Semantic MediaWiki networked Infrastructure for Text and Data Mining with ISTEX

COLLNET 2016 Nancy

ISTEX

L'excellence documentaire pour tous

- French national initiative
  - Investments for the future
  - *Investissements d'avenir*
- Catch word
  - *Documentary excellence for everbody*
- Equipment
  - Budget : approximatively 60.000.000 €
- Archives and data
  - target: 50 000 000 documents
  - At this time: 17 000 000
  - Springer, Wiley, Elsevier, Oxford University Press…
  - Various formats: text (OCR), metadata, XML…
- A portal is available for French academic people

ISTEX

L'excellence documentaire pour tous

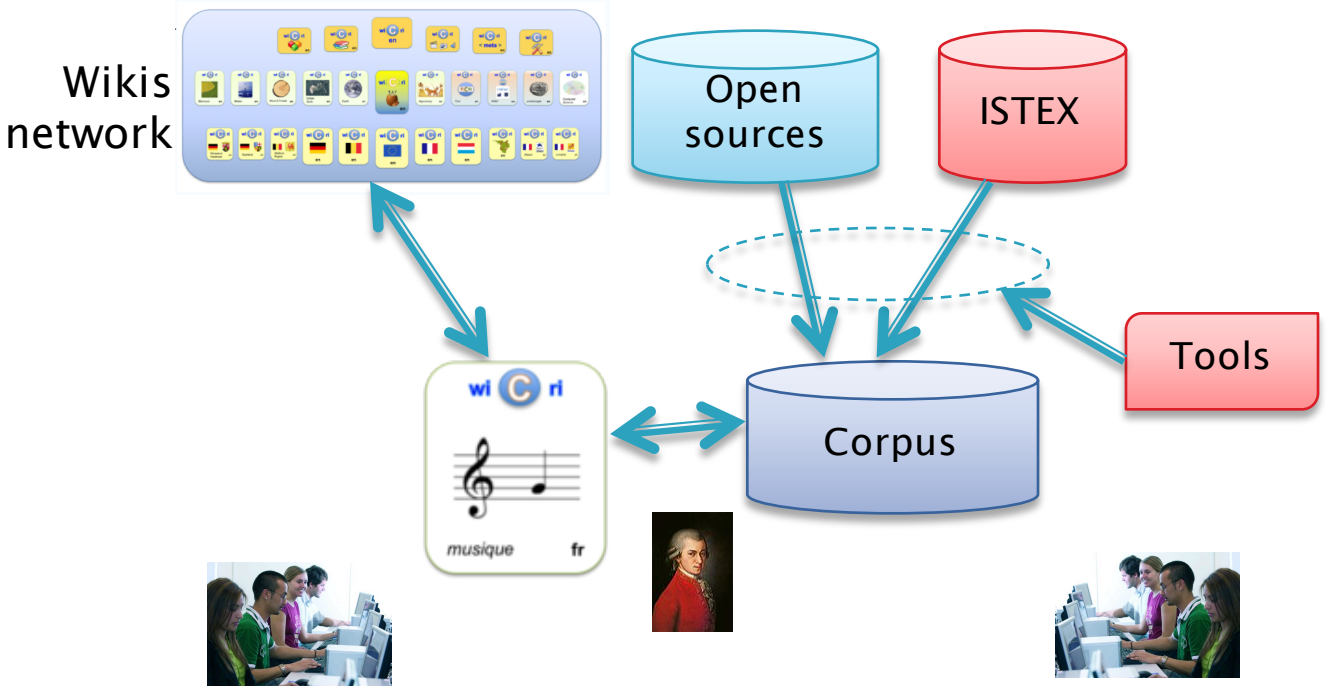
INVESTISSEMENTS D'AVENIR

Two questions:

- What can a young researcher do with 10,000 documents related to his thesis topic?
  - In France, how many librarians and teachers need to be trained: 1,000 ? 10.000 ?
- Due to copyright constraints:
  - How ISTEX could be usefull for the French society ?
  - How ISTEX could be used servir In the framework of international cooperation?

LorExplor: exploring new practices

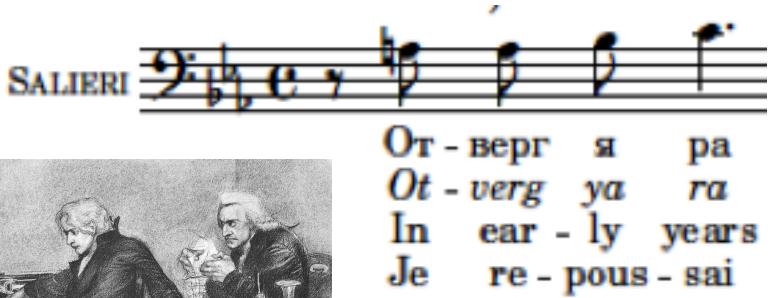
- Text & data mining with time constraints
- Co-construction of scientific or cultural portal

