BIOTIM

Exploitation de Gisements Texte-Image en Biodiversité

Éric de la Clergerie
Eric.De_La_Clergerie@inria.fr

ATOLL/INRIA

Journées PaRISTIC
Labri, 21-23 Novembre 2005

BIOTIM

ACI Masses de Données (2003 - 2006)

Objectifs :

- Analyse et structuration de masses de textes et de masses d'images
- Acquisition d'une surcouche sémantique commune pour l'intégration entre modalités
- Interrogations pluri-modales exploitant les complémentarités entre textes et images

Participants

- Imedia (INRIA) [Coordinateur] : traitement des images
- Atoll (INRIA) : traitement textuel
- Vertigo (Cedric/CNAM) : BD, passage à l'échelle
- Équipe Contraintes et Apprentissage (LIFO, Univ. Orléans) : apprentissage
- Unité Biodival (IRD, Orléans) : Corpus botaniques et expertise
- UMR Génomique Végétale (INRA, Évry) : Corpus images et expertise

Jacaranda mimosaefolia

2. Jacaranda mimosifolia D. Don

Bot. Reg. : *tab. 631* (1822).– Sandwith, Kew Bull. 1953 : 455 (1954).– Gentry, Ann. Missouri Bot. Gard. 60 : 803 (1973).

Arbre atteignant au moins 15 m de hauteur et un diamètre de 45 cm ; ramilles cylindriques. *Feuilles* composées bipennées, longues de 15-30 cm, avec 13-31 pennes, chaque penne longue de 5-10 cm avec un rachis ailé et 13-41 folioles sessiles, celles-ci 3-12 x 1-4 mm, étroitement elliptiques, abruptement acuminées, base cunée , papyracées, glabres ou faiblement pubérulentes le long de la nervure médiane et du bord.

Inflorescence en panicule terminale ouverte, à ramifications courtement pubérulentes. *Calice* réduit, largement campanulé, 5-denticulé, long de 1 mm, large de 1,5-2 mm, subpubérulent au moins le long du bord. *Corolle* bleu violacé avec tube blanc intérieurement, tubulaire-campanulée au-dessus d'un col rétréci qui est peu visiblement courbé et faiblement élargi vers la base, longue de 3-4 cm, large de 0,7-1,2 cm à l'ouverture du tube, lobes longs de 0,5 cm environ, pubescente extérieurement, spécialement vers la base, pubescente intérieurement au niveau de l'insertion des étamines, et éparsement avec de longs trichomes dans le tube. *Etamines* didynames, anthères monothèques, seconde thèque réduite à un très petit appendice, chaque thèque longue de 2 mm, filets pubescents sous le milieu, staminode long de 2-2,5 cm, zone moyenne et sommet glandulaire-pubescents. *Pistil* long de 2,2-2,5 cm, ovaire aplati cylindrique, long de 3-3,5 mm, large de 1,5-2 mm, glabre, ovules 6-8 sériés dans chaque loge. *Disque* court, pulviné.

Capsule orbiculaire comprimée, souvent peu profondément émarginée au sommet et tronquée à largement mais peu profondément subcordée basalement, longue de 3,2-5,8 cm, large de 3,7-5,5 cm, brun rougeâtre en séchant, bords non ondulés. *Graines* minces, 0,9-1,2 x 1,1-1,7 cm, aile ± entourante et nettement démarquée du corps de la graine.– Pl. 17, p. 55.

Arbre ornemental très communément cultivé dans les régions subtropicales du Monde ; moins communément cultivé dans les régions tropicales de basse altitude. Originaire du Nord de l'Argentine et de la partie adjacente de Bolivie.

NOM USUEL : Jacaranda.

MATERIEL CAMEROUNAIS ETUDIE :

Dang 520, Meiganga, P, YA.

3. Podranea ricasoliana (Tanf.) Sprague

in Thiselton-Dyer, Fl. Cap. 4 (2) : 449 (1904).– Gentry, Flora of Ecuador 7 : 122 (1977).
– *Tecoma ricasoliana* Tanfani, Bull. Soc. Tosc. Ortic. 1887 : 17 (1887).
– *Pandorea ricasoliana* (Tanfani) Baill. Hist. Pl. 10 : 40 (1888).

Plante grimpante ou *arbuste* sarmenteux. *Feuilles* composées imparipennées ; folioles ordinairement 7-9, sans vrille, ± ovées, obtuses à courtement acuminées, dentées, tronquées à cunées ou un peu atténuées, 2,5-3,8 x 1,5-2 cm, pétiolulées.

Inflorescence en panicule terminale. *Calice* large, cupulaire-campanulé, long de 1,5-2 cm, fortement 5-denté, avec des lobes apiculés, lavande pâle sur le frais. *Corolle* lavande pâle ou rosâtre avec une tache magenta à la base des

Pl. 17.– **Jacaranda mimosifolia** D. Don : 1, rameau fleuri x 1/2 ; 2, corolle face interne x 1 ; 3, étamine x 3 ; 4, fruit x 2/3 ; 5, graine x 1 *(1-3, Dang 520 ; 4, 5, Audru 3130,* Sénégal).

