LorExplor au séminaire technique ISTEX

Jacques Ducloy
Retraité CNRS (INIST, LORIA, DRRT…)
Consultant pour l'Université de Lorraine

Plan

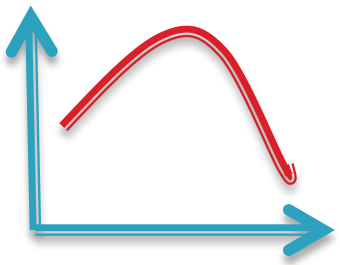
- Introduction : ISTEX, un devoir d'ambition
- Wicri/LorExplor : démonstrateur d'une cyberinfrastructure de la connaissance
- Bibliothèque XML DILIB (lien API ISTEX)
- Wikis sémantiques et curation de données
- Conclusion : Apprendre le numérique en construisant

ISTEX, un devoir d'ambition

- Budget : 60.000.000 € dans une situation de crise et de réduction de moyens
  - Un devoir d'ambition…
  - Quelles retombées pour le contribuable ?
- Construire le socle de la bibliothèque scientifique numérique, mais à couverture nationale
  - La recherche est internationale
  - Concilier national et international
- Des dizaines de millions d'articles en texte intégral
  - Pour quel usage ?
    - Que peut faire un acteur de la recherche ou de la culture avec 20000 documents en XML ?
  - L'ensemble de l'ESR est concerné
    - Et pas seulement les équipes de recherche en TAL

Humanités numériques et IST en France : TOP /Crise

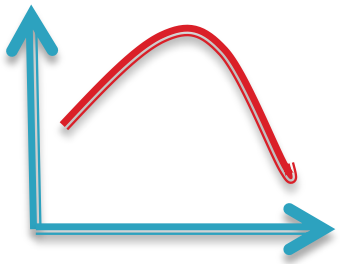
- De 1950 à 1975 dans le top 4 mondial avec :
  - TLF, Pascal, Questel, Cyclades, Mistral…
- Arrêt R&D, concepteurs non remplacés
  - Passage Gamma 60 vers Iris 80 sans réingénierie
  - Sous-traitance de l'ingénierie (Jouve, Questel)
- La rentabilité prime sur les missions initiales
  - Chicago TLF gratuit US, payant en France
  - RefDoc et la bulle Internet
    - 1987 => 1000 commandes par jour
    - 2000 => 3000 commandes par jour
    - 2015 => 50 à 100 commandes par jour
    - 2014 arrêt de la chaine Pascal Francis



Humanités numériques et IST en France : Crise / Reprise

- Dans la crise des signes positifs
  - TLF : Frantext
    - CDST infométrie, numérisation
    - 87 – 91 INIST
      - Top 4 en FDP, Pascal fait « jeu égal » avec Medline
      - Nathalie Dusoulier dans le réseau normatif LC + NLM…
      - pionnier sur SGML appliqué aux formats MARC/ISO 2709
- Reprise à l'ATILF
  - Synergie recherche + services + infomatique + linguistique
- INIST 2014, un espoir :
  - Redémarrage de la R&D big data avec ISTEX…

Priorités ISTEX / Devoir d'ambition



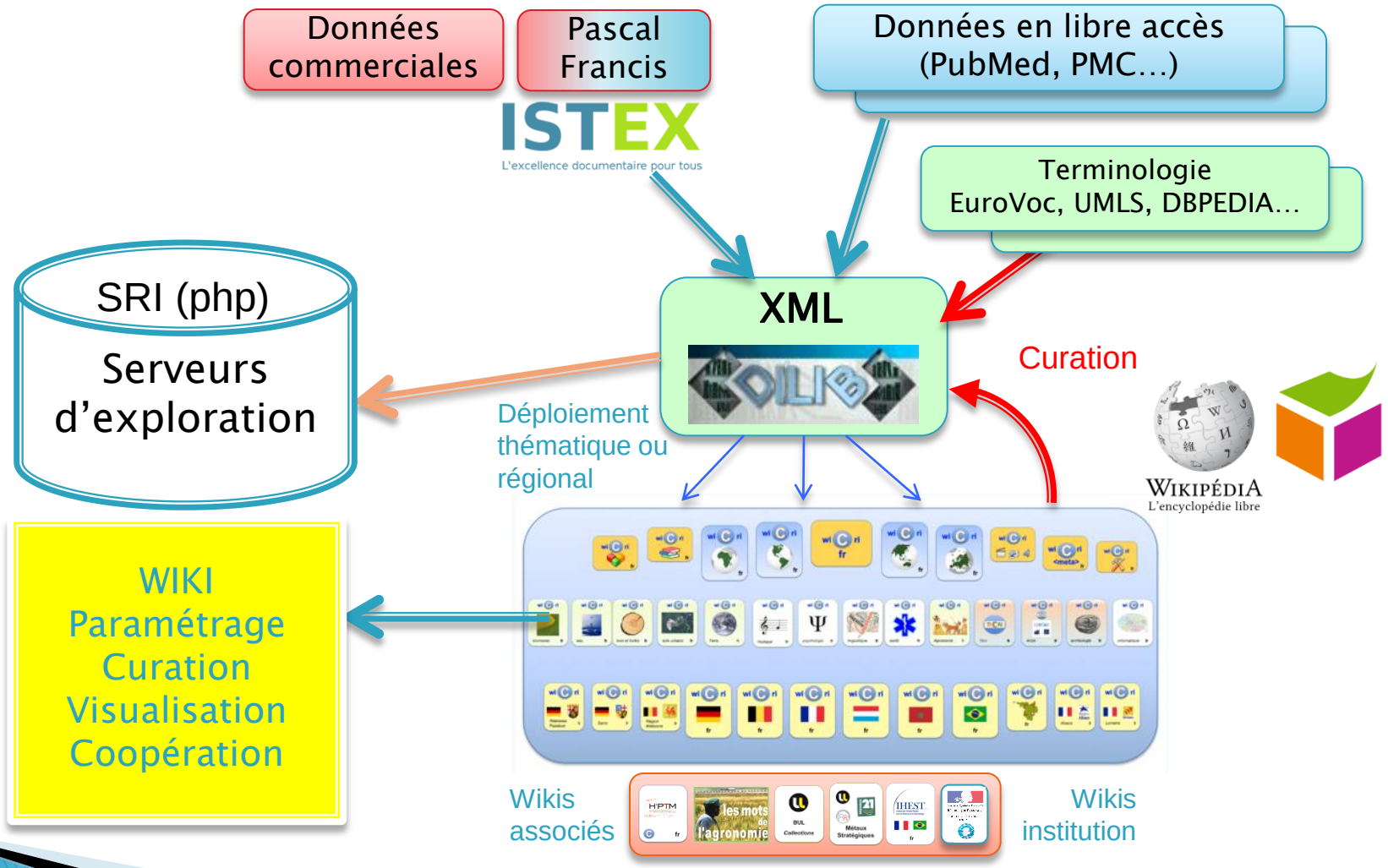
- Accès au document ???
  - Avec RefDoc/
  - 50 à 100 demandes/jour