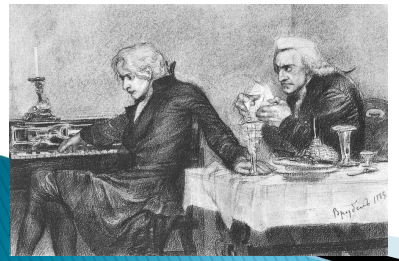


Building Knowledge with MediaWiki

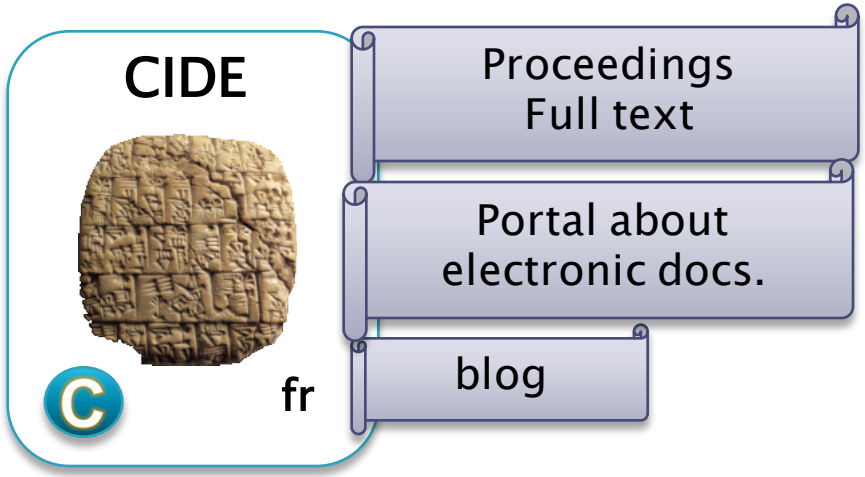
- MediaWiki : Wikipedia engine
  - The excellence of the scientific text
  - In a hypertext landscape
  - Cyberinfrastructure for collaborative knowledge building (ex Wikipedia)

A French text from Pouchkine on the music of Rimsky Korsakov (Mozart and Salieri)





About electronic documents

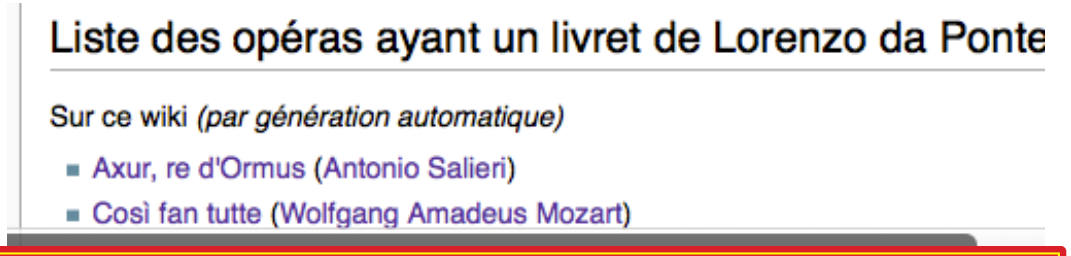


Structuring Knowledge with SMW


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

```
with a libretto by:
[[Has libretto creator::Lorenzo da Ponte]].
```

Has libretto creator




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

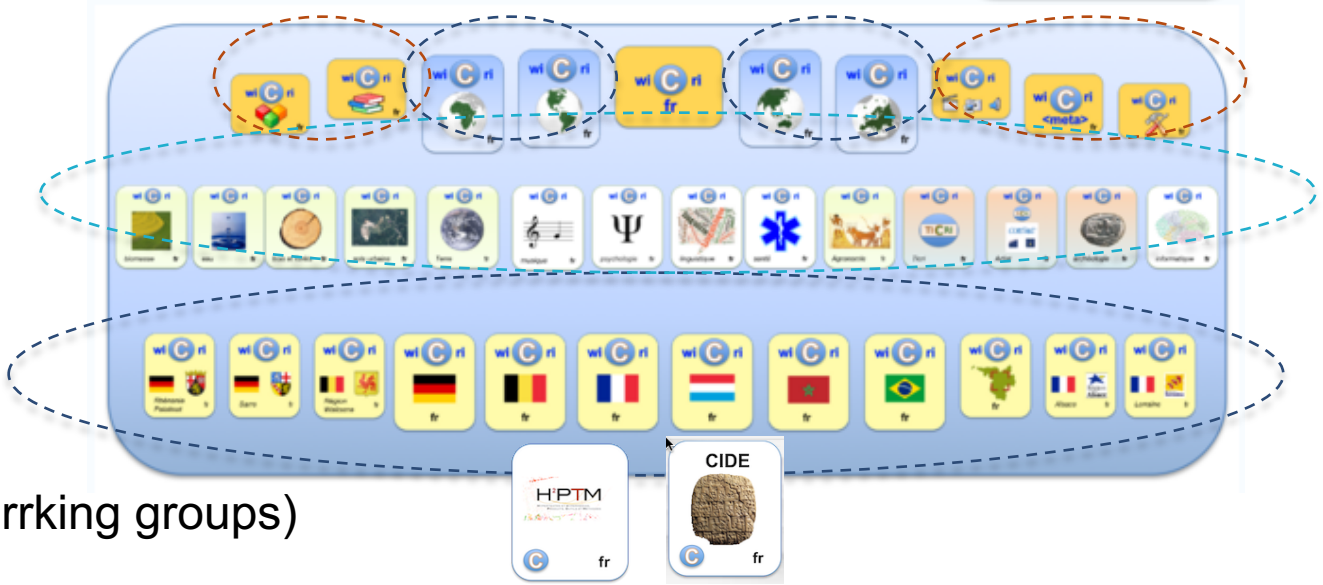
Reading and writing science in an interdisciplinary world

