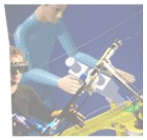


De 2 à 22 millions d'images; Création, Indexation et Recherche par le contenu

avec Piria

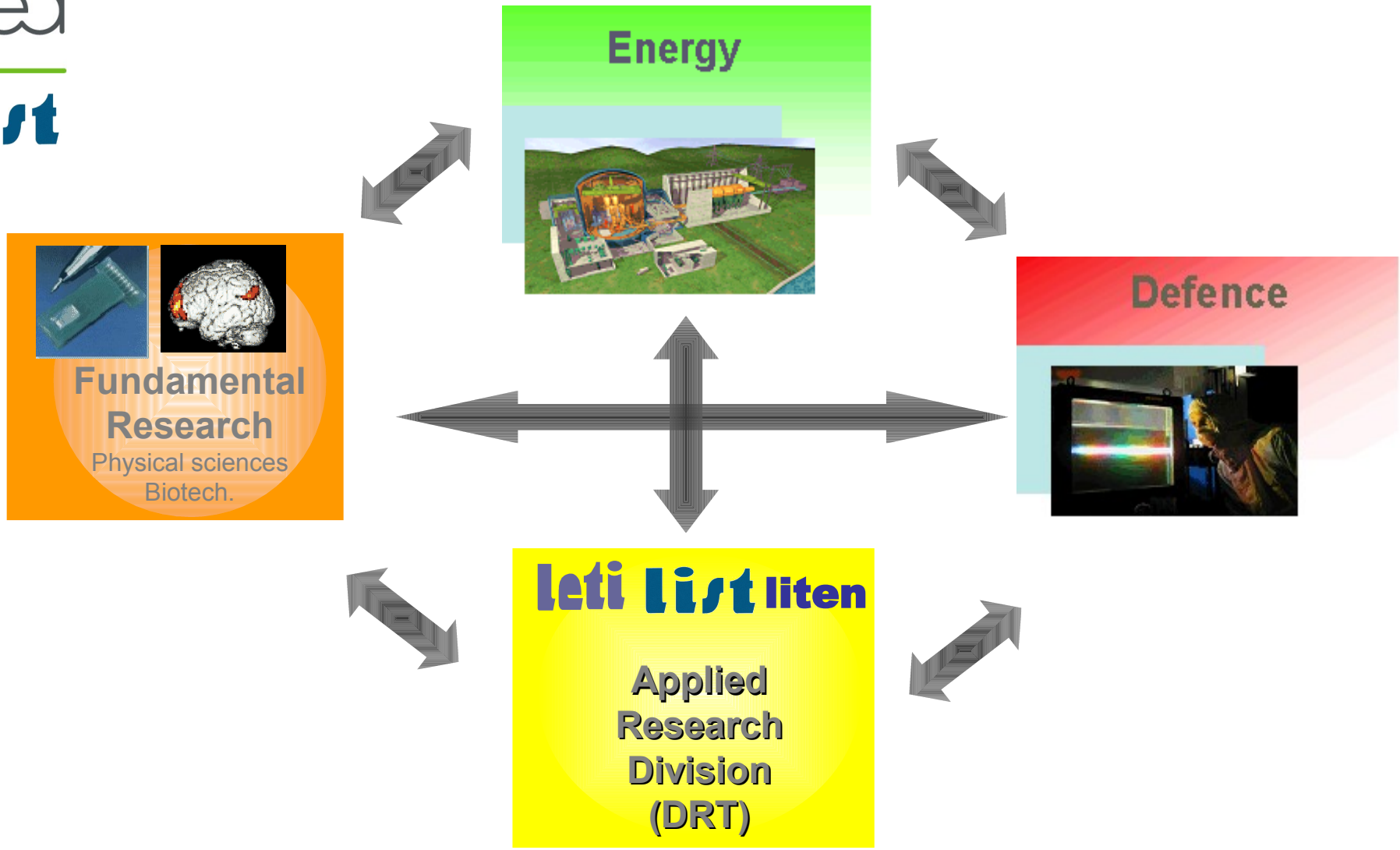
contact : patrick.hede@cea.fr

Commissariat à l'Energie Atomique
List

Plan

- Le CEA, le Lic2m
- Le moteur image Piria
- Le projet Fame II
  - ✓ Les partenaires, architecture BULL
- Le grand challenge image
- Travail réalisé
  - ✓ Le corpus (base d'images)
  - ✓ Piria MPI
    - Indexation
    - Recherche
- Performances
- Conclusions & perspectives

Le CEA: Commissariat à l'Energie Atomique



Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Création en 2002 (Christian Fluhr)

Héritage: moteur de recherche cross lingue Spirit, compétences sur le traitement linguistique et la recherche d'information scientifique et technique.

Héritage complémentaire de compétences sur l'analyse et l'indexation d'image

Effectif:

20 personnes, 13 permanents

Acteur majeur intégré au réseau de recherche française en TAL et en recherche d'information multimédia

CNRS, INRIA, Universités, Ecole de Mines, ENST

Engagement sur projets (MuSCLE, MEDAR,imageval,Fame2)

Transfert de technologie

Partenariat avec partenaires industriels sur projets collaboratifs (Bulll, Thales, Alcatel..)

A l'origine de la start-up NewPhenix

Activités de recherche

Image

descripteurs locaux, fusion de descripteurs, indexation, recherche

catégorisation, modèles 3D

Vidéo

structure, résumé

Texte

analyse sémantique

extraction d'information, catégorisation, résumé, filtrage

question réponse

Analyse et indexation document multimédia

moteur de recherche multimédia (fusion de descripteurs, fusion de résultats de comparaison)

analyse de la structure de documents multimédia

Extensions

constitution de ressources

passage à l'échelle: indexation de gros volume de données

amélioration de la transcription (syntaxe+sémantique)

constitution de ressources (pour la traduction: alignement de texte, dictionnaires bilingues, pour la sémantique: carte sémantique générale et métier)

Moteur indexation/recherche

- **CBIR Piria**
  - ✓ **Programme d'Indexation et de Recherche d'images par Affinité créé en 2002**
  - ✓ **Moteur développé au CEA LIST écrit en C++ utilisant les STL**
  - ✓ **Caractéristique réponse rapide,**

    convertit les signatures numériques des descripteurs en langage naturel, manipule l'image le texte la vidéo ; bientôt un moteur multimédia.

1milliard de secondes = …

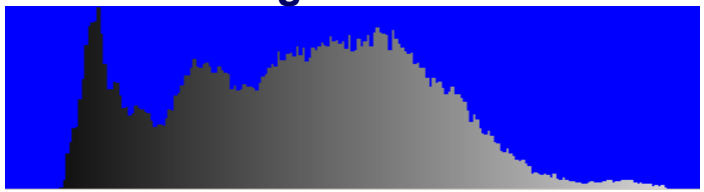
Analyse des images: Création de signatures