- Pascal + Francis (300 personnes sur 50 ans)
  - 300 * 50 * 100.000 € = **1.5 milliard €**
- ISTEX peut-il favoriser l'émergence d'un grand projet concurrentiel au niveau mondial ?
  - A partir des besoins stratégiques qui avaient motivé Pascal Francis ?

Wicri/LorExplor

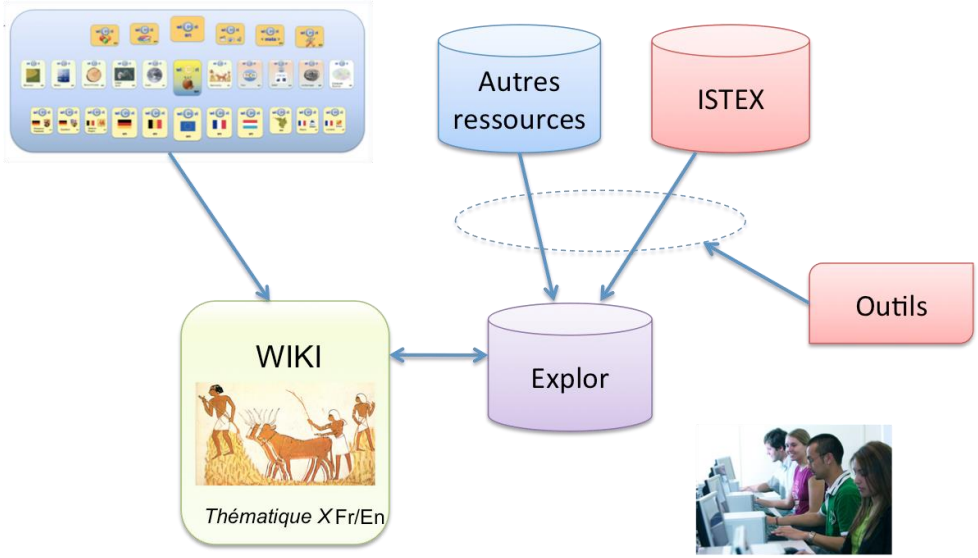
- Un démonstrateur
  - d'une cyberinfrastructure
  - de la connaissance scientifique, technique ou culturelle
- Initié par des besoins de valorisation de la recherche (ANL, DRRT)
- Dopé par ISTEX
- Inspiré par les réseaux, le génie logiciel, l'interopérabilité, et pratiques coopératives

Infrastructure LorExplor



Pratiques LorExplor

- Construire de la connaissance structurée, sémantique par des explorations de corpus
- Sensibilisation, formation, appropriation, construction collective



Le Réseau Wicri

Wikis de service

Wikis thématiques

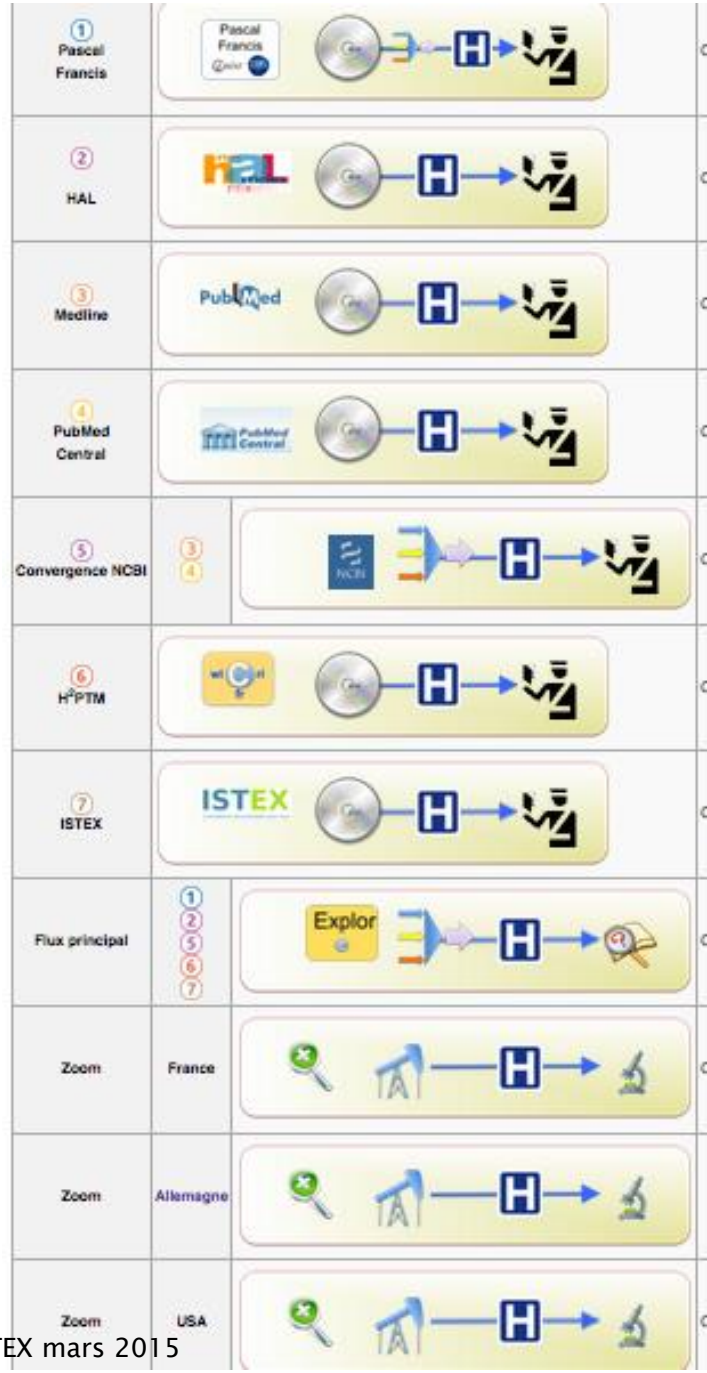
Wikis régionaux

Wikis associés

Wikis institutionnels

Gérer l'hétérogénéité

- Parser XML en KIT
- Moteur de recherche en KIT
- Interfaces outils
- Interfaces données
- Générateur de serveurs d'exploration