Technical wikis

Thematic wikis

Regional wikis

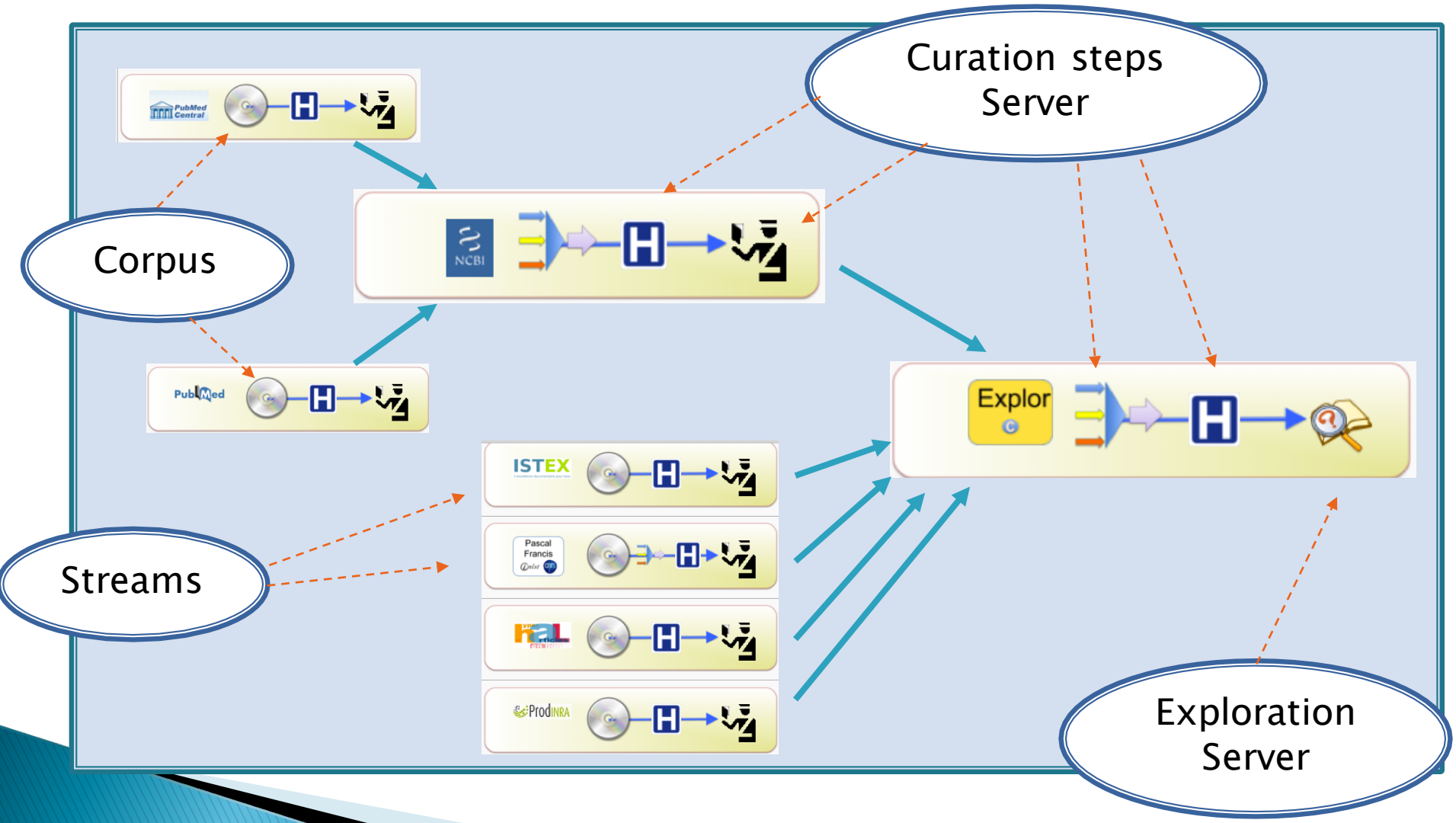
Associated wikis (worrking groups)



