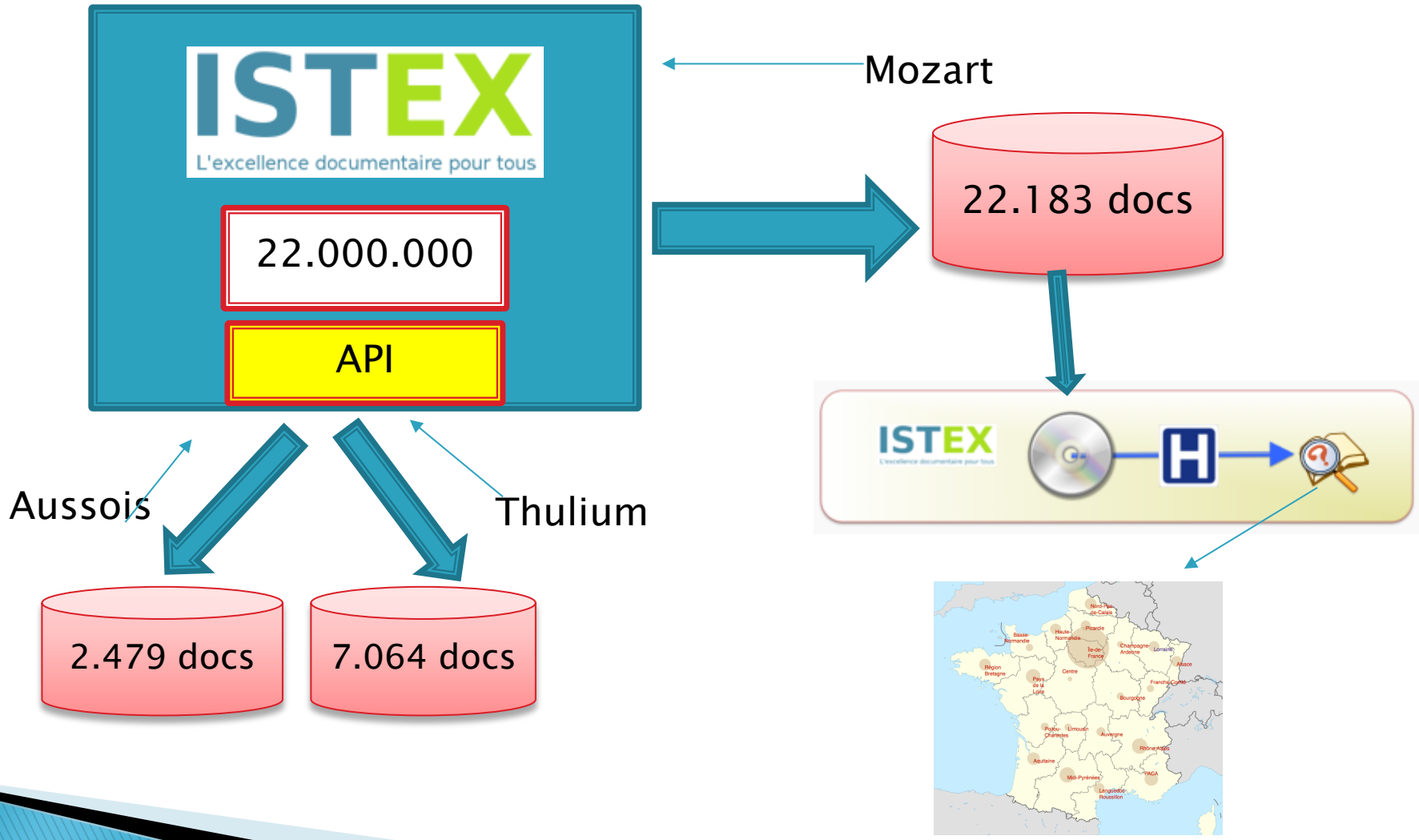
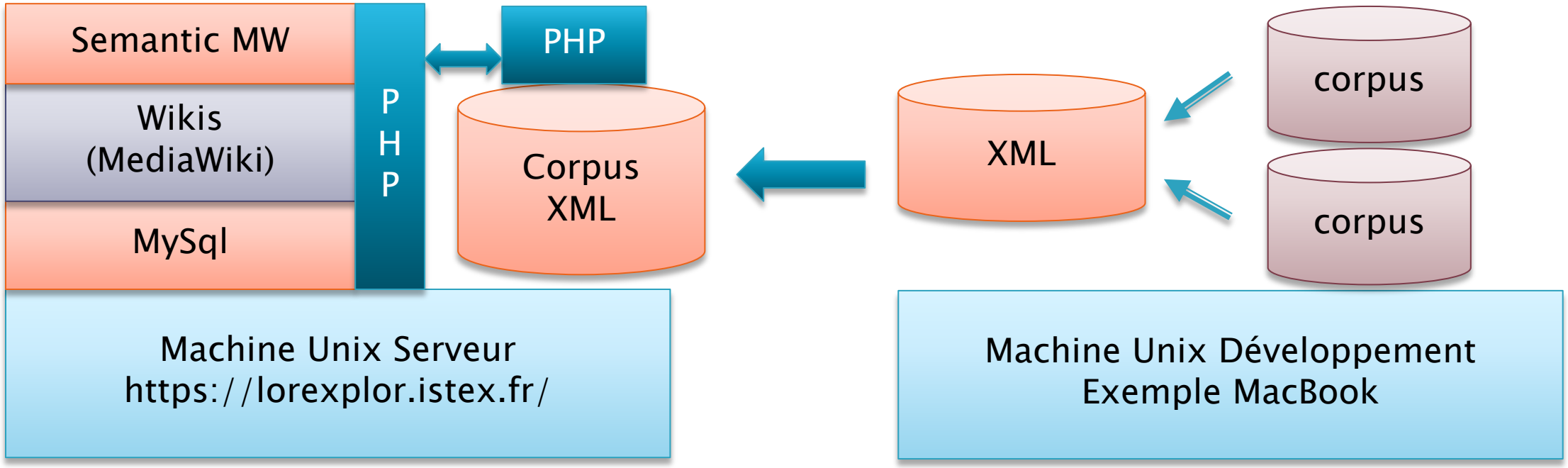
Gérer la volumétrie avec l'ingénierie XML

