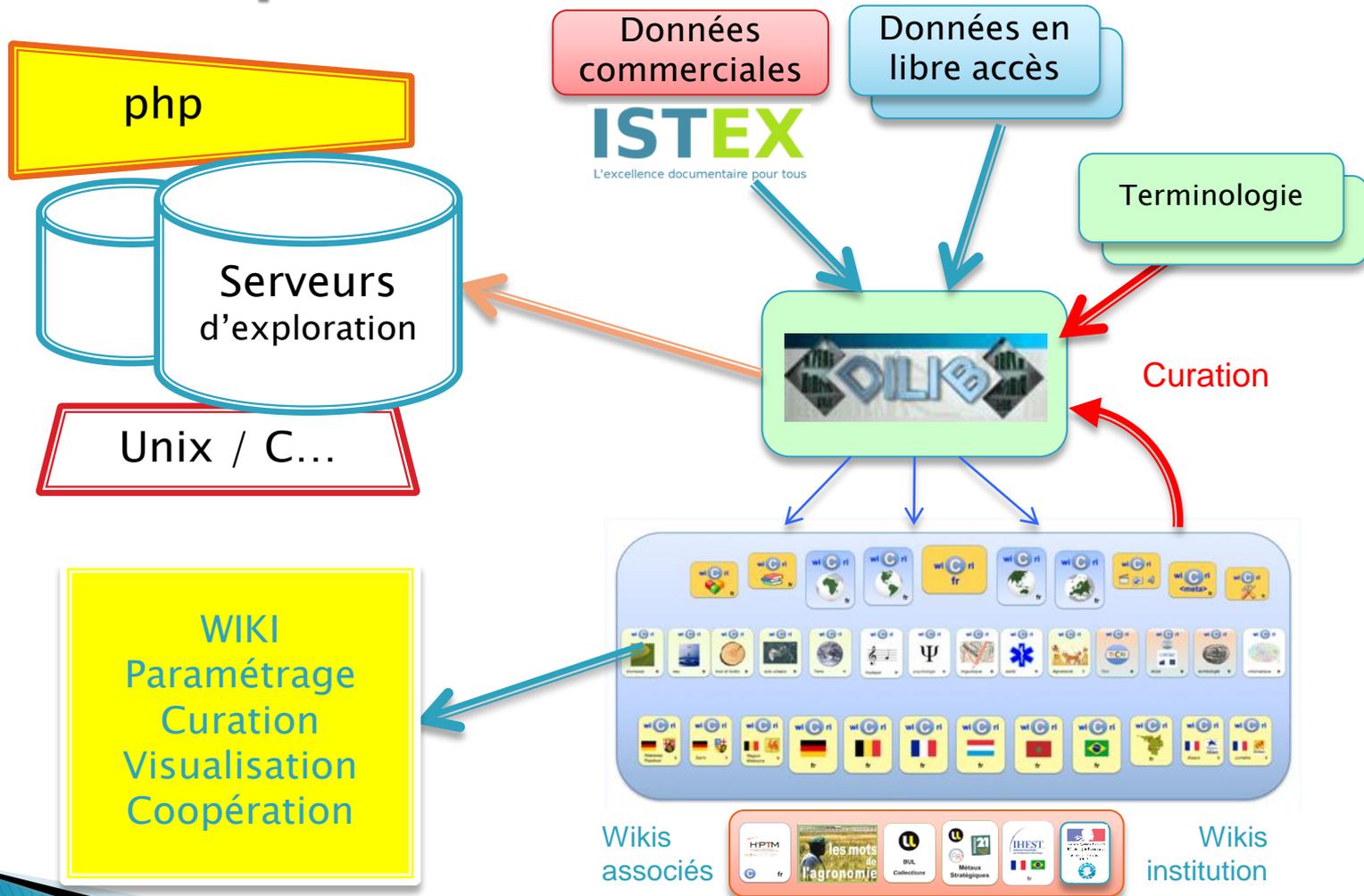
Liste des opéras ayant un livret de Lorenzo da Ponte

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 :
```

Un atelier flexible pour explorer des corpus



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>
```

Filtrage du texte

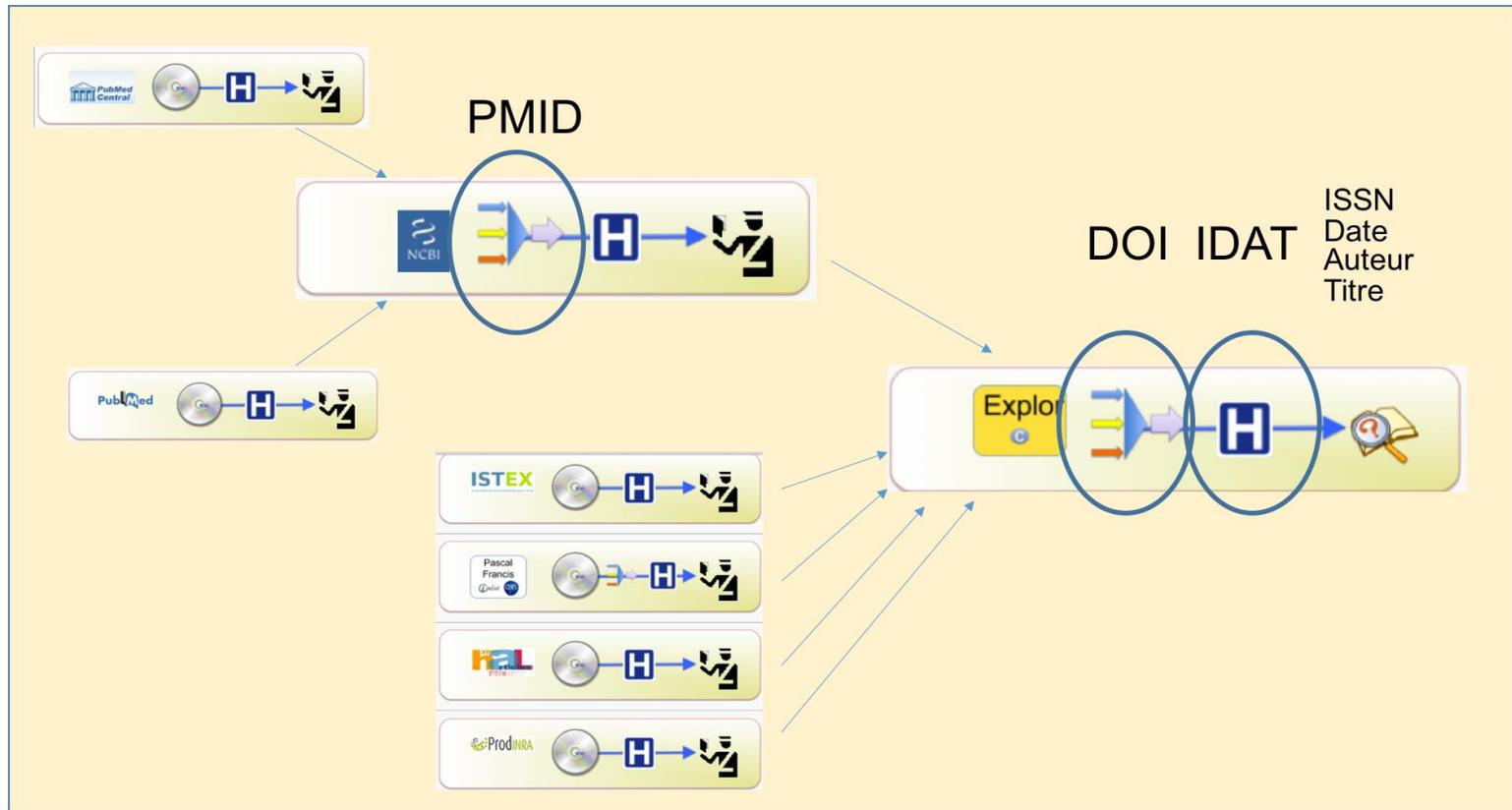
Heuristiques :

- ▶ 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