Jacaranda mimosaefolia

2. Jacaranda mimosifolia D. Don

Bot. Reg. : *tab. 631* (1822).— Sandwith, Kew Bull. 1953 : 455 (1954).— Gentry, Ann. Missouri Bot. Gard. 60 : 803 (1973).

Arbre atteignant au moins 15 m de hauteur et un diamètre de 45 cm ; ramilles cylindriques. *Feuilles* composées bipennées, longues de 15-30 cm, avec 13-31 pennes, chaque penne longue de 5-10 cm avec un rachis ailé et 13-41 folioles sessiles, celles-ci 3-12 x 1-4 mm, étroitement elliptiques, abruptement acuminées, base cunée , papyracées, glabres ou faiblement pubérulentes le long de la nervure médiane et du bord.

Inflorescence en panicule terminale ouverte, à ramifications courtement pubérulentes. *Calice* réduit, largement campanulé, 5-denticulé, long de 1 mm, large de 1,5-2 mm, subpubérulent au moins le long du bord. *Corolle* bleu violacé avec tube blanc intérieurement, tubulaire-campanulée au-dessus d'un col rétréci qui est peu visiblement courbé et faiblement élargi vers la base, longue de 3-4 cm, large de 0,7-1,2 cm à l'ouverture du tube, lobes longs de 0,5 cm environ, pubescente extérieurement, spécialement vers la base, pubescente intérieurement au niveau de l'insertion des étamines, et éparsement avec de longs trichomes dans le tube. *Etamines* didynames, anthères monothèques, seconde thèque réduite à un très petit appendice, chaque thèque longue de 2 mm, filets pubes-

Jacaranda mimosaefolia (pictures)

Collections de textes et d'images

Existence de collections couplant images et descriptions textuelles courtes :

- collections en biodiversité (flores **[BIOTIM]**, faunes, insectes, poissons, oiseaux, . . .)
- descriptions d'artefacts culturels (fournitures, peintures, . . .)
- images médicales commentées (echographies, Rayons-X, RMN, . . .)
- images satellite avec commentaires et/ou accès à des descriptions géographiques (cartes, légendes, . . .)

Collections de textes et d'images

Existence de collections couplant images et descriptions textuelles courtes :

- collections en biodiversité (flores **[BIOTIM]**, faunes, insectes, poissons, oiseaux, . . .)
- descriptions d'artefacts culturels (fournitures, peintures, . . .)
- images médicales commentées (echographies, Rayons-X, RMN, . . .)
- images satellite avec commentaires et/ou accès à des descriptions géographiques (cartes, légendes, . . .)

Principales caractéristiques :

- Images nécessitant interprétations et identifications
- Descriptions concises et précises, avec un vocabulaire spécialisé la plupart des termes correspondant à des propriétés visuelles
- identification (d'un taxon) au travers d'un arbre de décision

Traiter des collections scientifiques multimodales

- Préservation de connaissances scientifiques (anciennes)
  - migration du support papier à électronique,
    complété par une surcouche sémantique (acquise) pour des accès plus riches.
    (Projet Google Library ou BNF Gallica + WEB Sémantique)
  - connaissances précieuses (biodiversité, usage médical des plantes)

Traiter des collections scientifiques multimodales

- Préservation de connaissances scientifiques (anciennes)
  - migration du support papier à électronique,
    complété par une surcouche sémantique (acquise) pour des accès plus riches.
    (Projet Google Library ou BNF Gallica + WEB Sémantique)
  - connaissances précieuses (biodiversité, usage médical des plantes)

- Ouvrir l'accès à la connaissance (grand public, étudiants, décideurs, ...) au travers d'interfaces de recherche/navigation en texte & image
  $\Rightarrow$ réduire la fracture entre images et descriptions spécialisées

Traiter des collections scientifiques multimodales

- Préservation de connaissances scientifiques (anciennes)
  - migration du support papier à électronique,
    complété par une surcouche sémantique (acquise) pour des accès plus riches.
    (Projet Google Library ou BNF Gallica + WEB Sémantique)
  - connaissances précieuses (biodiversité, usage médical des plantes)

- Ouvrir l'accès à la connaissance (grand public, étudiants, décideurs, ...) au travers d'interfaces de recherche/navigation en texte & image
  $\Rightarrow$ réduire la fracture entre images et descriptions spécialisées

- Exploiter les connaissances des experts présentes au seinn des descriptions textuelles dans des programmes d'analyse.
  - ainsi, analyse automatique d'images satellite ou médicales, experience INRA d'expression de gènes
  - toujours plus d'images, attendant d'êtres intelligemment indexées, classifiées, et traitées (Google Earth)

Traitement textuels : des livres aux bases de connaissances

Flore du “Cameroun” :