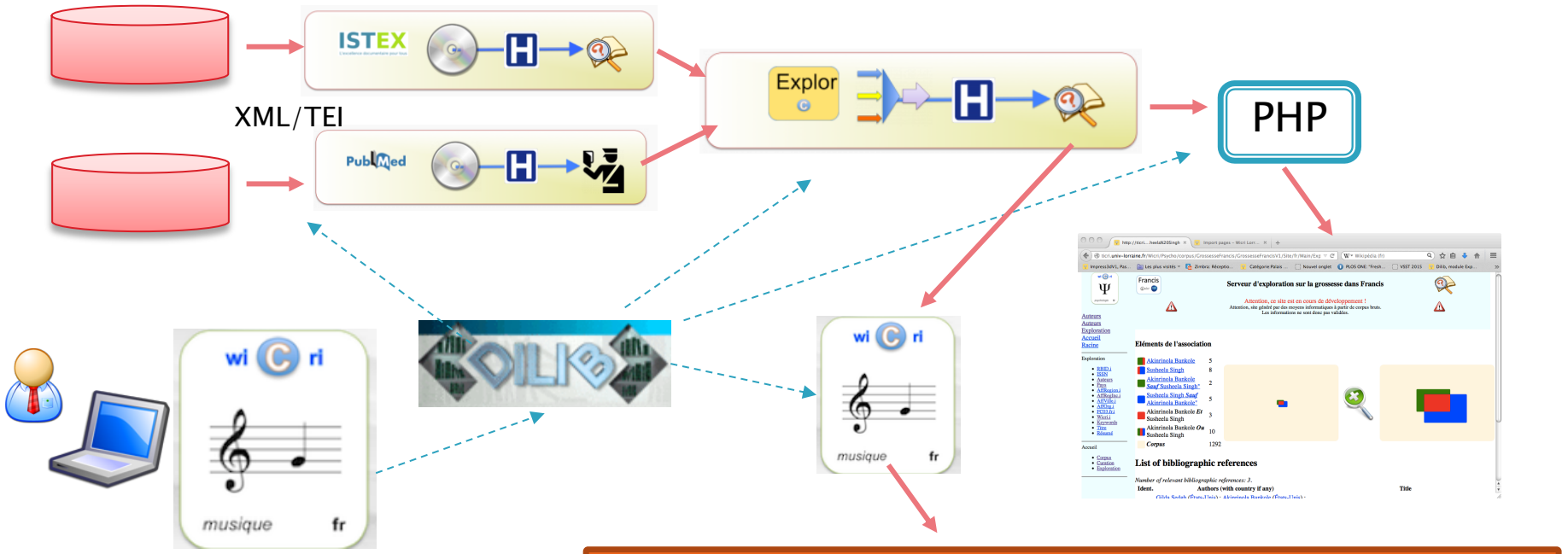
ISTEX – Serveurs d'exploration



Machine serveur – machine développement Unix, langage C, PHP, XML, JSON, etc…



ISTEX – Serveur – génération



{{Explor plateforme MozartV1 /Carte France|taille=400}}

MediaWiki
Because ideas want to be free.

Pays	Région	Villes
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)	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)	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	**Mots des titres**	**ISSN/revue**
: 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)	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)	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