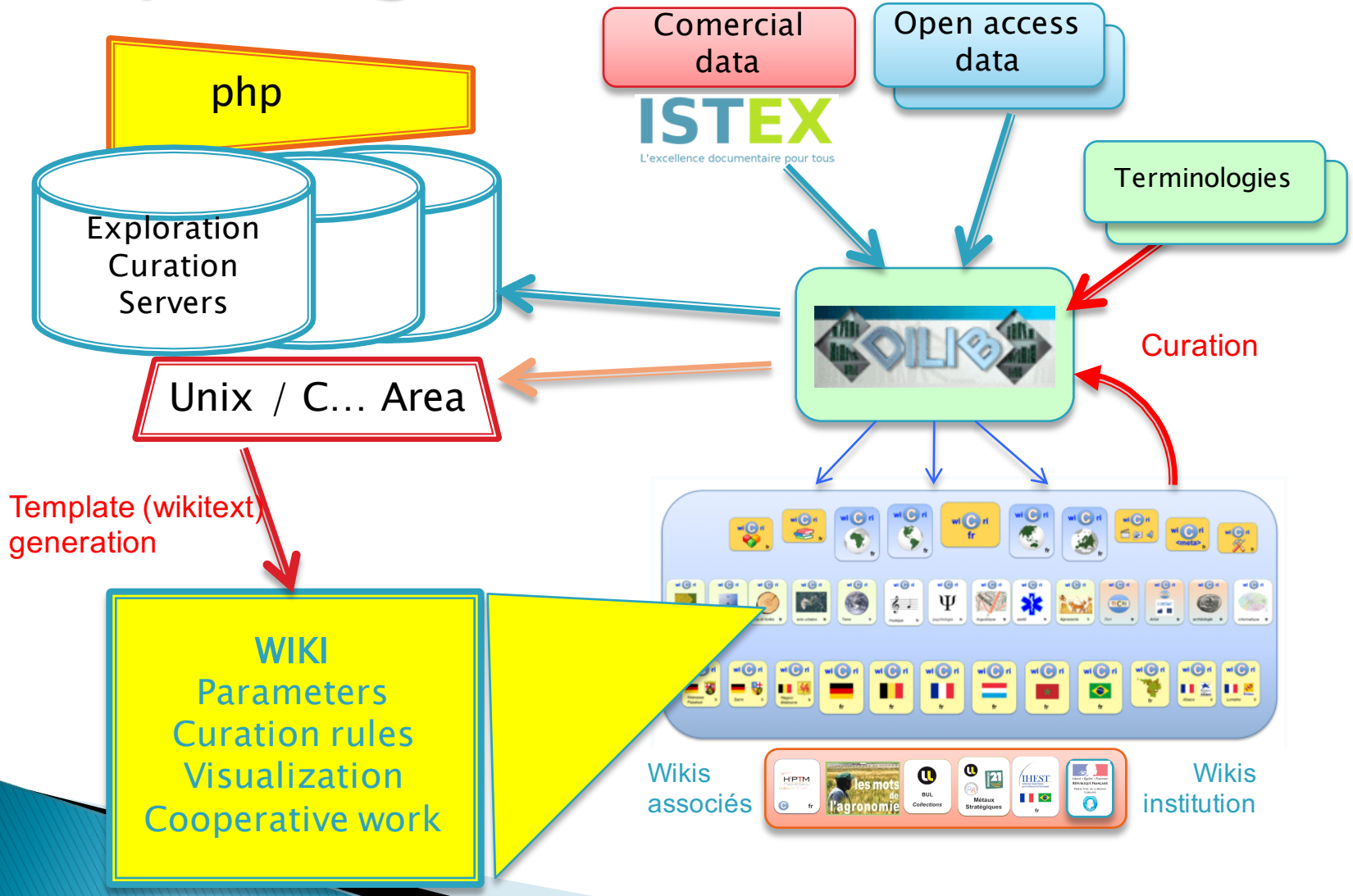
No anonymous !!!!

Any registered professional (researcher, practitioner, engineer ...) can contribute to any wiki

Exploration/curation area



Flexible workshop for corpus exploring

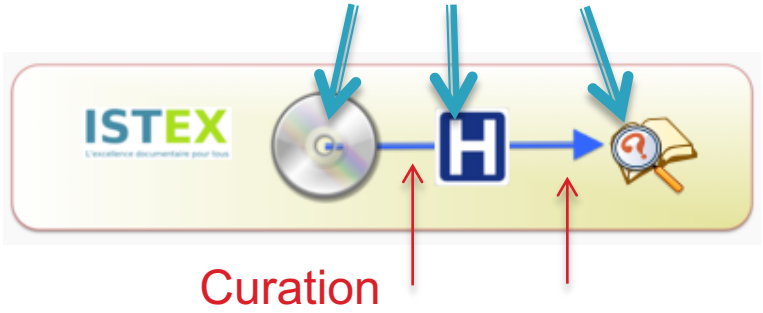


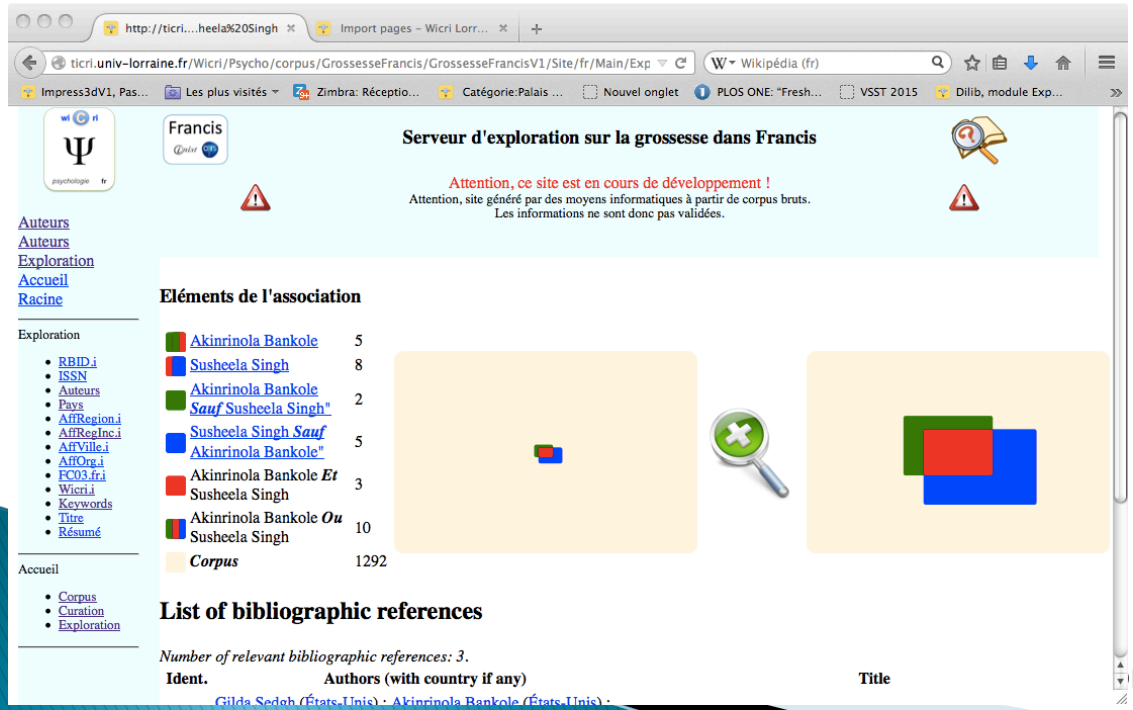
Information retrieval – vs – Knowledge exploration