Extraction de caractéristiques

Texture, Couleur, Forme, Points d'intérêt

Information globale

Ex: Histogramme



Information locale

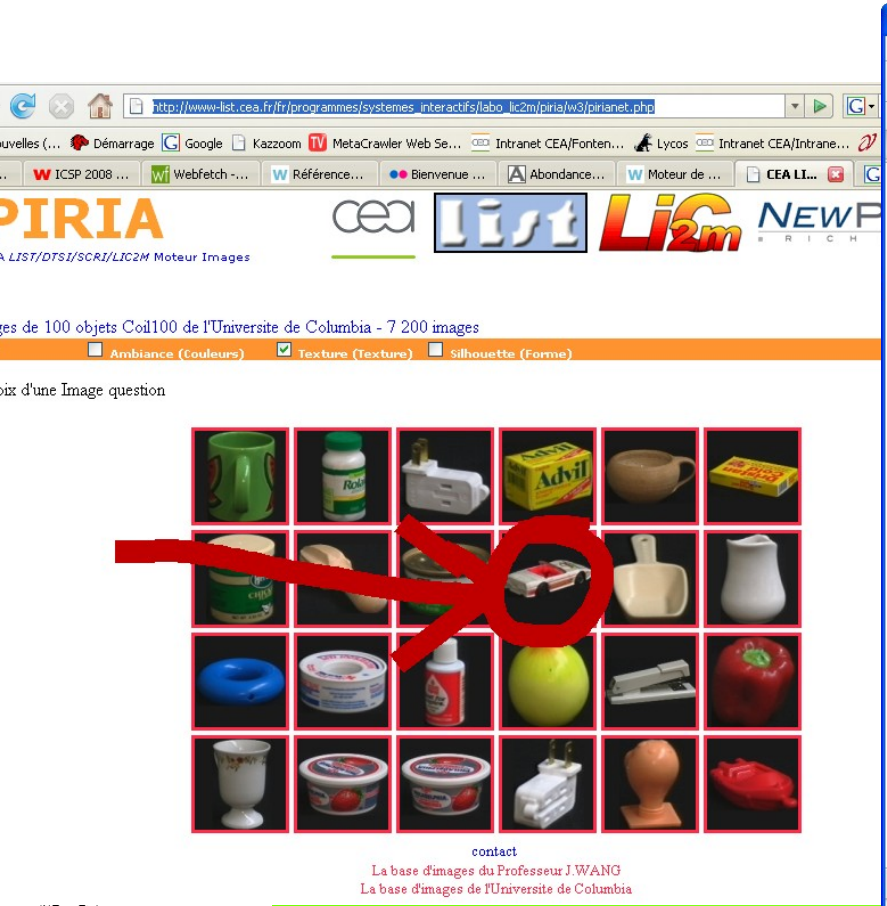
Ex :Points singuliers

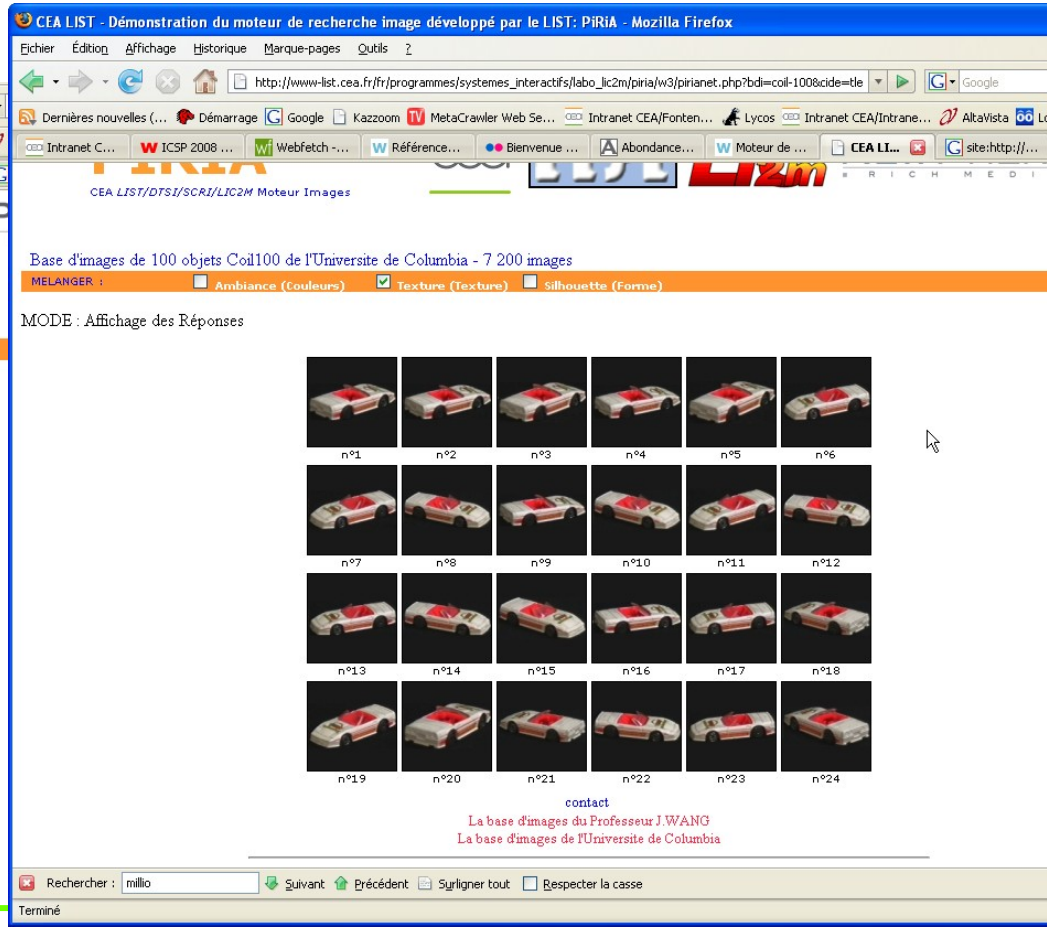
Avec segmentation systématique

signature de l'image

Moteur indexation/recherche

- CBIR Piria
  - ✓ Recherche par le contenu (Hors Google, FlickR)





Pourquoi indexer?

- En aout 2005 Google recence 2,1 milliards d'images sur le web sans doute environ 5 milliards en Février 2008
- En Octobre 2006 FlickR 3000 images/jour , 4000 en Février 2008

- Ina, Nasa, les particuliers sont des gros générateurs de contenu numériques. Mais très peu, moins de 10% du web annotent.
- Un disque dur de 1Tera Octets pour moins de 500€ en grands magasins

volume

Recherche de documents

- **Problème : On recherche une certaine image(ou document:texte multilingue, vidéo, sons)**

  Solution : Sans doute une version proche existe sur internet!

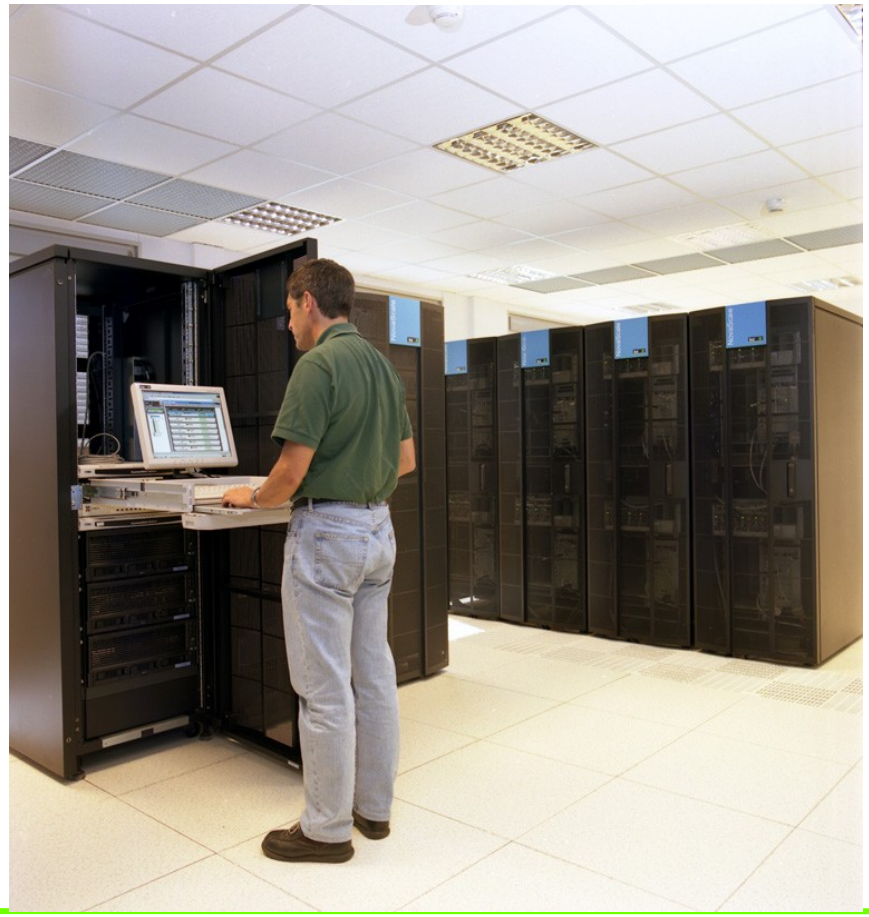
  Savoir que l'on dispose d'une information sans pouvoir y accéder revient a ne pas disposer de cette information.

  Sous Problème : comment la trouver?

- **Solution utiliser un moteur de recherche**

Fame2

- 88 cœurs de calcul hiérarchie de parallélismes intra et inter socket et inter modules
- Accélérateurs spécifiques
- 50 Terabytes de disques



Fame2

- Le grand challenge image
- Etat de l'art :Le système Cortina (Univ. Santa Barbara)
  - ✓ Indexation : 11millions en ligne
  - ✓ Recherche : 15 secondes
- Piria
  - ✓ Indexation : la plus grosse base 50 000 images
  - ✓ Recherche : 5s
- But: faire mieux!