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)

Combinaison d'index : AutAff

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

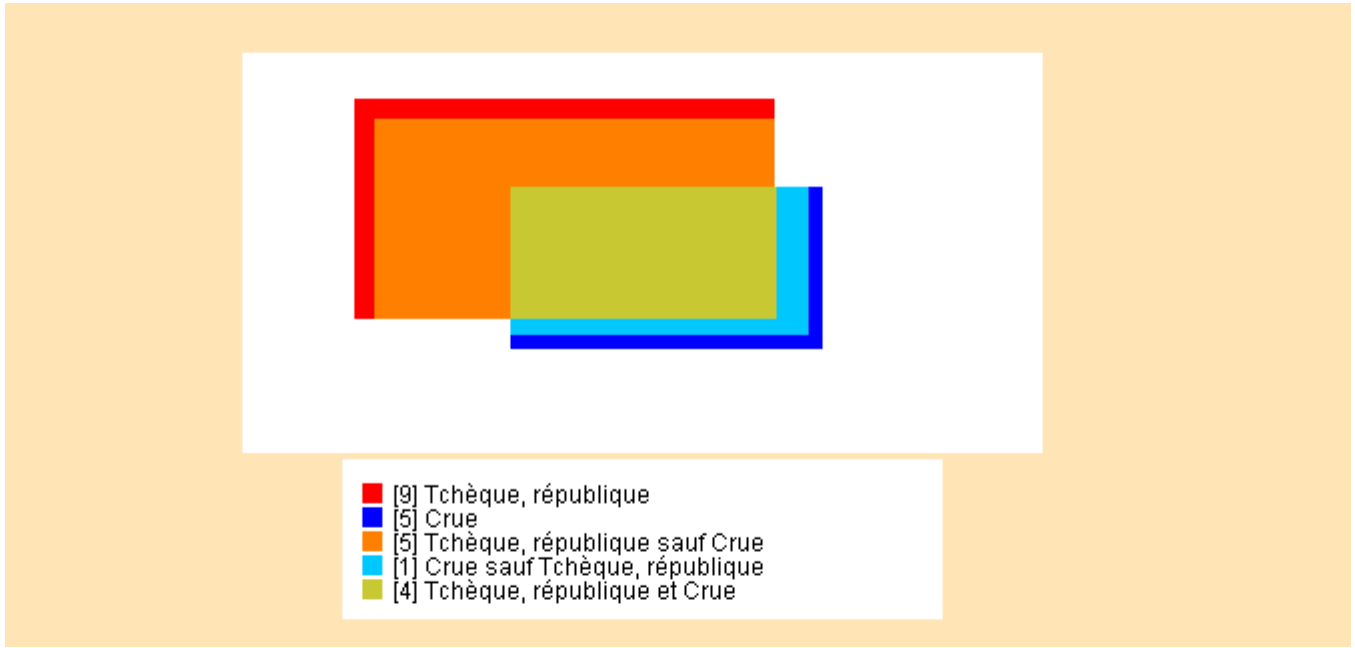
	Department of Neurology, Juntendo University School of Medicine, Tokyo	004177
	Department of Neurology, Juntendo University School of Medicine, Urayasu Hospital, Tokyo, Japan	002388
	Department of Neurology, Juntendo University, School of Medicine, Tokyo, Bunkyo-ku, Japan	004111
	Department of Neurology,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