Dilib, historique

- Antériorité :
  - TLF, bande magnétique = flux, performances
    - Mistral Système de recherche avec ontologie
  - ANL = Unix génie éditorial, génie Logiciel, IA
  - Geac = système ISO 2709
- Ilib, INIST 91,
  - Prototype spécialisé (fichiers Marc codés en SGML / lex)
  - Normalisation approximative (non XML)
  - 20 ans de retombées : MIRIAD, Stanalyst
- Dilib V0.1, Loria 93
  - Préfiguration DOM (Sgml bien formé sans DTD)
  - Bibliothèque de composants pour infométrie
- Dilib V0.2, Loria 98 -> Inist 2003
  - Cohabitation SGML, XML ; interfaces cgi
  - Projets MedExplore, Biban, prototype Servist
- Dilib V0.5, UL 2013
  - Sxml + PHP + couplage Semantic MediaWiki + UTF8

Conventions Sxml

- Objectif : gérer sous Unix des flux Xml éventuellement volumineux
- Principe :1 document XML = 1 ligne Unix
  - grep, sort, cat… deviennent des outils XML
- Sauts de ligne et tabulations interdits
  - Palliatif entités &#15; …

Exemple d'emploi

Fichier inverse pays (AffPays) du serveur d'exploration sur l'hypertexte

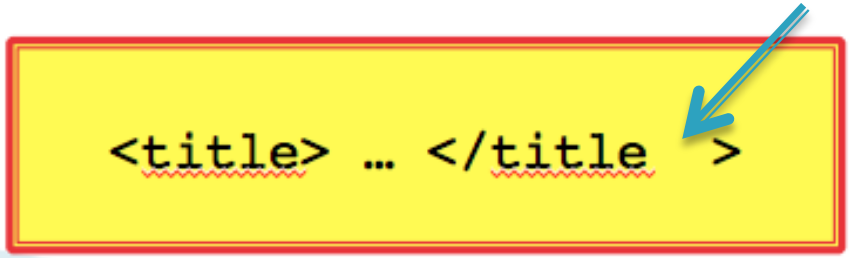
```
000000   <idx><kw>Afrique du Sud</kw><f>23</f><l><e>000345</e>…
000001    <idx><kw>Algérie</kw><f>5</f><l><e>001714</e><e>001715</e>
```

```
HfdCat Ticri/H2ptm/corpus/HypertextV5/Data/Main/Exploration/AffPays.i.hfd \
    |    SxmlSelect -g idx/f/1 -p @g1 -g idx/kw/1 -p @g2          \
    |    sort -rn
```

```
1823    États-Unis
916     Royaume-Uni
850     Allemagne
821     France
519     Italie
333     Espagne
310     Canada
293     Japon
266     République populaire de Chine
```

Parser Xml, les défis ISTEX

- Ilib : Sgml pour 1 type de document (ISO 2709)
- Dilib 0.1 : Xml à partir de métadonnées structurées non Xml
- Dilib 0.2 : métadonnées Xml de provenance diversifiée
- Dilib 0.4 : 1 DTD document (JATS-NLM)
- Dilib 0.5 : ISTEX
  - Manipuler des gros flux de documents avec des multiples DTD avec toutes les variantes syntaxiques



Parser Xml

- Approches diversifiée
  - Batch contexte « big data » : langage C
    - BufferParserXml repère des éléments XML.
      - Cohabitation possible avec d'autres parser (libxml)
    - SxmlParser analyse une chaine contenant un objet XML
  - Web : utilisation du parser PHP (DOM)
    - Interface avec le moteur de recherche (HFD)
  - Corpus moyens : solution Python à l'étude
- Extensions possibles :
  - SxmlContainer :
    - nombre réel, xpath,
    - table,
    - formule (mesure)

```
<place>
  <placeName>Meuse (département)</placeName>
  <dilib:geoPath>
    <path>
      <M>
        <x>
          <![double[5.27192]]>
        </x>
        <y>
          <![double[49.53714]]>
        </y>
      </M>
```

TEI vs JATS

- TEI : Text Encoding Intitiative
  - Référence pour les humanités numériques
- JATS : **Journal Publishing Tag Suite (NLM)**
  - **Référence pour bio-médical**
- Corpus vieillissement
  - 6000 doc sur 9000 utilisent JATS
  - DILIB traite JATS via PubMed Central
- Proposition / réflexion :
  - cohabitation
    - TEI (ex Wicri/Linguistique)
    - JATS (ex Wicri/Santé)
  - Enrichissement d'autant meilleur qu'il est spécialisé

Organisation HFD

- Hierachic Organisation for Documents
- 1000000 docs =
  - 100 répertoires
    - De 100 fichiers
      - De 100 documents (Sxml)
- Fichiers inverses
- Outils

```
HfdIndexSelect –k France –h AffPays.i
```

- Contraintes ISTEX
  - Dépasser le million de métadonnées: FFFF99
  - Et à la fois 1 document full text par fichier