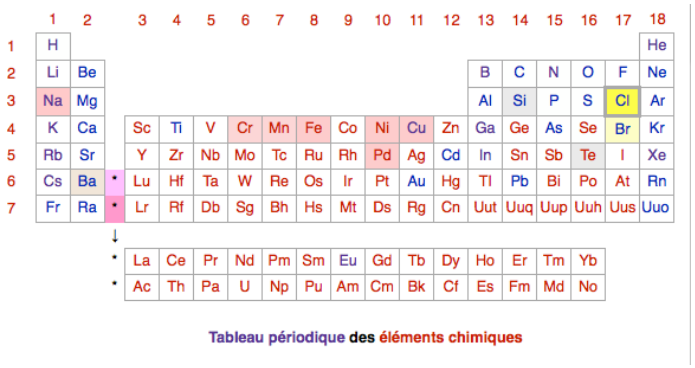
- An article by Martin Hatzinger on the death of Mozart?
- What is the oldest reference to hypertext? What is the most quoted work of Mozart?
- Concerning Kazakhstan:
  - What are the main cooperations between Lorraine and Kazakhstan?
  - With which international laboratories can the University of Lorraine ally in order to cooperate with this country?
- What are still wild fish that can be domesticated?

Exploration area and servers

Several servers by stream (curation steps)







Dilib, an Sxml Toolbox

- Sxml
  - XML light
  - Unix pipe compatible
    - a document = a unix line…
- 1990: Ilib
  - Needs: ISO 2709 (MARC formats)
  - SGML with an XML way
  - A « Lego » for Information retrieval systems
- 2010 : Dilib 0.5…
  - Needs: fulltext corpus
  - Many documents with Many DTD


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

Exploration servers
Index Browsing with wiki summary

Template generation

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)

Index combination: AutAff

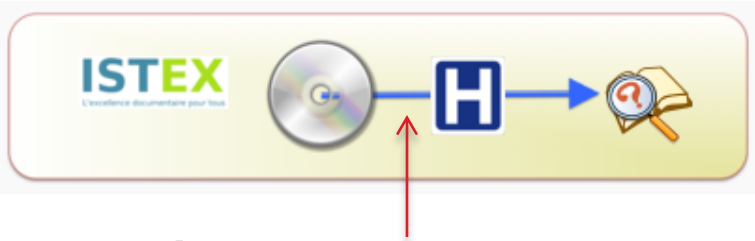
- Author « root name »
  - Last name + initial
- + Affiliations
- Initial objective:
  - Curation
  - Homonymies solving
- Pragmatic issue: actors discovering

	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

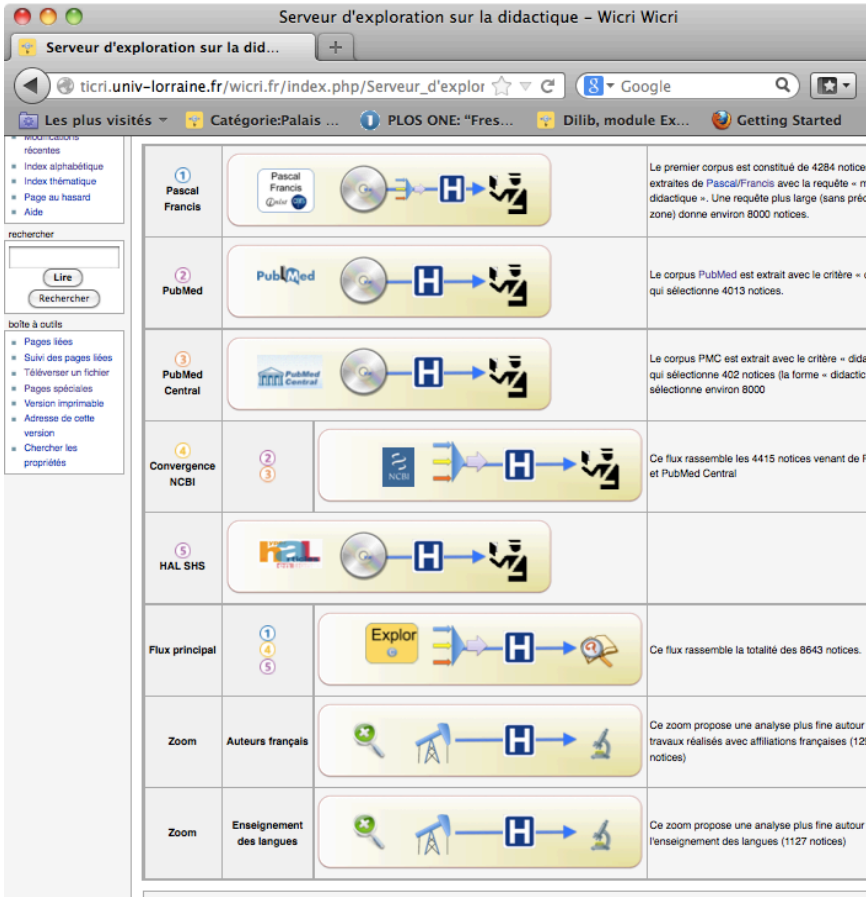
365 Lees A
325 Lang A
303 Louis E
279 Poewe W
251 Bhatia K
248 Quinn N
218 Goetz C
216 Jankovic J