Associations

- Permettent de visualiser les relations liant 2 concepts

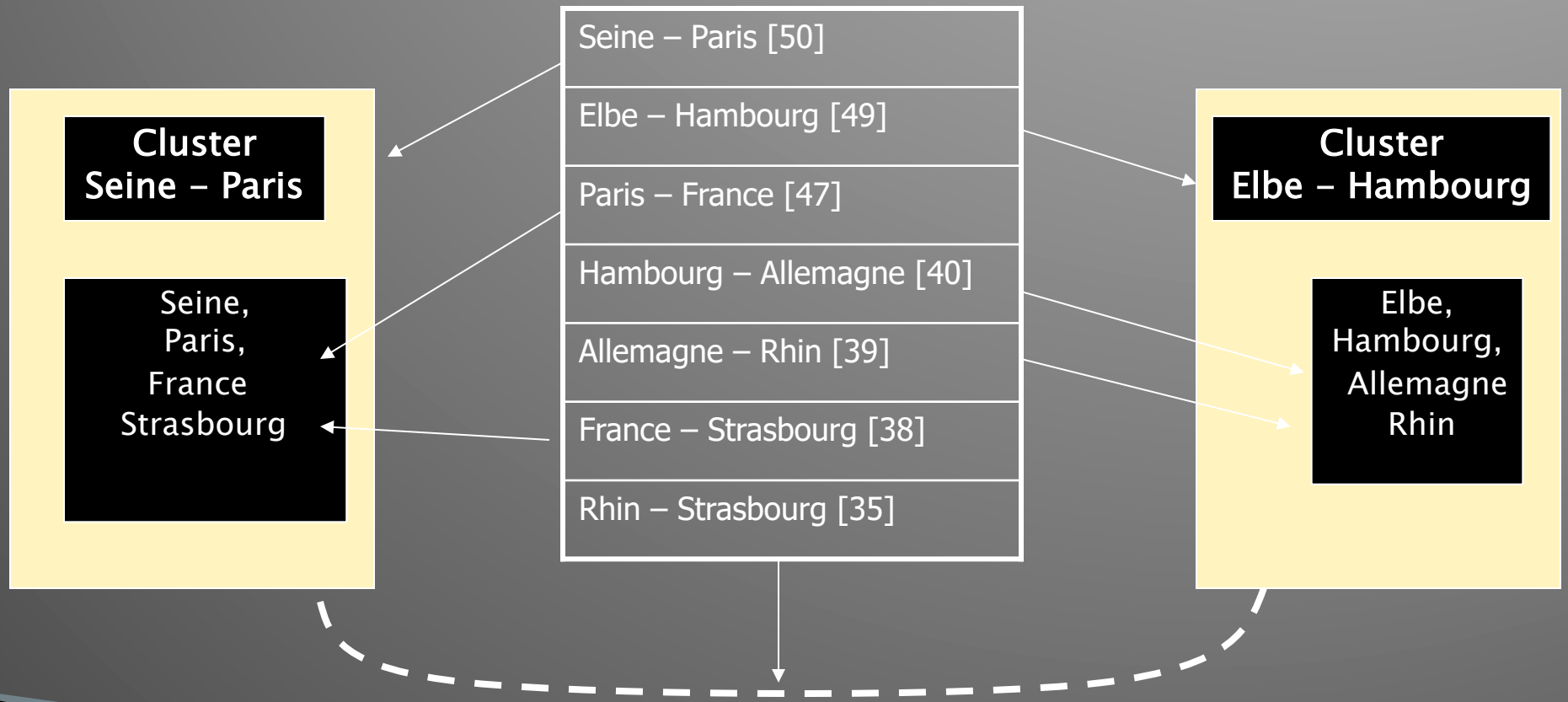


Liste d'associations