Existant

- Source du Moteur sous Win32 .net
- 10 descripteurs globaux, 3 locaux
- Tests sur des bases de 1000 a 50 000 images
- http://www-list.cea.fr/fr/programmes/systemes_interactifs/labo_lic2m/piria/w3/pirianet.php

Constitution de la base

- **Difficulté N°1 : comment trouver plusieurs millions d'images © ?**

- **Multiplication des données**
  - ✓ **CLIC (CEA Lic2m Image Collection)**
    - Transformations géométriques et chromatiques 1million d'images a partir d'un noyau de 15 200.
- **Fame II**
  - ✓ **Corpus multimédia multilingue: wikipédia fr, gb,**

Constitution de la base

- **Fame II**
  - ✓ **Corpus multimédia multilingue: wikipédia fr, gb aspiré**
    - Un peu moins d'un million d'images
  - ✓ **Filtrage des ©, récupération des images**
  - ✓ **Mise à dimension 320x200 maxi et conversion au format JPEG**
  - ✓ **En conservant l'arborescence limitant 10 000 images maxi par répertoire**
  - ✓ **Application de 25 transformations sur toutes les images**
  - ✓ **Obtention de 22 millions d'images**

Indexation

- ✓ **Choix d'un descripteur de Piria existant**
- ✓ **Amélioration de ce descripteur**
  - Réduction de la dimension de sa signature par 2
  - Accélération du code de calcul par optimisation
  - Élimination de fuite mémoires non détectées avant le passage a l'échelle.
- ✓ **Conception et Codage d'une méthode permettant l'indexation de plusieurs répertoires**
- ✓ **Écriture d'un wraper permettant via les MPI l'exécution simultanée de plusieurs instances du moteur Piria**
- ✓ **Monté en volume par itération successives**

- ✓ **Moins de 100 heures de calcul**

Descripteur

- **Couleur Contour BIC**
  - ✓ **Ref:*A compact and efficient image retrieval approach based on border/interior pixel classification* Stehling,Nascimento,Falcão CIKM 2002**

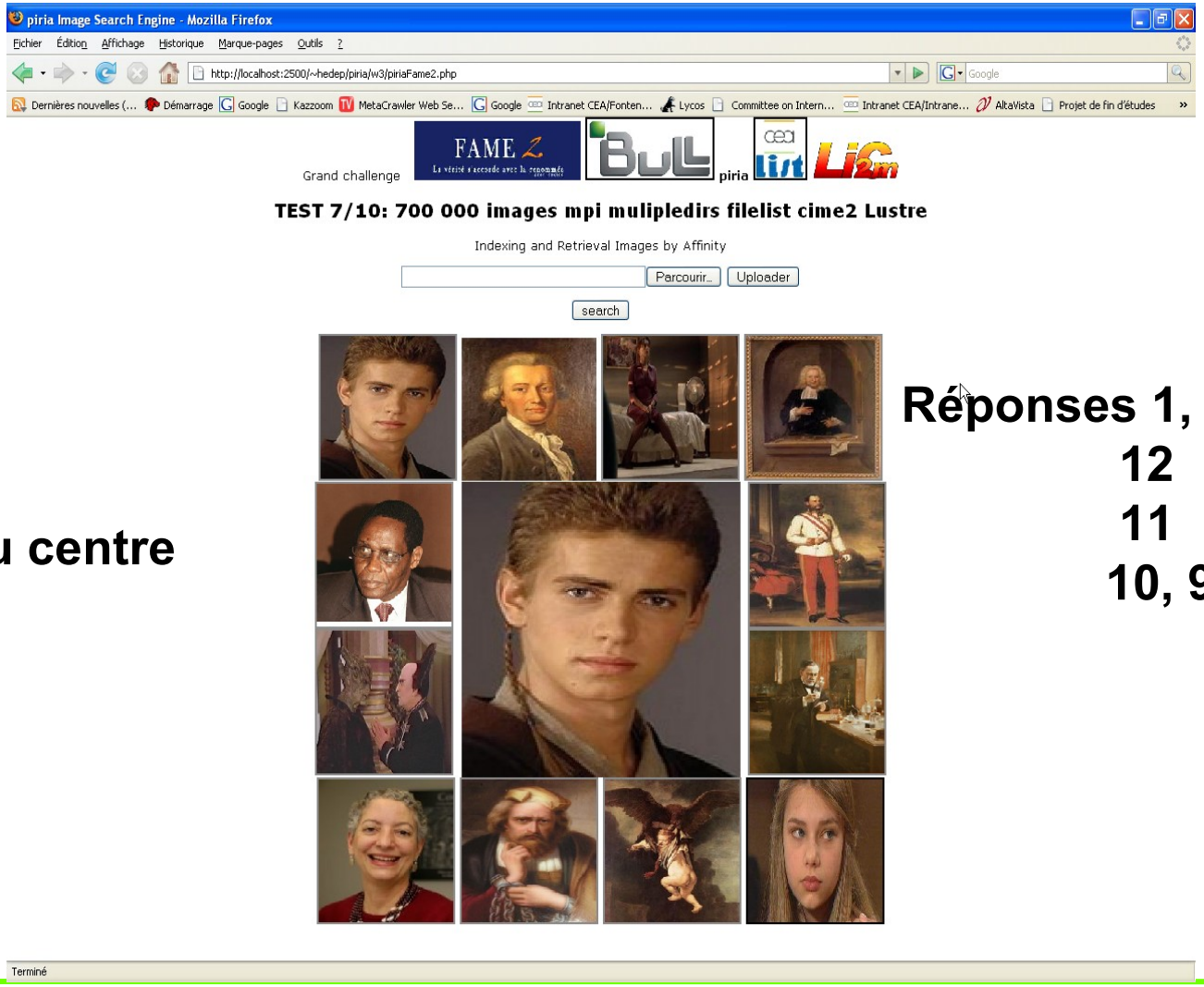
Travail réalisé

- Portage Linux P4 (Woodcrest), I64 (Montécio)
- Aspiration des images de wikipédia (français, anglais, commons)
  - Mise en forme (jpeg, resize)
- Création d'une base de 22 millions par transformations,duplication
- Amélioration d'un descripteur (réduction de la taille de sa signature par 2)

- Écriture d'un wrapper MPI Bull v 1.
- Architecture logicielle 'scalable'
  - En nombre de cœur de 1 a 'n' (test: 1, 16, 48 et 80 coeurs)
  - Arborescence d'image quelconque (lustre plus de 700 répertoires)
- Indexation de 10 000, 60 000, 101 000, 700 000, et 22Millions d'images 3To de données
- Tests de pertinence et de temps de réponse
- Conception et écriture d'une Interface php (Avec Bull serveur web) pour l'interrogation et l'affichage.