Interfaces logiciel

- Exemple MediaWiki

Laboratoire des sciences du génie chimique	CNRS : UPR6811 ; Laboratoire des sciences du génie chimique	country : France ; region @type=region @nuts=2 : Lorraine ; settlement @type=city : Nancy ; orgName @type=institution : Centre national de la recherche scientifique

```
WicriGetPage -l wicri-france.fr -p "Wicri:Liste d'unités propres du CNRS" \
   | WicriTableOrgFromWiki| SxmlIndent
```

```
CNRS : UPR6811
------ 1
<org>
  <orgName>Laboratoire des sciences du génie chimique</orgName>
  <country>France</country>
  <orgName type="institution">Centre national de la recherche scientifique</orgName>
  <placeName>
    <settlement type="city">Nancy</settlement>
    <region type="region" nuts="2">Lorraine</region>
  </placeName>
</org>
```

Interface logiciel : API Istex

```
IstexGetCorpus –s1-q …
```

- Estimer la taille d'un corpus

```
IstexFlashCorpus –s20 -q …
```

- Première idée du contenu

```
IstexExplorCorpus –s 1000 -q …
```

- Construction d'un serveur d'exploration de base

```
IstexExplorCorpus –s 4000 -q …
```

ıramétrage

- Contraintes :
  - zéro défauts au niveau parsing
  - Chaque éditeur amène des problèmes spécifiques
    - Exemple : Texte XML en métadonnées dans Elsevier
      - Plusieurs semaines (mois?) d'adaptation pour usage courant

Générateur de serveur d'exploration

- Procédure à améliorer :
  - Le wiki contient des tables de paramètres téléchargées sur un ordinateur de développement
  - Le serveur est généré en 2 parties :
    - Données : un ensemble de HFD
    - IHM : des pages PHP générées à partir des paramètres
  - Transfert de fichiers dans le file system des wikis
- Procédure simplifiée IstexExplorCorpus

list#listIndexes	Istex/Exploration	RBID.i ; Author.i ; AffOrg.i ; AffPays.i ; AffRegion.i ; AffVille.i ; AffRegInc.i ; Wicri.i ; KwdEn.i ; Title.i
list#listAssoc	Istex/Exploration	KwdEn.a ; Author.a
list#listCluster	Istex/Exploration	Author.cf ; Author.cc ; KwdEn.cc ; KwdEn.cf

Interface données : ex. Pascal

Reformatage ;
- Pascal/ilib -> Xml Sxml
- Serveur -> Sxml
- Entités -> UTF 8 (pb XML)
- Dé-doublonnage Francis

En cours d'analyse :
- Format convergent (ISO 2709)
- Mise à jour possible sur wiki
- Enrichissement MESH ?

Remarques Api ISTEX

- Sommes nous dans le Monde de Mickey ?
  - Tout nouveau flux demande adaptation
    - Elle peut être conséquente
  - Les services en lignes évoluent (HAL, NLM)
  - NLM plus de 20 ans de r&D conséquente
    - Pendant 20 ans l'INIST a sous-traité…
  - Que devient le site ISTEX le jour du départ des CDD ?
  - Sans support logistique ni transfert de compétence, Dilib s'arrête du jour au lendemain…
- Par rapport à la qualité des données éditeur
  - Un tiens vaut mieux que deux tu l'auras (cf Pascal)
  - L'ingénierie de la connaissance repose sur l'art de traiter des données pourries…
  - Arrêtons, en France, de vouloir appliquer à la connaissance les méthodologies de la gestion administrative où toutes les données doivent être validées ab initio.

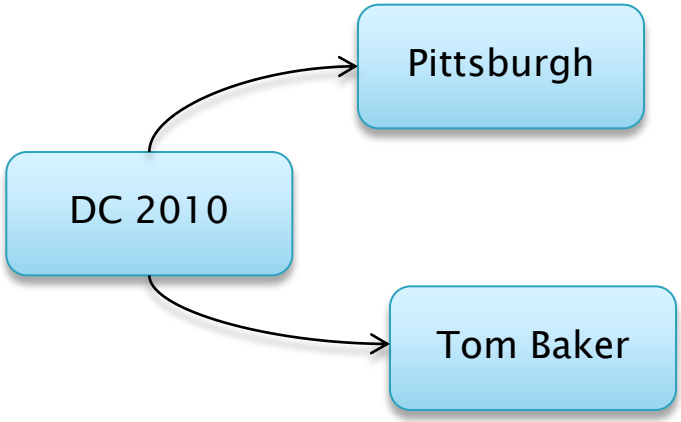
Curation des données, aspects termilogiques

- 2 mots sur SMW (Semantic MediaWiki)
- Exemples de curation sur données géographiques
- Réflexions sur la formation

Liens simples/ sémantiques

Liens simples

```
DC 2010 takes place in [[Pittsburgh]]
==Program Committee==
* [[Thomas Baker]]
```



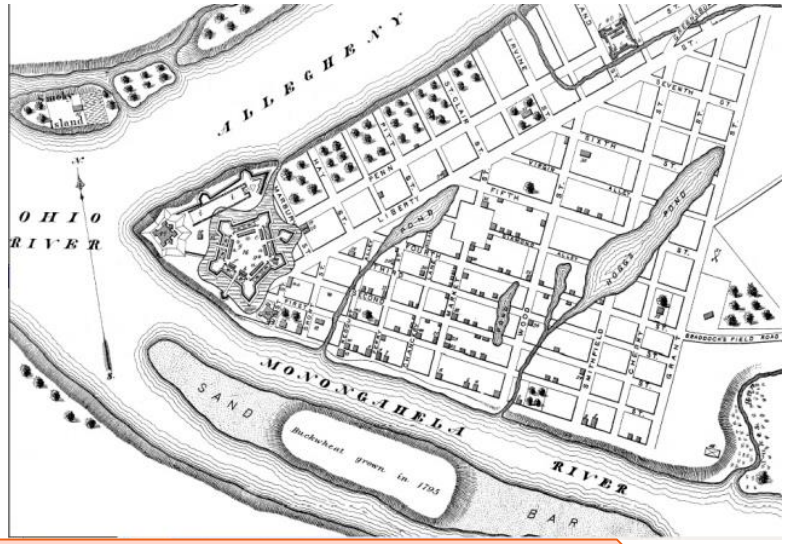
- A chaque lien on associe
  - + un attribut (property)

- Extension Semantic MediaWiki

```
DC 2010 takes place in [[Has location city::Pittsburgh]]
==Program Committee==
* [[Has PC member::Thomas Baker]]
```

Liens sémantiques

Pittsburgh est située au confluent des rivières Allegheny et Monongahela qui forment l'Ohio, un affluent du Mississippi .