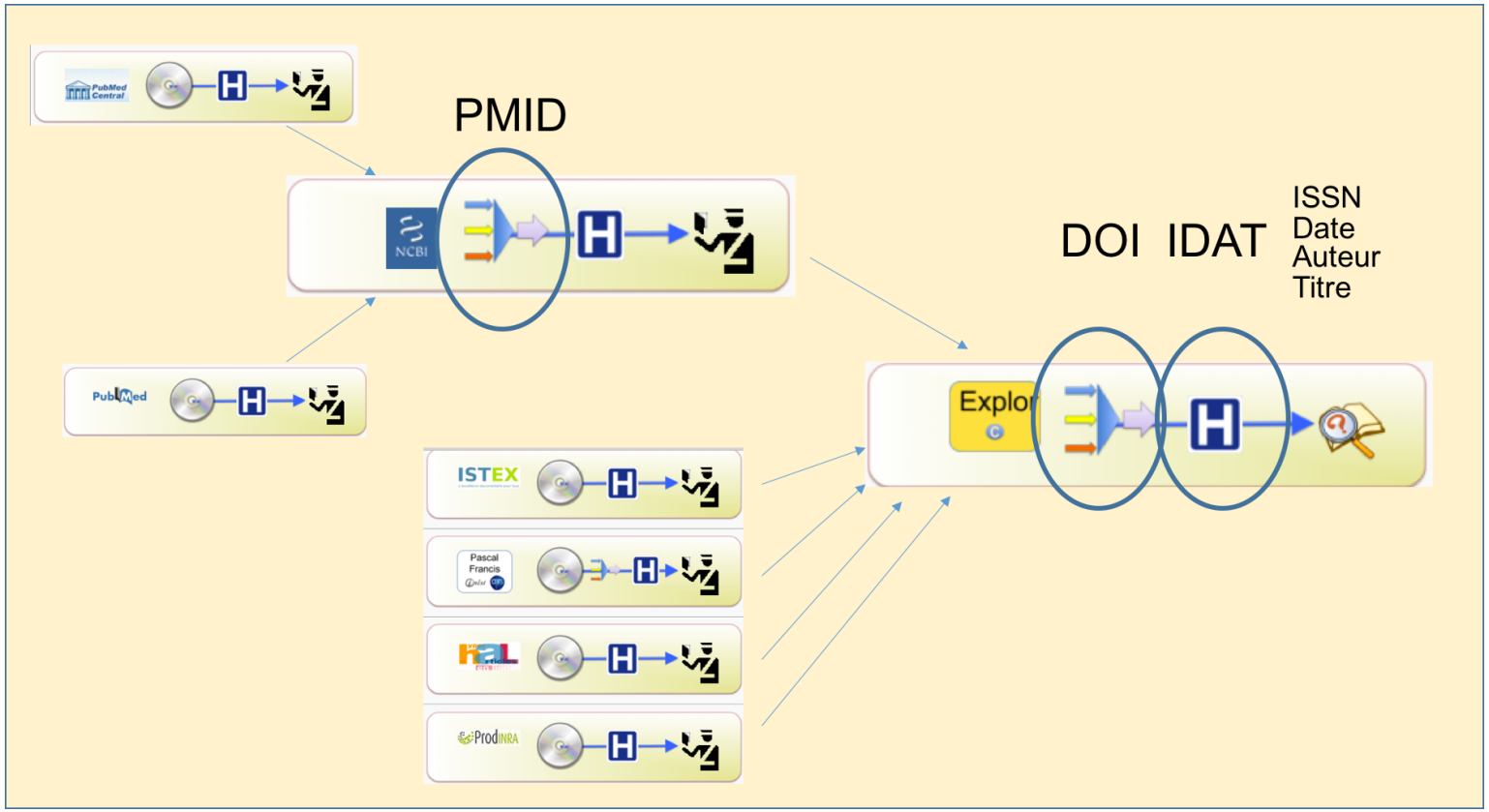
- Intéressant mais difficilement utilisable
- Exemple : base sur l'hydrographie en Allemagne