Démonstrateur

Soyez indulgent…



Requête au centre

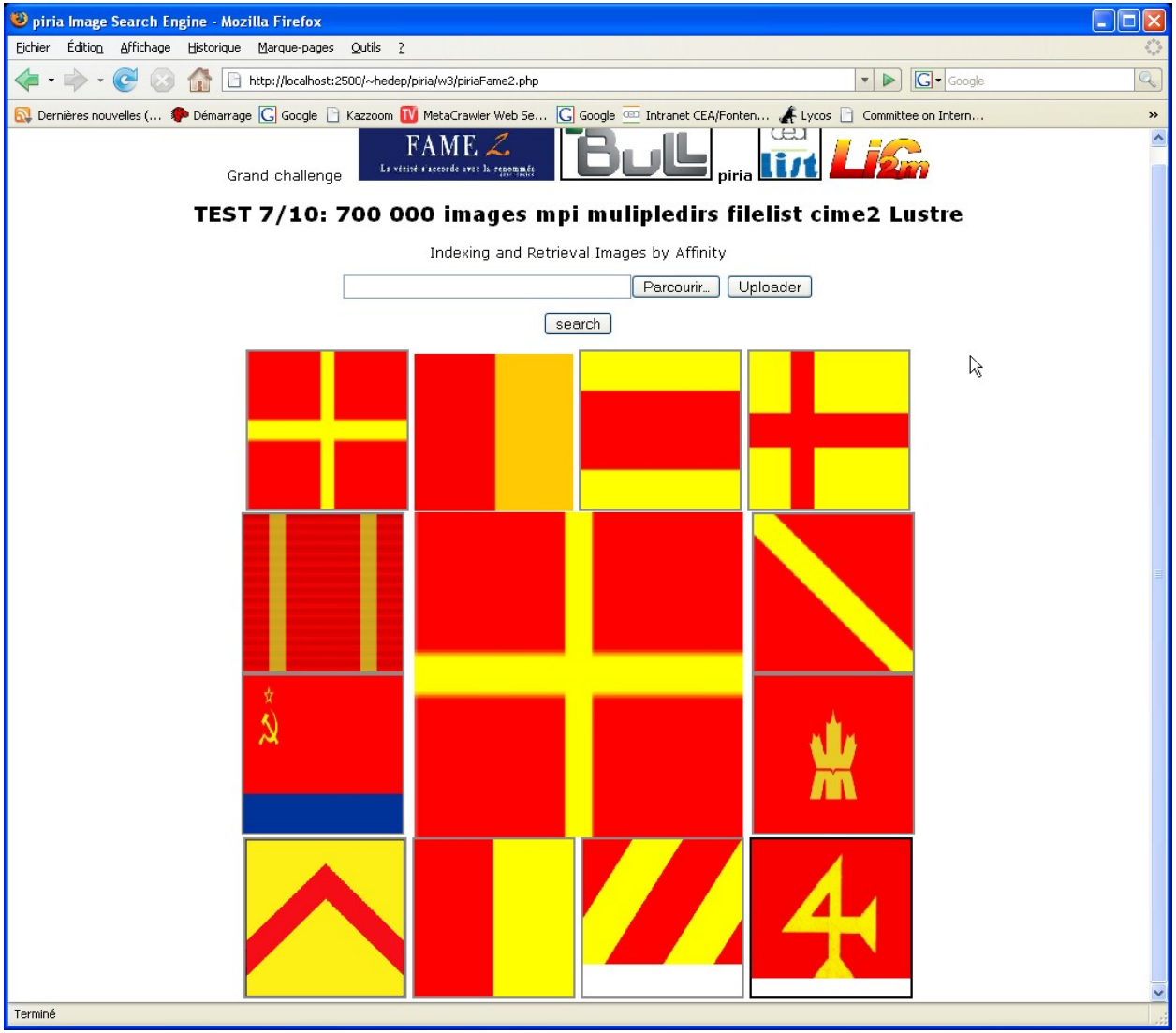
Réponses 1, 2, 3, 4

12 5

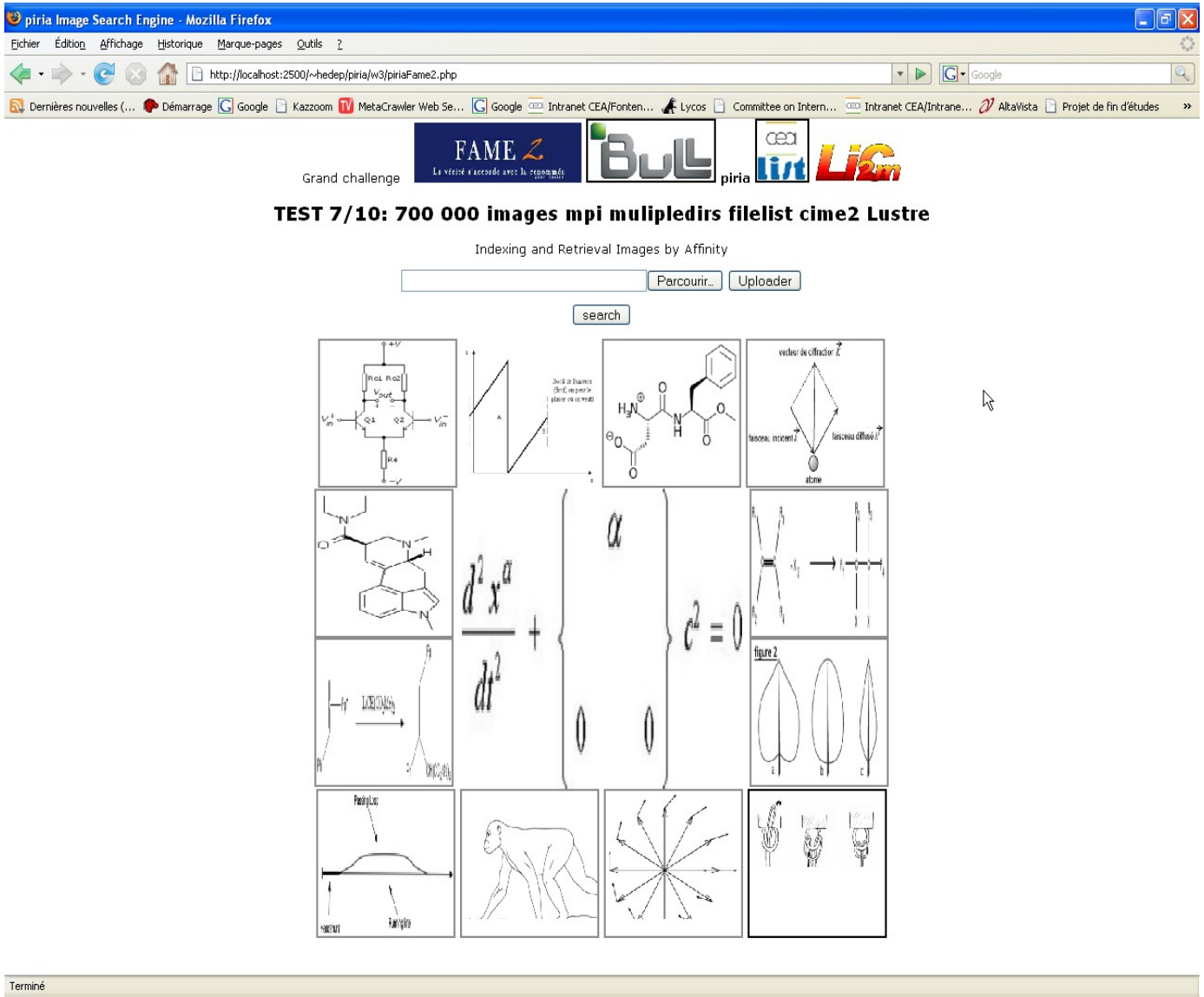
11 6

10, 9, 8, 7

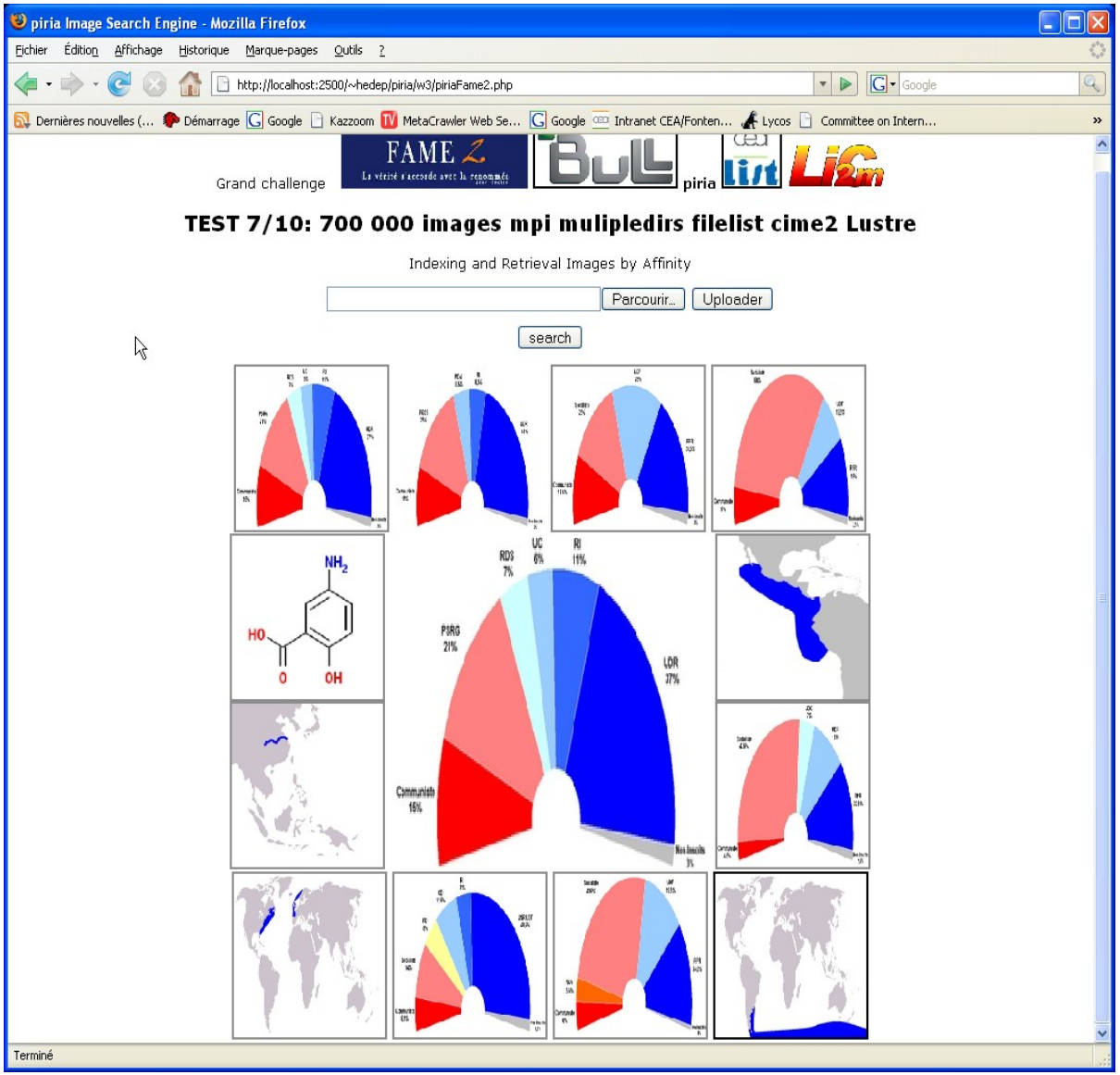
Recherche


piria Image Search Engine - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://localhost:2500/~hedep/piria/w3/piriaFame2.php
Grand challenge
FAME Z
BULL
piria
TEST 7/10: 700 000 images mpi mulipledirs filelist cime2 Lustre
Indexing and Retrieval Images by Affinity
Parcourir...
Uploader
search
Terminé

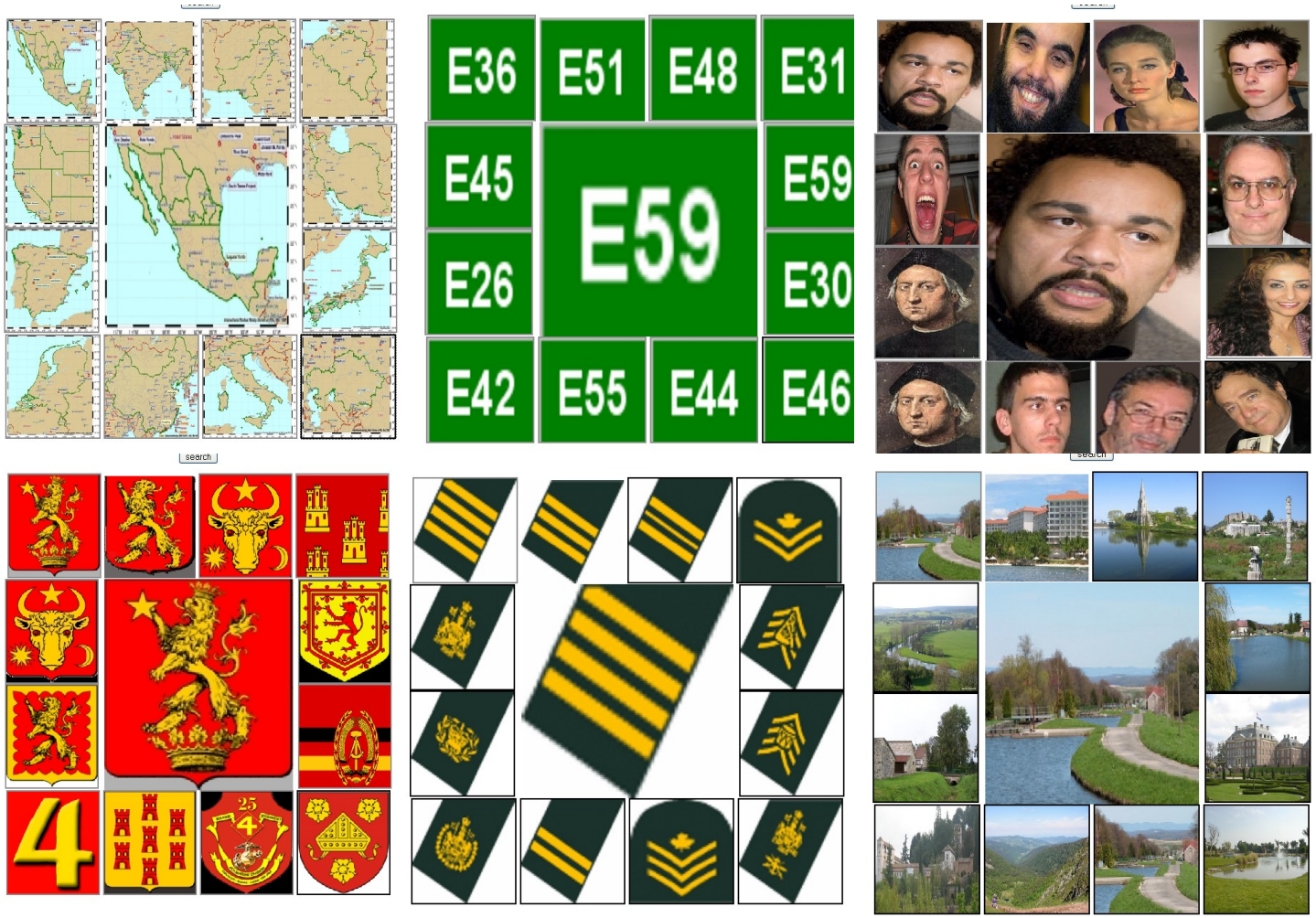
Recherche


piria Image Search Engine - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://localhost:2500/~hedep/piria/w3/piriaFame2.php
Grand challenge
FAME
piria
TEST 7/10: 700 000 images mpi mulipledirs filelist cime2 Lustre
Indexing and Retrieval Images by Affinity
Parcourir...
Uploader
search
Terminé

Recherche


piria Image Search Engine - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
http://localhost:2500/~hedep/piria/w3/piriaFame2.php
Grand challenge
FAME 2
BULL
piria
TEST 7/10: 700 000 images mpi mulipledirs filelist cime2 Lustre
Indexing and Retrieval Images by Affinity
Parcourir...
Uploader
search
RDS 7%
UC 6%
RI 11%
PIRG 21%
LDR 37%
Communisté 16%
Terminé