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

Data curation



- First samples: country names in an heterogeneous context



Data curation – countries

- Pascal: ISO codes (exemple Pascal)
- Wikipedia page adaptation (towards sematic Web)

```
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*

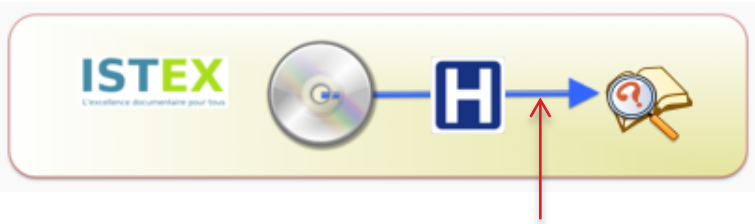
Data curation (countries) Address analysis from wiki tables

(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

Data curation: by region level



These maps could ba available on each servers

Data curation: region level

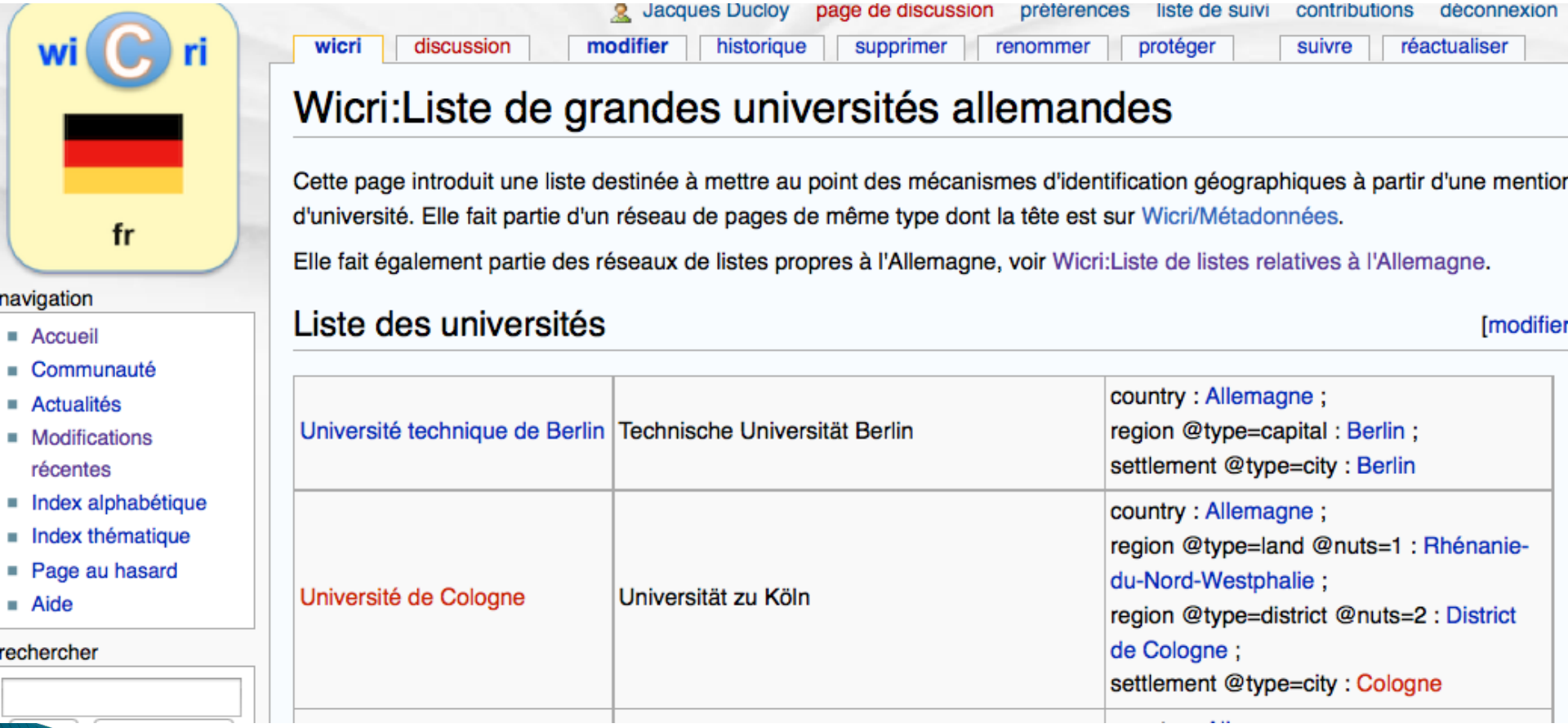
Wicri/Allemagne (Wicri/Germany)

ville	code 4 chiffres	code 5 chiffres	formes courantes	district/land	N
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				
Bayreuth	W-8580				
Berlin	W-1000				
Bonn	W-5300				