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
Fluviatile - Allemagne	16
Ecoulement - Allemagne	15
Baden-Württemberg - Allemagne	15
Hessen - Allemagne	14
Végétation - Allemagne	13

Clusterisation

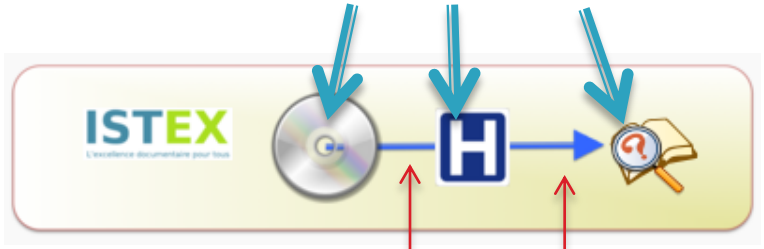


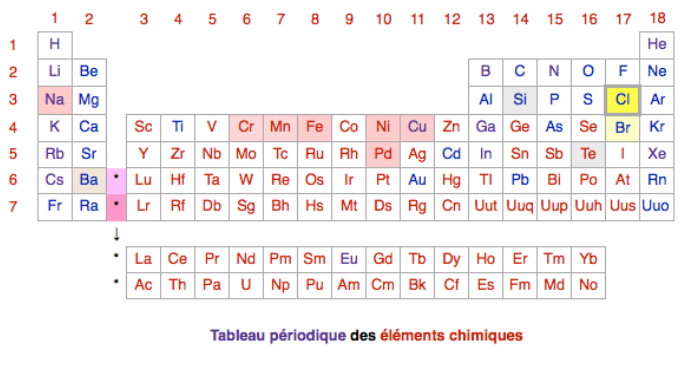
Enrichissement : dédoublonnage ISTEX / Pascal / Hal / MEDLINE…



Serveur d'exploration

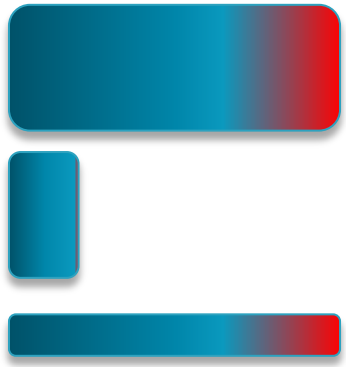
Système d'information orienté exploration





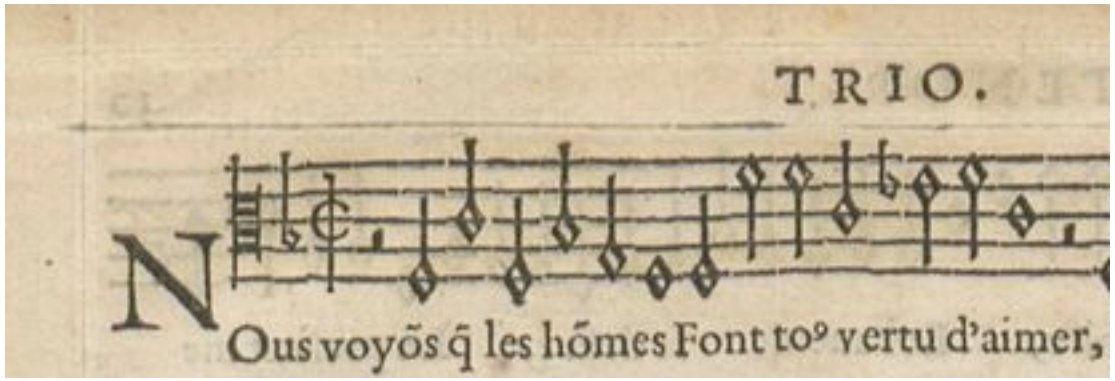
Corpus : méfiance / curation

- 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.