Pittsburgh est située au confluent des rivières [[sur le cours d'eau::Allegheny (rivière)|Allegheny]] et [[sur le cours d'eau::Monongahela]] qui forment l'[[sur le cours d'eau::Ohio (rivière)|Ohio]], un affluent du [[Mississippi (fleuve)|Mississippi]] .

Faits relatifs à Pittsburgh — **Recherche de pages similaires avec** + .

Voir comme RDF

Sur le cours d'eau Allegheny (rivière) +, Monongahela + et Ohio (rivière) +

Utilisation des liens sémantiques

- Naviguer sur une propriété

Pages utilisant l'attribut « Est un affluent de »

Afficher les 13 pages utilisant cette propriété.

A

Allegheny (rivière) + Ohio (rivière) +

D

Durbion + Moselle (rivière) +

E

Euron + Moselle (rivière) +

F

Fensch + Moselle (rivière) +

M

Meurthe + Moselle (rivière) +
Monongahela + Ohio (rivière) +
Moselle (rivière) + Rhin +

O

Ohio (rivière) + Mississippi (fleuve) +
Orne (Moselle) + Moselle (rivière) +

Utilisation des liens

- Poser des requêtes sémantiques

```
{{#ask:[[est un affluent::{{Ohio (rivière)]]
| format=ul
| sep=,_
| intro=Rivières citées sur Wicri Eau :_ }}
```

L'Ohio

==Les affluents de l'Ohio==

("liste calculée")

{{#ask:[[est un affluent::{{PAGENAME}}]]
| format=ul
| sep=,_
| intro=Rivières citées sur Wicri Eau :_ }}

==Les villes traversées par l'Ohio==

("liste calculée")

{{#ask:[[sur le cours d'eau::{{PAGENAME}}]]
| format=ul
| sep=,_
| intro=Villes citées sur Wicri Eau :_ }}

Exemple H²PTM 2011

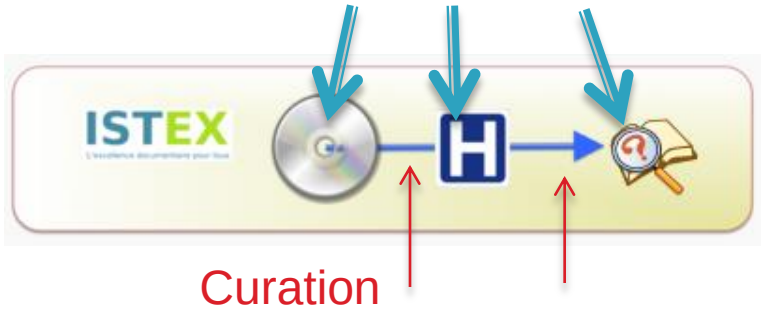
Faits relatifs à H2PTM 2011 Metz — Recherche de pages similaires avec + . Voir comme RDF

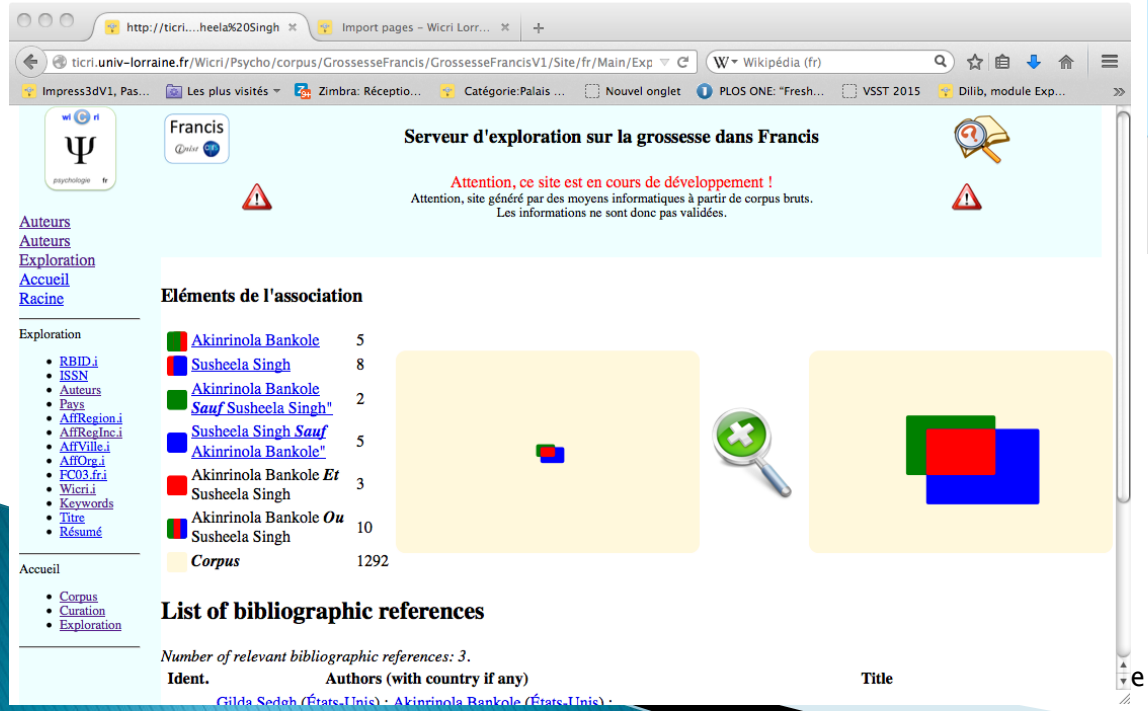
A pour affiliation de président de comité de programme	Université Paris 8 +, Université Paul Verlaine - Metz +, Université de Valenciennes et du Hainaut-Cambrésis +, CELSA + et Université Paris 1 Panthéon-Sorbonne +
A pour conférencier invité	Fabien Granjon + et Luc Massou +
A pour intervenant	Imad Saleh +, Luc Massou +, Sylvie Leleu-Merviel +, Yves Jeanneret +, Chiara Giaccardi +, Béatrice Drot-Delange +, Laurent Collet +, Aurélie Brouwers +, Nour El Mawas +, Jean-Pierre Cahier +, Aurélien Benel +, Sarah Labelle +, Olivier Mauco +, Alban Gregoire +, Aude Seurrat +, Luc Ploquin +, Fanny Georges +, Nicolas Auray +, Justine Simon +, Ève Ross +, Marc Veyrat +, Alexandra Saemmer +, Philippe Bootz +, Pierre Morelli +, Brigitte Simonnot +, Fredj Zamit +, Sébastien Hock-Koon +, Nam-Jun Pyun +, Khaldoun Zreik + et Manuel Zacklad +
A pour membre du comité de programme	Ghislaine Azemard +, Jean-Pierre Balpe +, Claude Baltz +, Belhassen Baddredine +, Didier Baltazart +, Ali Ben Cherif +, Mokhtar Ben Henda +, Serge Bouchardon +, Philippe Bootz +, Hafida Boulekbache-Mazouz +, Eric Brangier +, Christian Bastien +, Eric Bruillard +, Marie Chagnoux +, Jean Clément +, Tristan Cazenave +, François Denieul +, Giovanni De Paoli +, Jacques Ducloy +, Aude Dufresne +, Raja Fenniche +, Sébastien Genvo +, Gino Gramaccia +, Bernard Idelson +, Madjid Ihadjadene +, Bernard Jacquemin +, Christian Jacquemin +, Catherine Kellner +, Christophe Kolski +, Jean-Marc Labat +, Michel Labour +, Pierre Levy +, Charles Max +, Arnaud Mercier +, Abderrazek Mkadmi +, Yves Mineur +, Jocelyne Nanard +, Marc Nanard +, Stéphane Natkin +, Sophie Pene +, Pierre Quettier +, Vincent Quint +, Pierre Rabardel +, Jean-Hugues Réty +, Everardo Reyes +, Estrella Rojas +, Eve Ross +, Ioan Roxin +, Alexandra Saemmer +, Emmanuel Sander +, Mohammed Sidir +, Basel Solaiman +, Chantal Soulé-Dupuy +, André Tricot +, Saïd Tazi +, Brigitte Trousse +, Christoph Vatter +, Geneviève Vidal +, Gilles Venturini +, Thilo Von Pape +, Jacques Walter +, Roberto Willrich + et Khaldoun Zreik +
A pour organisateur	Université Paul Verlaine - Metz +, Paragraphe (laboratoire) +, Université Paris 8 +, Université de Cergy-Pontoise +, Centre de recherche sur les médiations +, Université Nancy 2 +, Université de Haute-Alsace +, Laboratoire des sciences de la communication, DeVisu +, Université de Valenciennes et du Hainaut-Cambrésis +, Groupe de recherches interdisciplinaires sur les processus d'information et de communication +, CELSA + et Université Paris 1 Panthéon-Sorbonne +
A pour pays	France +
A pour président de session	Sébastien Genvo +