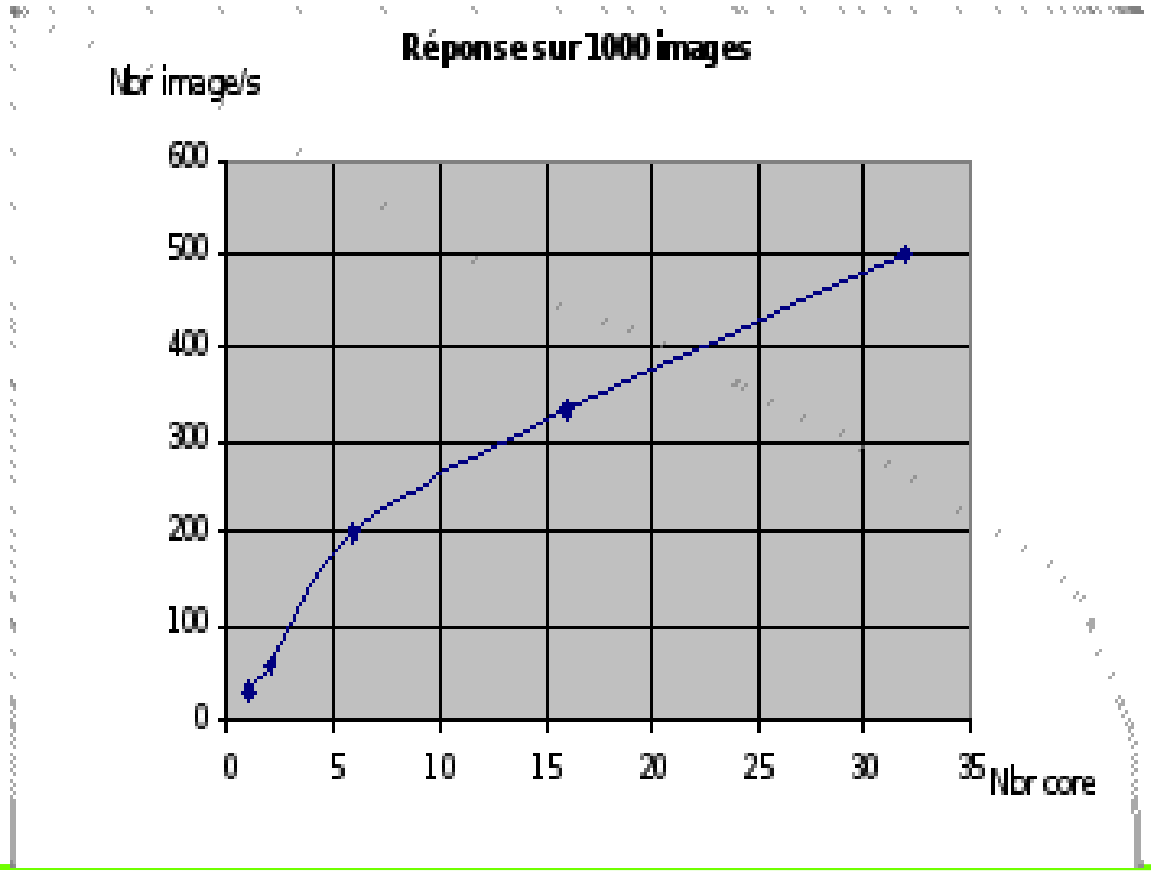
Recherche


E36
E51
E48
E31
E45
E59
E59
E26
E30
E42
E55
E44
E46
search
search

Performances

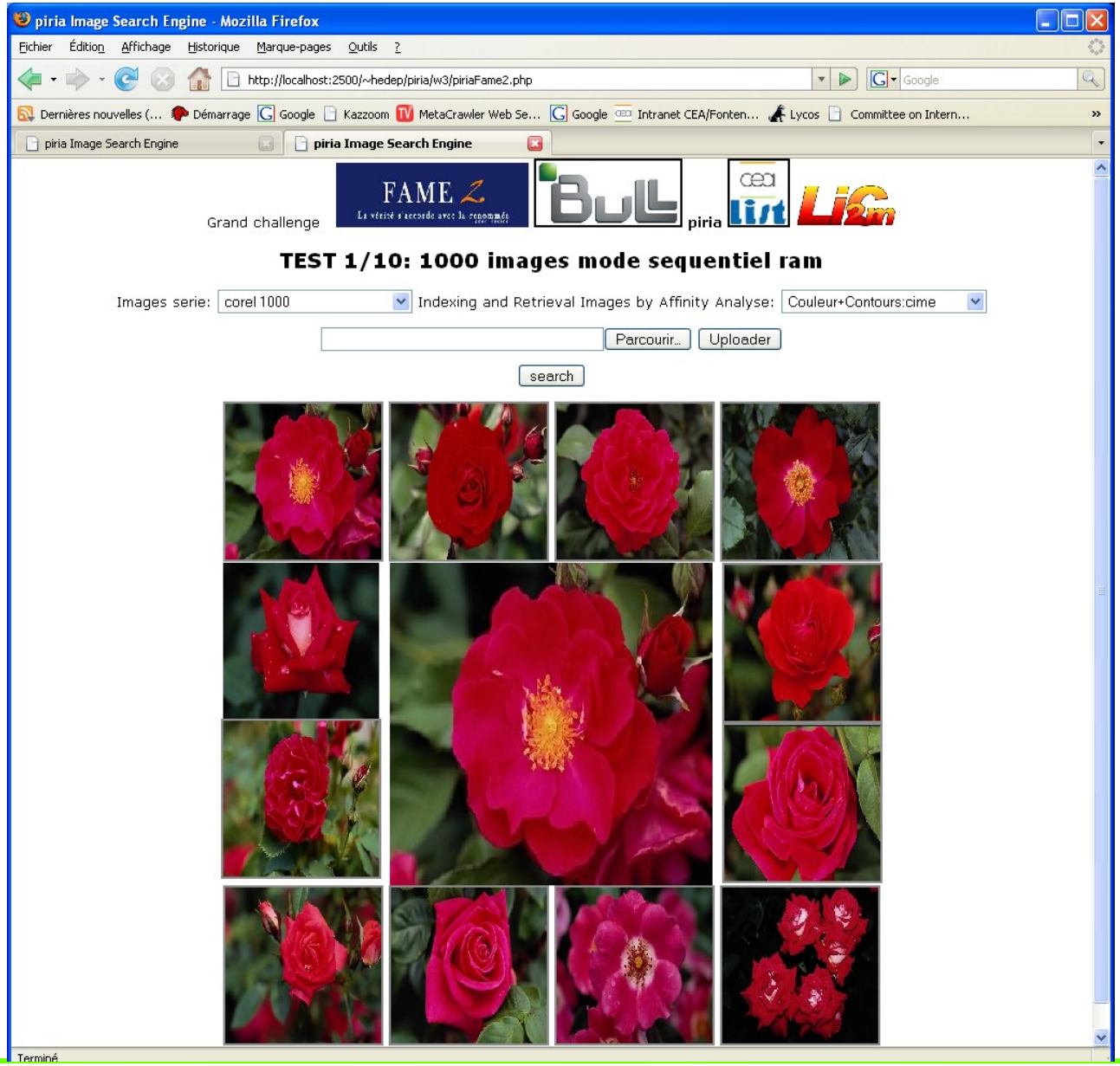
3Millions d'image / seconde en recherche

Nov 04: 500.000 img en 30h **Indexation**
Jun 07: 500.000 img en 30 min (gain de 60) **Indexation**
PiriaFame2 22Millions 6s **Recherche**

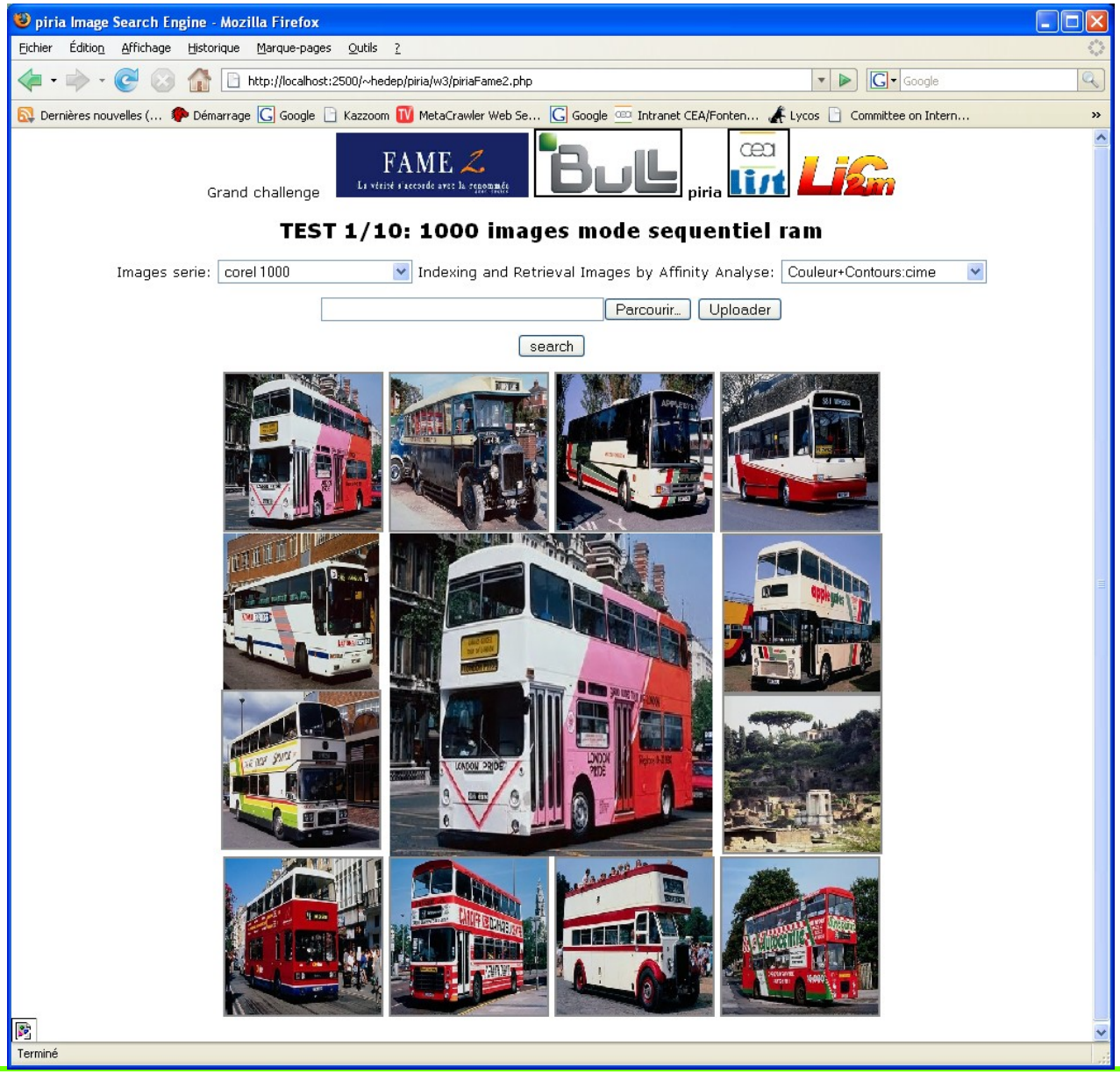


Démonstrateur (sur corel1000)