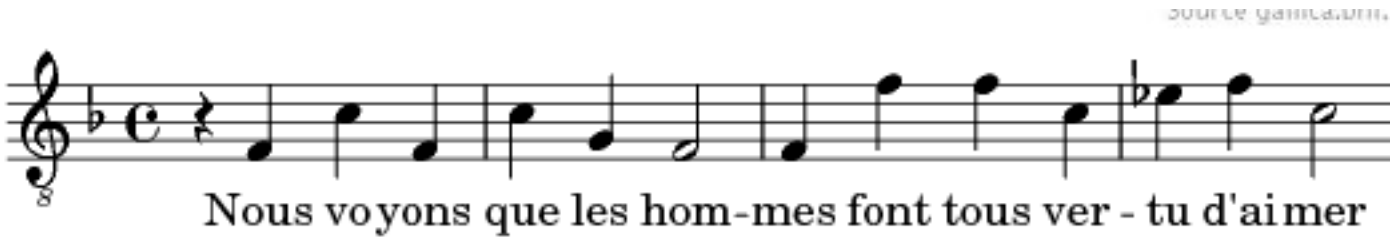


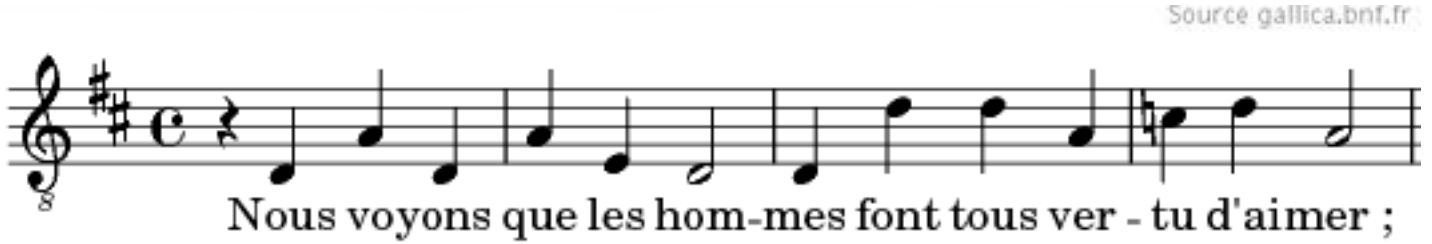
- **Quelle formation donner à un bibliothécaire pour accompagner un chercheur dans une démarche de curation?**

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();
}
```





Dilib, une boîte à outils Sxml

- SXML : XML lite (mais JSON+)
  - Compatible avec les outils Unix
    - Un document = Une ligne Unix
- Origine
  - 1990 : Ilib : ISO 2709 (MARC, Pascal…)
    - Un LEGO pour les corpus
  - 2000 : Dilib : métadonnées hétérogènes
- 2018 : LorExplor
  - 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>
```

Dilib, quelques exemples

- Extraction des exemples du Trésor de la Langue française pour une conversion en WikiTexte

♦ *Grand orgue* ou *grandes orgues.* Orgue le plus important d'une église, placé souvent dans une tribune au fond de l'église, par opposition à *petit orgue* ou *orgue de choeur,* orgue de dimensions restreintes, souvent placé dans le choeur. *Le* Gloria in excelsis *divisé entre le grand et le petit orgue, l'un chantant seul et l'autre dirigeant et soutenant le choeur, exultait d'allégresse* (HUYSMANS,*En route,*t. 1, 1895, p.51).

- *Le* Gloria in excelsis *divisé entre le grand et le petit orgue, l'un chantant seul et l'autre dirigeant et soutenant le choeur, exultait d'allégresse* (HUYSMANS, *En route,*t. 1, 1895, p.51).