```
<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>
```

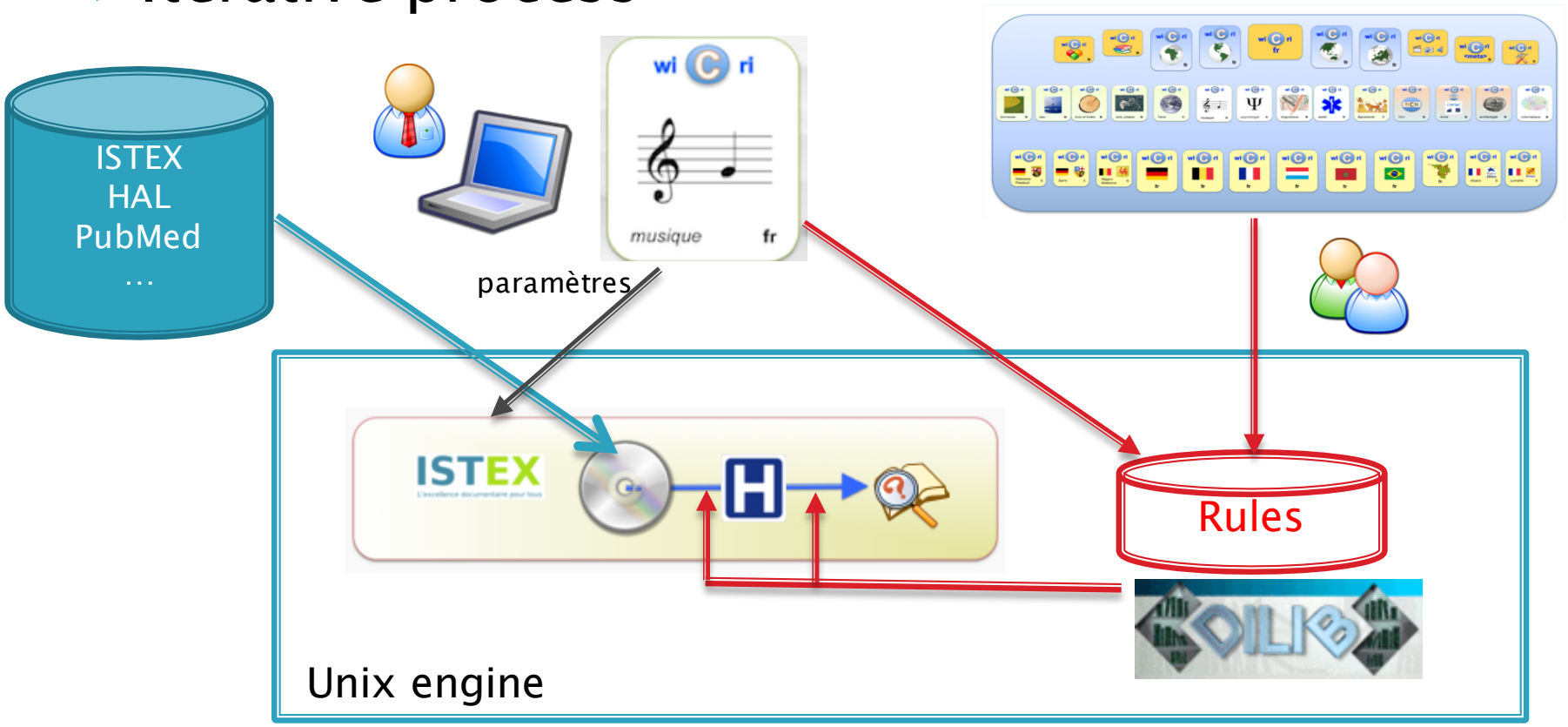
Result:
Sxml rules

Data curation: research & tools

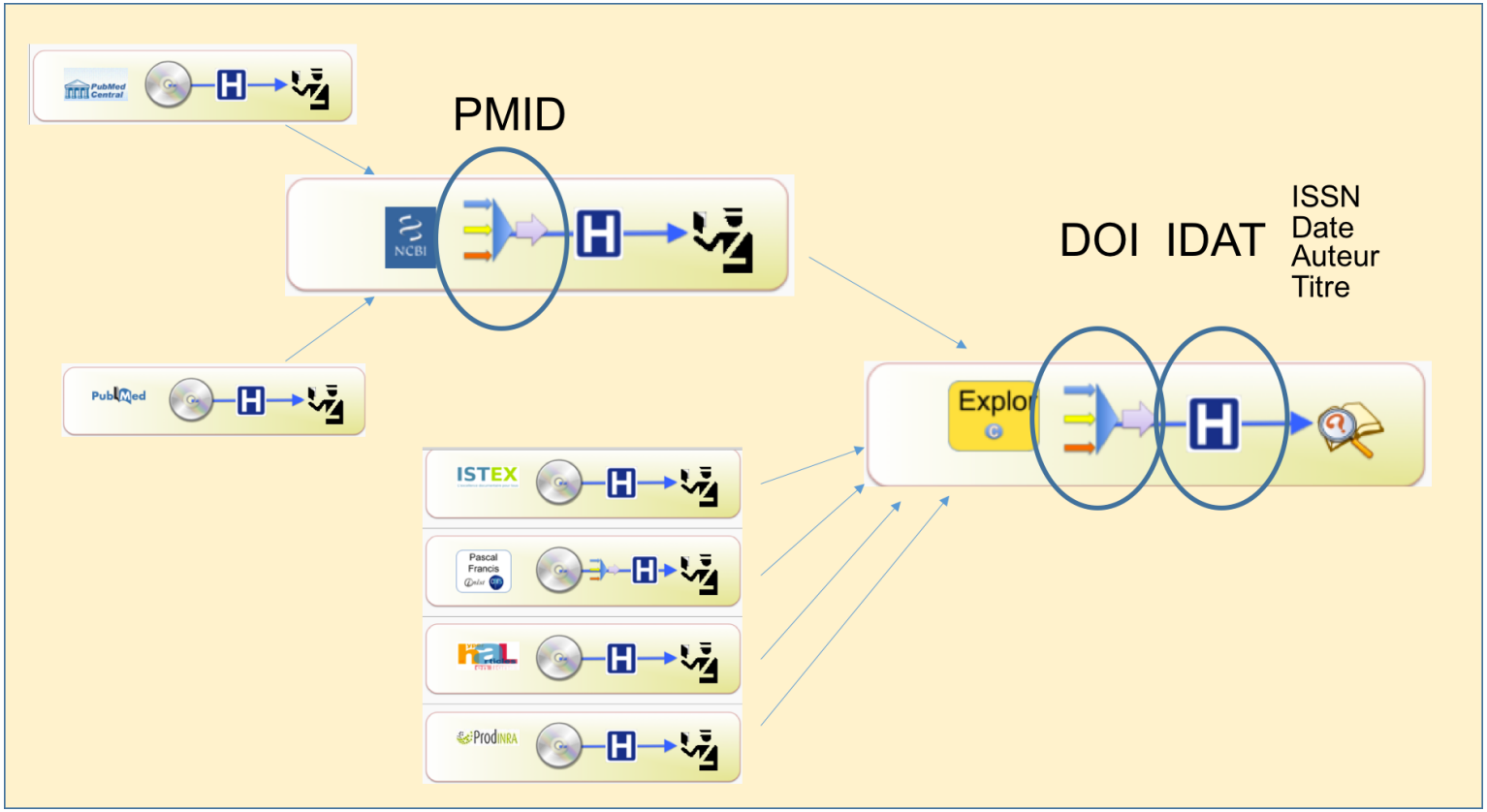


Area generation

- Iterative process



Pairs solving
ISTEX / Pascal / Hal / MEDLINE…



Filtering text

Heuristics, an example:

- *What are the most frequently quoted Mozart works in a corpus ?*
- General idea: use the Köchel catalog like: *Sonata KV. 448*
- Regular expressions for KV…

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