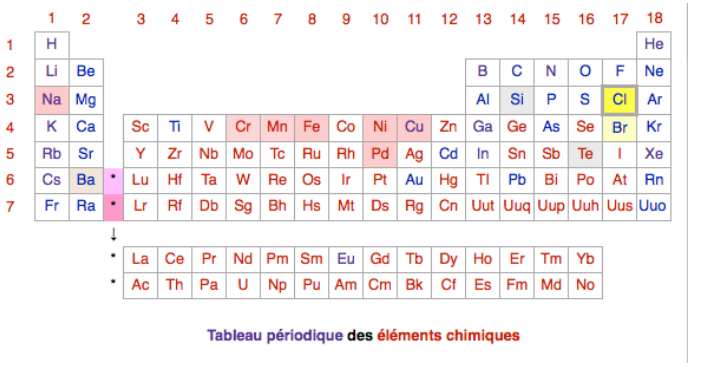
76 propriétés, 5400 valeurs de propriétés….

LorExplor – Serveur d'exploration

Système d'information orienté exploration



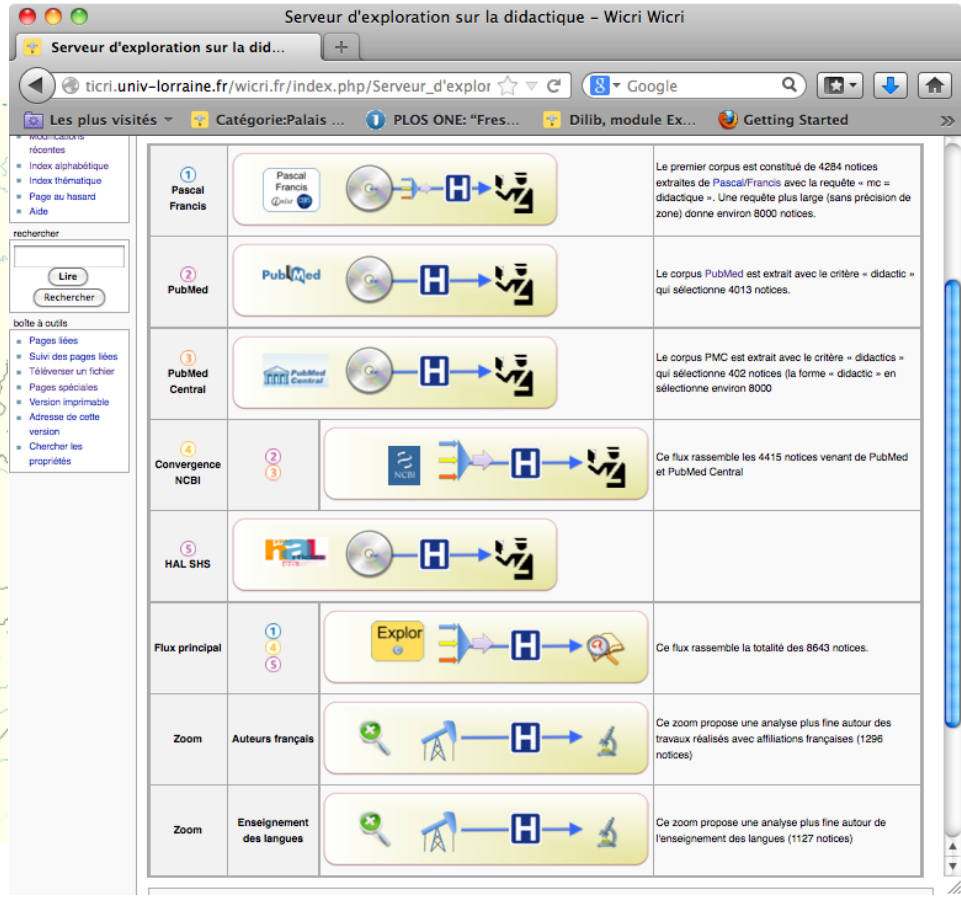




e ISTEX mars 2015

Curation des données

- Exemple : identifier les pays dans un contexte hétérogène



Curation des données – pays

- Codes ISO (exemple Pascal)