- Conversions de chartes en TEI
- Extraction de documents sur Gallica

Exploration, Filtrage

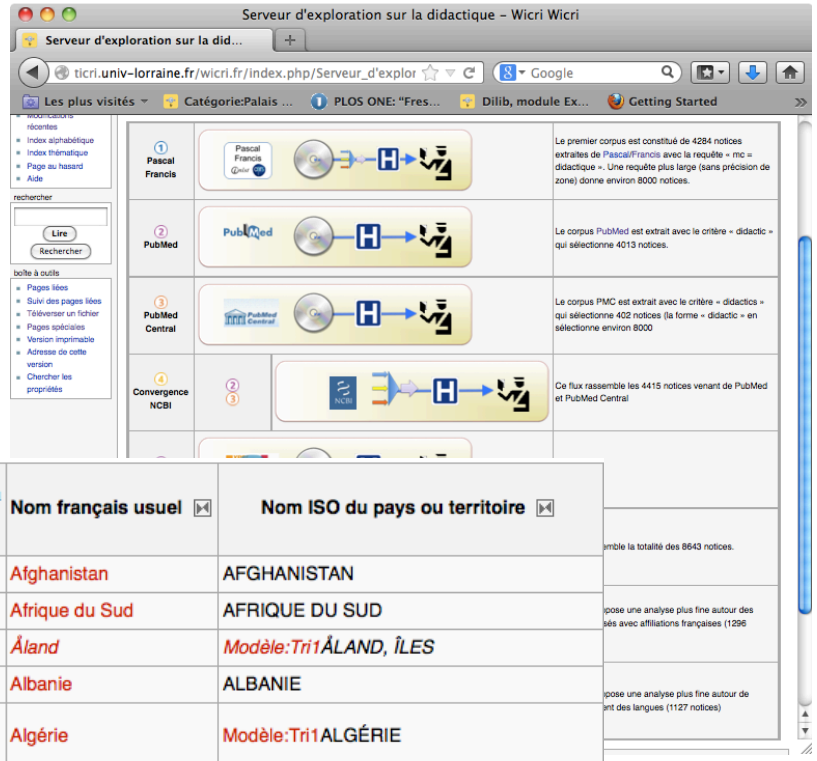
- Quelles sont les œuvres de Mozart les plus citées dans un corpus ?
  - Idée générale : utiliser le catalogue Köchel
    - Résultat : Sonate KV. 448

```
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

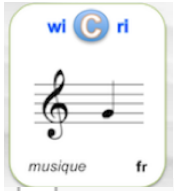
Curation des données

- 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*

Règles de curation des données



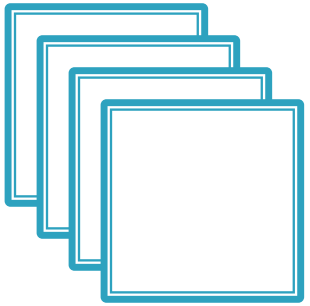
Myriam Chimènes	Myriam Chimènes	affiliation : Institut de recherche en musicologie
Denis Herlin	Denis Herlin	affiliation : Institut de recherche en musicologie ; affiliation : Université de Tours
Paul Henry Lang	0027-4631:P. H. L.	affiliation : Université Columbia
Edward Lowinsky	Edward E. Lowinsky	affiliation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

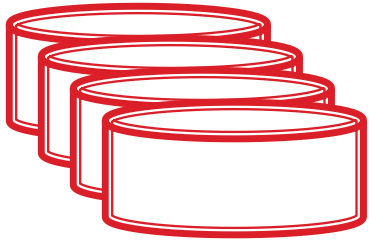
Santé : Serveurs à génération rapide (5')

```
NlmPubMedExplorCorpus –q influenza –s 1000
```

corpus



Pages paramètres



Base Xml

MeSH english -> Français

Santé : PubMed enrichi par ISTEX / HAL

- PubMed
  - 30.000.000 articles (métadonnées)
  - Indexés par des spécialistes (chercheurs, médecins)
- ISTEX
  - 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