```

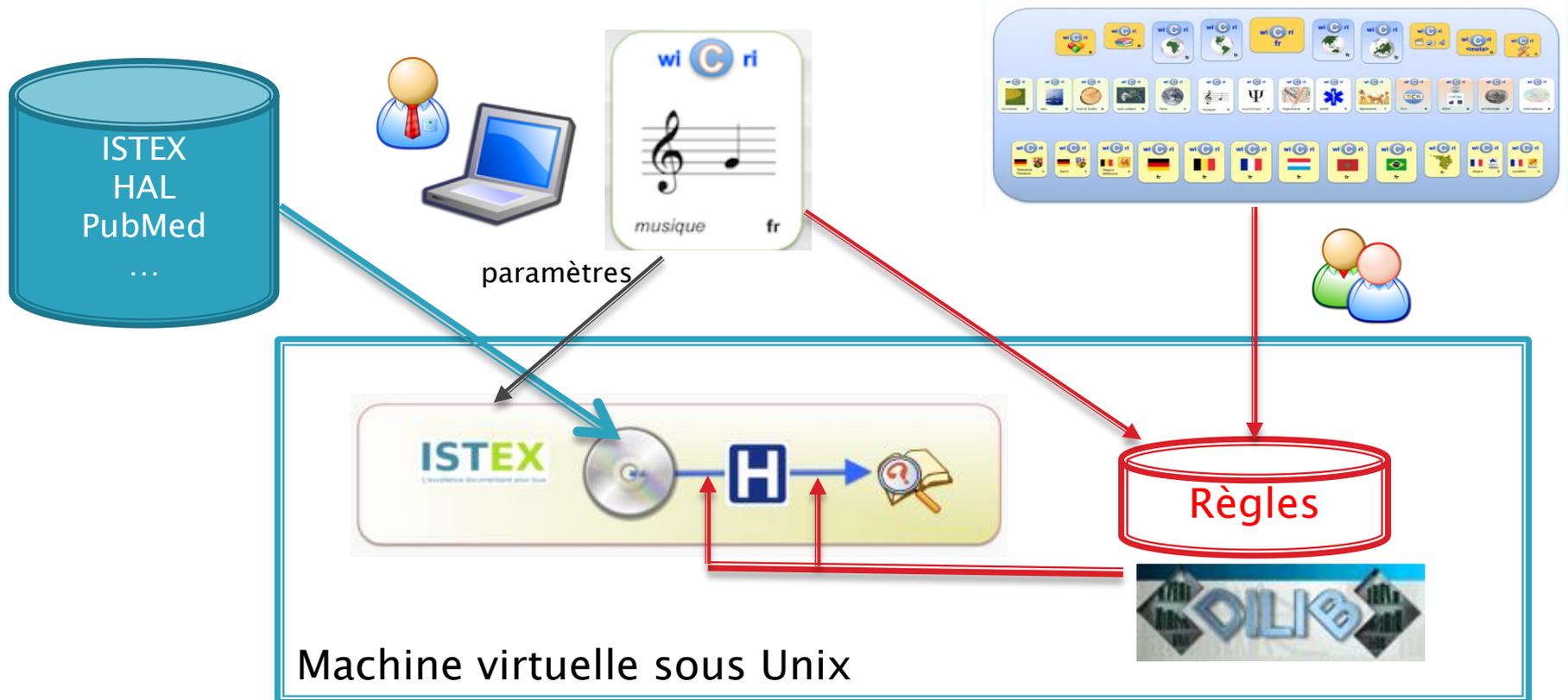
- ▶ Recherche de noms binomiaux
 - Recherche de chorégraphes (nom-prénom)
- ▶ Identification du cœur d'un corpus
 - (termes pondérés)

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



Génération de plateformes de curation et d'exploration

► Processus itératif mutualisé



Curation des données – 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 régions /codes postaux

Sur Wicri/Allemagne

ville <input type="checkbox"/>	code 4 chiffres <input type="checkbox"/>	code 5 chiffres <input type="checkbox"/>	formes courantes <input type="checkbox"/>	district/land <input type="checkbox"/>
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

Curation des universités



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