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:Tri1ALGÉ
276	DEU	DE	Allemagne	ALLEMAGNE
020	AND	AD	Andorre	ANDORRE
024	AGO	AO	Angola	ANGOLA
660	AIA	AI	*Anguilla*	*ANGUILLA*

```
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 imaging, and
                      digital typography : Saint Malo, 30 March - 3 April
                      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 Wissensverarbeitung,
                      Universität Hannover, Lange Laube 3 @2 30159 Hannover
                      @3 DEU @z 1 aut. @z 2 aut. @z 3 aut.
```

Curation des pays – Adresses

Adresses postales
(Springer, PubMed)

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

```
<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>
 <namePart type="family">Keller</namePart>
 <role>
  <roleTerm type="text">author</roleTerm>
 </role>
 <description>Matthias.keller@kit.edu</description>
 <affiliation>Steinbuch Centre for Computing (SCC),
 Karlsruhe Institute of Technology (KIT), D-76128,
 Karlsruhe, Germany</affiliation>
</name>
```

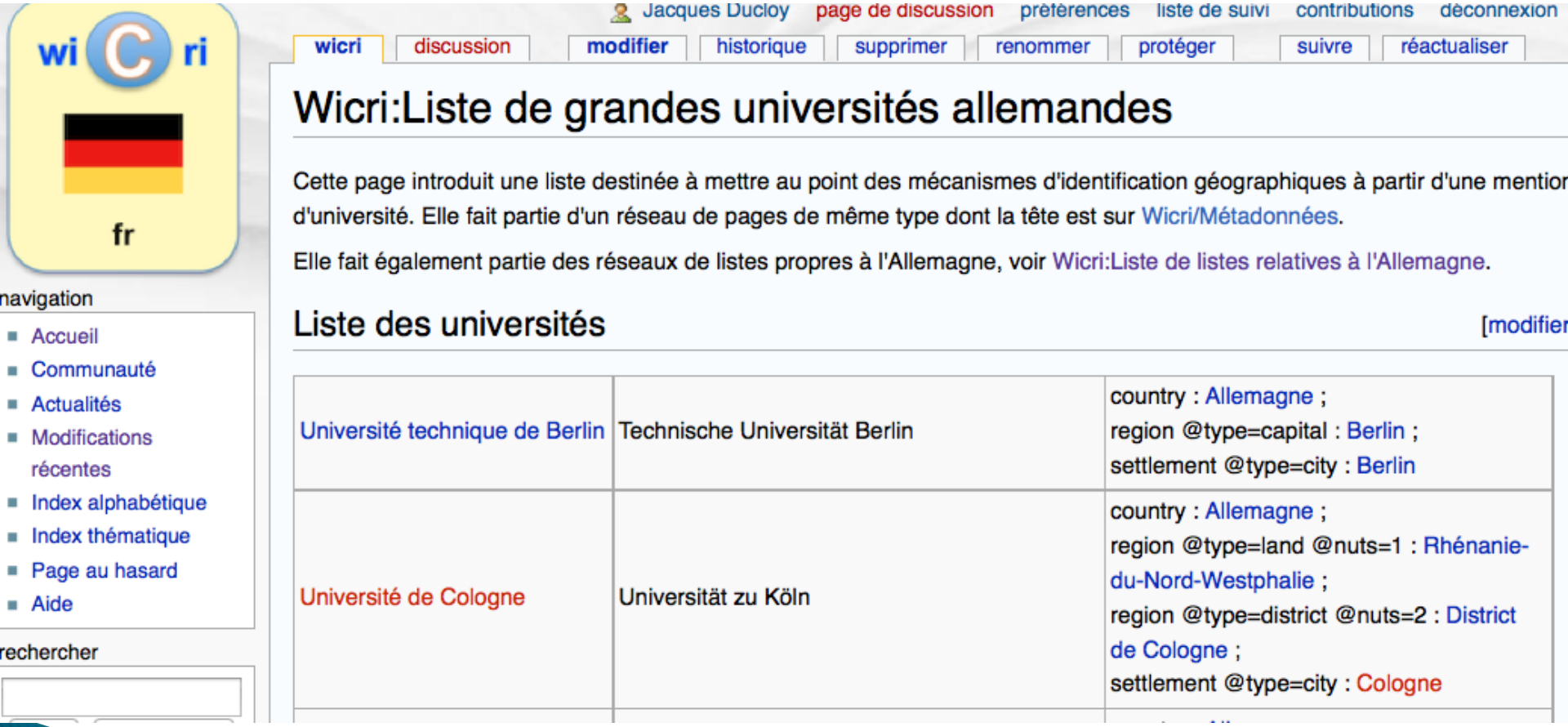
Curation des régions

Curation des régions

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	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 Haute-Franconie	
Berlin	W-1000	10115			
Bonn	W-5300	53111			

```
<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>
```

Curation des régions



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

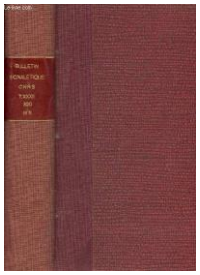
rechercher

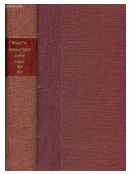
Enrichissement thématique

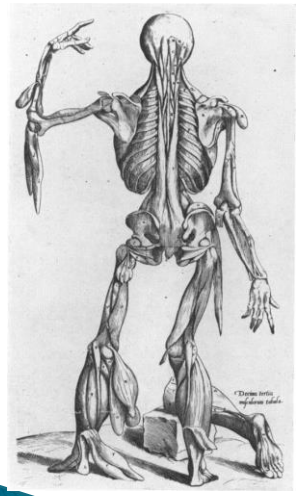
- Objectif à moyen terme :
  - Jointure ISTEX / Pascal / Francis / PubMed
  - Via les références (PMID parfois dans ISTEX)
- Curation Pascal/Francis
  - Sur la base ISO 2509 complétée
  - Mise à jour sur le réseau de wiki
    - Wiki de référence associé à un documents
    - Cohérence par robots et modèles.