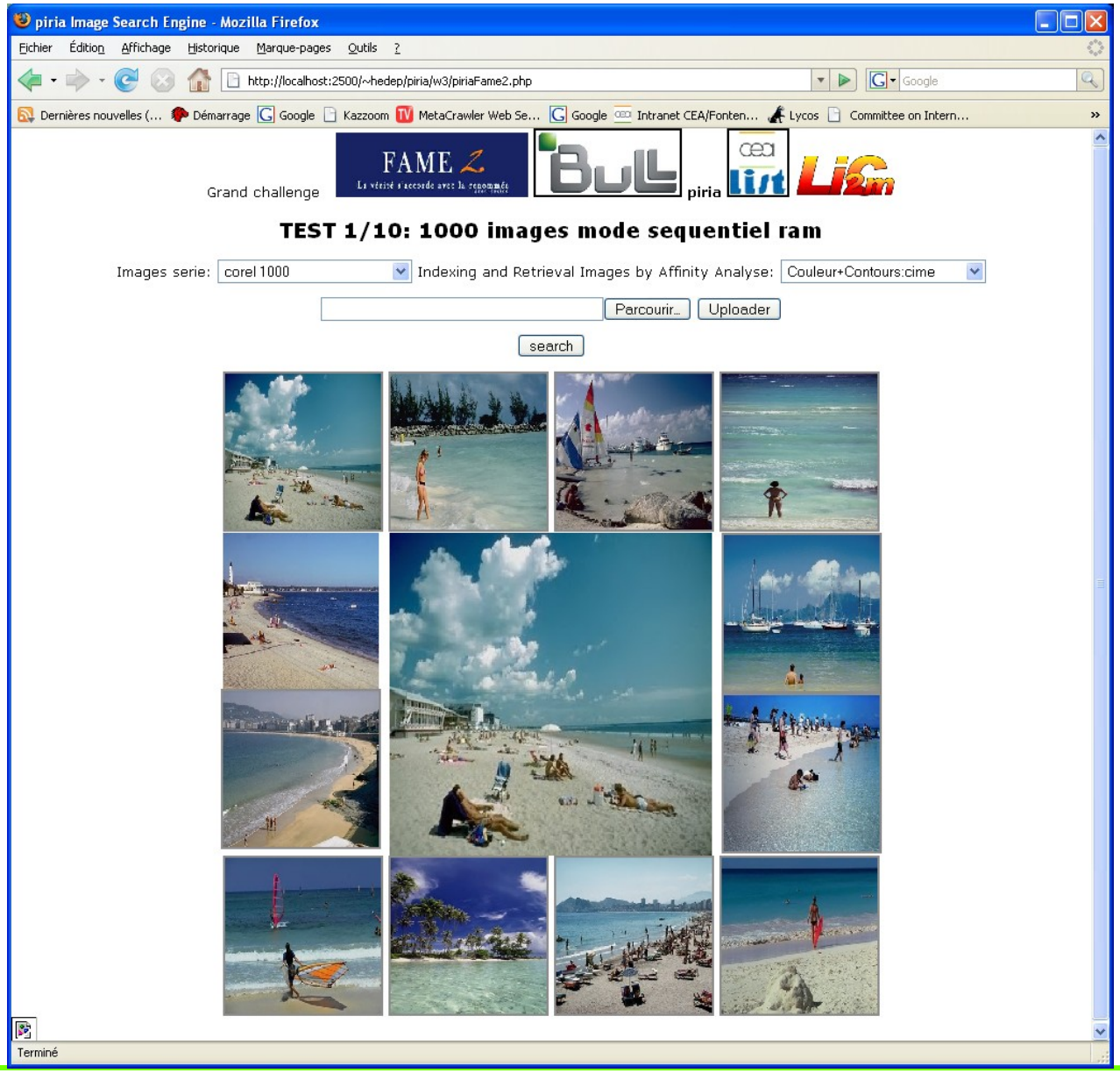
Démonstrateur (sur corel1000)



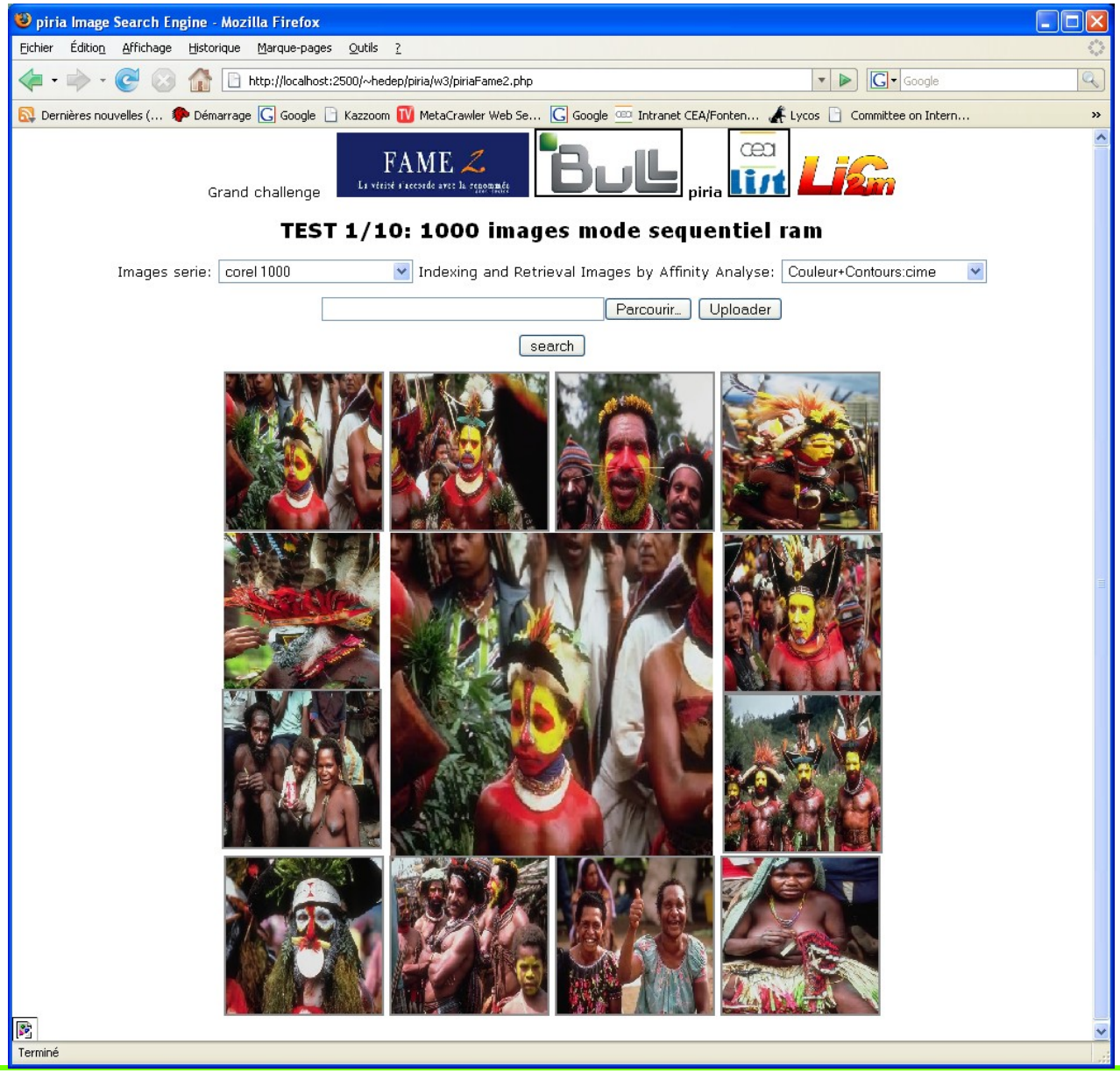
Démonstrateur (sur corel1000)