```
KV.223 000000
KV.125 000000
KV.448 000001
KV.448 000001
KV.025 000002
KV.448 000003
….
```

```
…
KV.448 000001
KV.448 000001
KV.448 000003
….
```

```
…
<index><k>KV.447</k><n>1</n><l>…
<index><k>KV.448</k><n>500</n><l>…
<index><k>KV.449</k><n>30</n><l>…
….
```

Filtering text: binomial names

- Same tools within different contexts
- On Wicri/Water: binomial names
  - Filtering papers about european fishes:
    - « *perca fluviatilis* » « *lota lota* »…
- On Wicri/Music and Wicri/Psychlogy
- American Journal of Dance Therapy
  - *What articles have an artistic point of view?*
    - Wikipedia => List of American choreographers
    - Filtering by « first name » « last name »

A key point with full text corpus: Curation,

- Fisrt sample, a corpus about Mozart:
- Humanities journals + medical ones:
  - Humanities: authors without affiliations
  - Medicine: authors with strong affiliations
- Global statistical results deal with medecine!
  - And not: Music!!!
- How to locate items where:
  - the « Mozart » programming environment
  - is used to process music data?

OCR: Curation, curation

- A test on « Scrum approach »
- 9,000 documents dealing with Scrum
- But 8,000 with OCR errors:
  - sérum or serum -> scrum

Information science: Curation, curation, curation

- *Looking for open science in Belgium*
- 4,000 documents for:
  - "open access" AND belg*
  - Only between 100 and 200 dealing strongly with open access.

  - Among the others:
  - Title:The EADGENE Microarray Data Analysis Workshop (Open Access publication)

Conclusion

- **Curation**
- Curation
- Curation
- Curation
- Curation
- Curation
- Thank you !