Pascal Francis – mission initiale
Accéder à l'essentiel de la science

1970 1988 2000 2015







Quelques références Wicri/LorExplor

- Dublin Core…
  - DC 2010 Pittsburgh 40.000 visites
  - Article en anglais 35.000 visites
- H2PTM avec Paris 8, CREM…
  - Actes H2PTM (environ 70 -> 300 articles)
  - Observatoire des recherches sur l'hypertexte (wikis sémantiques)
  - Terminologie, bibliographies , 2000 -> 10000 termes 5000 ->20000 relations
  - Serveur d'exploration 10.000 -> 20.000)
  - Bouquet envisageable : CIDE, VSST, ISKO + revue…
- IHEST : wiki France Brésil / observatoire / exploration
- Humanités numériques sur Nancy
  - (Chanson de Roland, chartes…)
- Les mots de l'agronomie (INRA°
- Ouverture Grande Région
- Matériaux…
- TP Master Université Lorraine, Paris 8

Apport des formations (TP)

- Chaque étudiant choisit son sujet pour lequel
  - Il teste des requêtes
  - Construit un serveur d'exploration
  - Peut améliorer sa requête
  - Analyse les acteurs connus et inconnus
  - Introduit des éléments de curation
- Très formateur pour les étudiants
  - Donne du recul sur la formulation d'une requête
- Extrêmement riche pour LorExplor… ISTEX
  - Variété des thématiques
  - Emergence de problèmes (API… Dilib…)
  - Observation des pratiques

Problèmes soulevés par les TP

- Retour à la réalité, 1 session de TP =
  - Un enseignant ultra motivé
  - 2 intervenants LorExplor pour 10 à 15 étudiants
  - Des mois/semaines de développement pour corriger les problèmes mis en évidence.
- Problèmes administratifs
  - Rémunérer un étudiant pendant 2 hommes mois…
- Problèmes de logistique
  - 1 session = des jours (semaines) de préparation
  - Blocages multiples (accès, pare-feux, DSI..)
  - Ensembles logiciels instables pour des années !
- Solution partielles proposées (TP ou sensibilisation)
  - Espace d'accueil pré installé (ex INIST kiosque)
  - Machine virtuelle pré installée

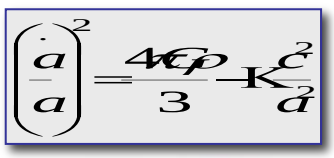
Sensibilisation / formation

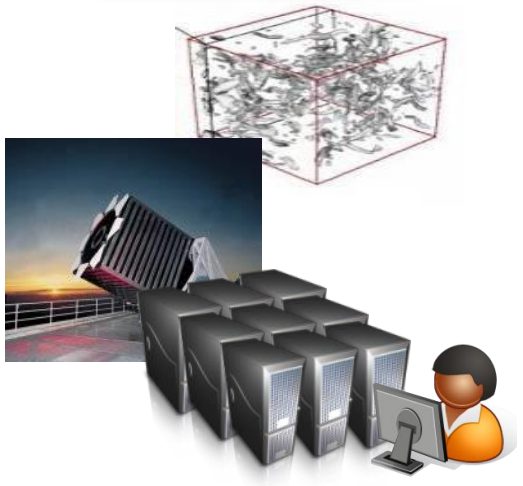
- LorExplor : Panorama potentiellement complet
  - Culture scientifique et technique
  - Edition numérique actuelle et ancienne
  - Terminologie
  - Exploration de corpus de métadonnées
  - Passage au texte intégral
  - Ressources
    - Pascal (1 milliard €), ISTEX (50 millions)
    - + corpus et terminologies open access
- JD seul : sensibilisation puis, plus rien !
- Infrastructure pour sensibilisation / formation
  - Formation (les leçons des TP et de la Mutation Technologique)
  - Ensemble éditorial sur l'ingénierie de la connaissance
  - Banc d'essai pour expérimentations
  - Support minimal : environ 3 personnes (IR, IE, T)
    - 50% logistique,
    - 50% formation, expérimentation

Explosion d'un nouveau Paradigme

- Thousand years ago – **Experimental Science**
  - Description of natural phenomena
- Last few hundred years – **Theoretical Science**
  - Newton's Laws, Maxwell's Equations…
- Last few decades – **Computational Science**
  - Simulation of complex phenomena
- Today – **eScience or Data-centric Science**
  - Unify theory, experiment, and simulation
  - Using data exploration and data mining
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
  - ➢ Scientists over-whelmed with data
  - ➢ Computer Science and IT companies have technologies that will help

(With thanks to Jim Gray)





… Merci à Tony Hey (Microsoft Research)

Ingrédients des réussites concurrentes

- Très fort taux de r&D publique
  - Ex NCBI, OCLC, DCC etc
  - Démarrage autour 1990 (comparable INIST DRPN)
- Priorité aux réseaux d'universités ou aux acteurs de terrain (NSDL, iPlant, DPLA)
- Approche technologique plus proche du calcul scientifique que de la gestion
- Très haut degré d'interdisciplinarité
- Rupture dans les modèles d'organisation

Sensibiliser, former

- ISTEX : révélateur du gap à combler…
- LorExplor : démonstrateur propositionnel
- En France, requête type Google = niveau collège
  - Sensibiliser 100.000 Masters ?
  - Formation de base : 10.000 thésards ?
  - En ingénierie de la connaissance : 1000 M+D ?
- Modèle Wikipédia : apprentissage par construction collective de connaissance
  - Un thésard peut produire 2 ou 3 pages + 5 à 10 refs
- Extensible au niveau européen et francophone
- Un projet big data à construire !
  - 1000 familles de wikis sur 20 sites exploitant 100.000 de documents…