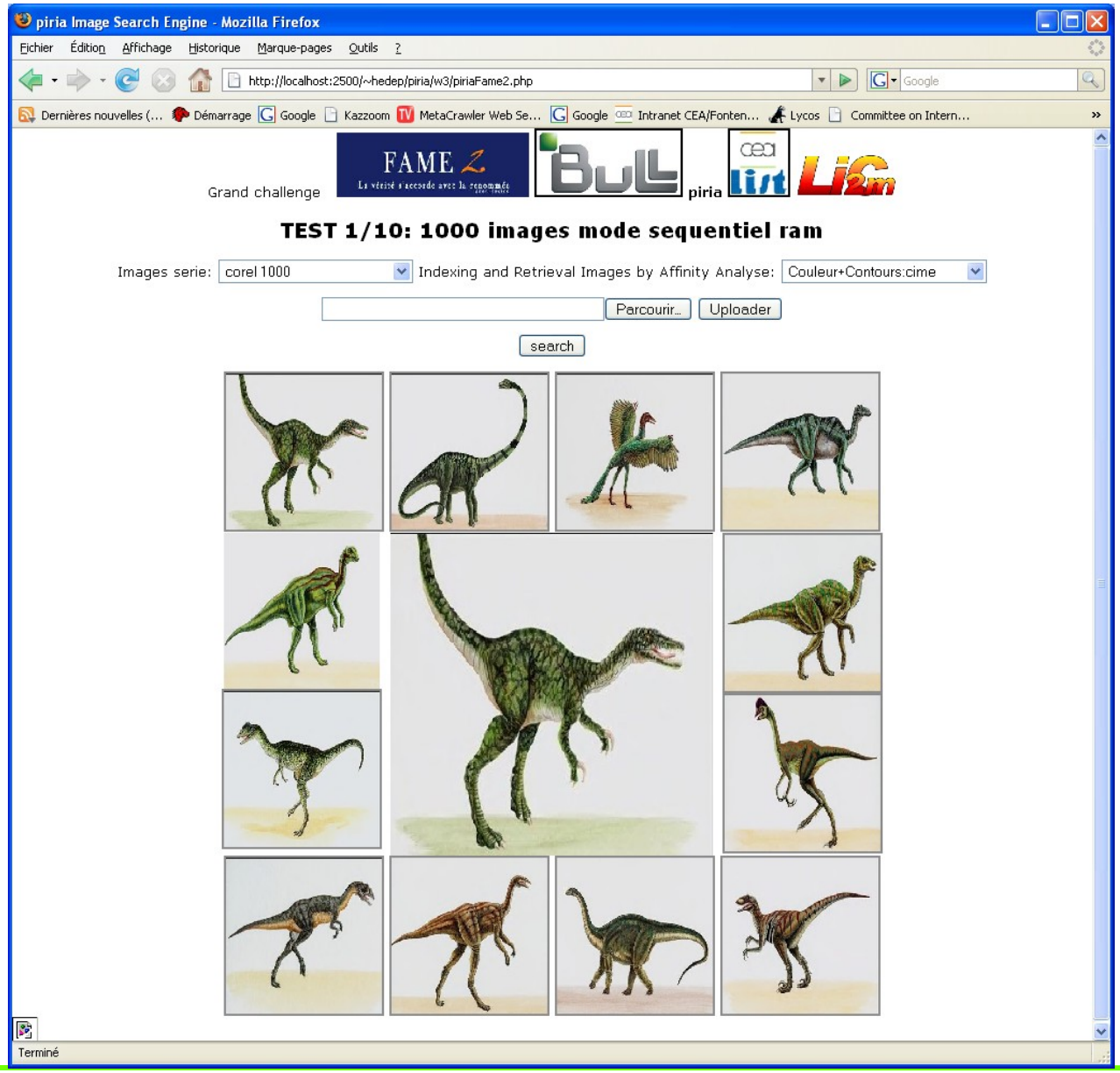
Démonstrateur (sur corel1000)



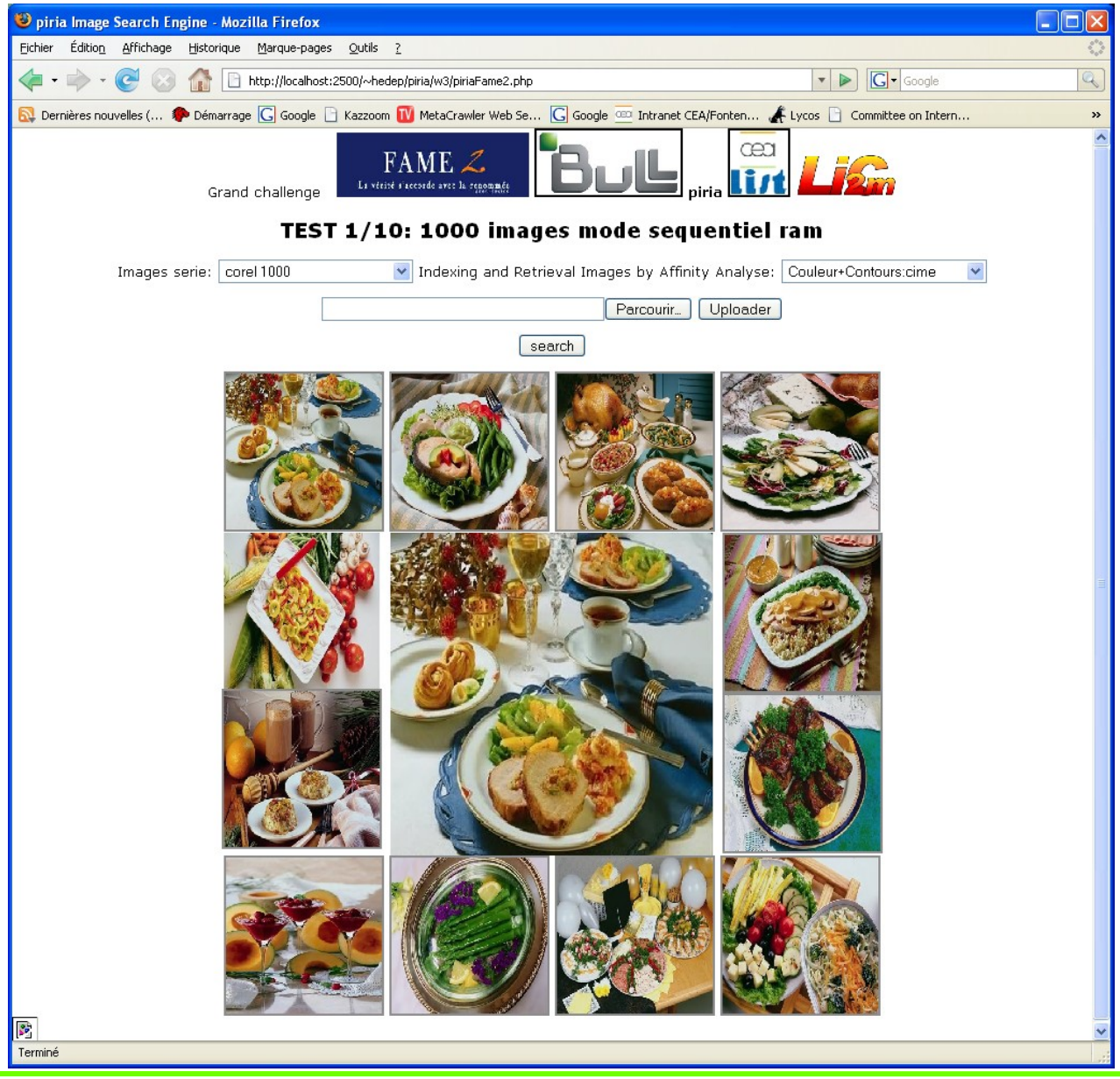
Démonstrateur (sur corel1000)



Démonstrateur (sur corel1000)



Démonstrateur (sur corel1000)



Conclusions Perspectives

- **Machine BULL**
  - **Performante**
  - **Record tout relatif**

- **Moteurs de demain:**
  - ✓ Indexer une taille proche du web (qq Milliards d'images)
  - ✓ Gérer tous les médias images (fixes, annimations 2D, 3D),vidéos, son, textes multilingue **POPS**
  - ✓ Descripteurs numériques plus proches de nos perceptions et de notre sémantique (réduction du fossé sémantique)
  - ✓ En temps réel

MERCI DE VOTRE ATTENTION