Volumes	Pages	Moy. Pages	Mots	Taxons
31	9466	305	1.5M	$\sim$ 2400

Traitement textuels : des livres aux bases de connaissances

Flore du “Cameroun” :

Volumes	Pages	Moy. Pages	Mots	Taxons
31	9466	305	1.5M	$\sim$ 2400

Tâches :

- Préparation des corpus :
- Traitement linguistique préliminaire : traitement morpho-syntaxique
- Extraction de la terminologie :
  liste des mots simples fréquents et complexes co- occurrents (N de N, N Adj, . . .)
  $\Rightarrow$ indexation de documents, base d’un thesaurus, lexique
- Acquisition d’une ontologie (lexicale), grâce à l’analyse syntaxique :
  $\Rightarrow$ lancéolé (adj) : **forme** de **feuille**
- Fouille de texte : obtenir les propriétés de chaque taxon

Beaucoup d’autres flores à traiter : Sénégal, Nouvelle-Calédonie, Guyane, Polynésie, . . .

Jacaranda – Texte OCR brut (avec typos)

2. Jacaranda mimosifolía D. Don

Bot. Reg. : tab. 631 (1822).- Sandwith, Kew Bull. 1953 : 455 (1954)
Missouri Bot. Gard. 60 ; 803 (1973).

Arbre atteignant au moins 15 m de hauteur et un diamètre de 45 cm ,
ramilles cylindriques. Feuilles composées bipennées, longues de 15-
. . avec 13-31 pennes, chaque penne longue de 5-10 cm avec un rachi
folioles sessiles, celles-ci 3-12x1-4 mm, étroitement elliptiques,
acuminéès, base cunée , papyracées, glabres ou faiblement pubérulen
de la nervure médiane et du bord.

inflorescence en panicule terminale ouverte, ;à ramifications court
pubérulentes. Calice réduit, largement campanule, 5-denticulé, <..

NOM USUEL ¡Jacaranda.

MATERIEL CAMEROUNAIS ETUDIE :

Dang 520, Meiganga, P, YA.

Corrections Post OCR – illustration

Méthodes : proximité entre mots du corpus words et entrées lexicales, complétées par mots (longs) du corpus à forte fréquence
+ apprentissage de tables de correction

Errors	Correction
pubescences	pubescentes
généralemint	généralement
Ierbier	herbier
nigerian	Nigerian
lordu	tordu

(a) Corrections

Errors	Correction
ß	B
6	é
i	l
l	t

(b) Table de Correction

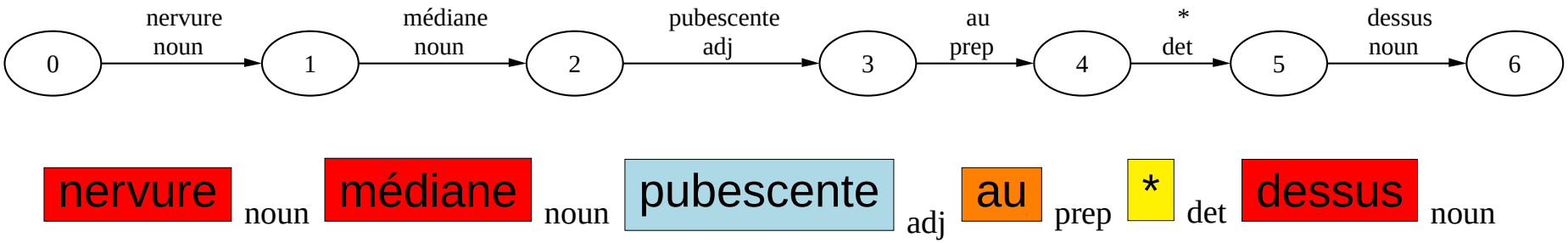
Jacaranda – Structuration logique

Méthode : détection de frontières + ouverture/fermeture d'éléments XML
Robustesse & flexibilité pour traiter des corpus légèrement différents

```
<genre id="section25" key="2" name="Jacaranda␣mimosifolía␣D.␣Don">

<biblio>
 Bot. Reg. : tab. 631 (1822).— Sandwith, Kew Bull. 1953 : 455
 (1954).— Gentry, Ann. Missouri Bot. Gard. 60 ; 803 (1973).
</biblio>

<p>Arbre atteignant au moins 15 m de hauteur ....</p>
<p>inflorescence en panicule terminale ouverte, .... </p>
<p>Capsule orbiculaire comprimée, souvent peu profondément émarginée au.
    sommet ... </p>
<p>Arbre ornemental très communément cultivé dans les régions
    subtropicales du Monde ... </p>
 ....
NOM USUEL ¡Jacaranda.

<materiel>
Dang 520, Meiganga, P, YA.
</materiel>
</genre>
```

Traitement Morpho-Syntaxique

Motivation : Segmentation en mots et phrase, reconnaissance des entités nommées (noms propres, dates, dimensions, . . .), information morphologique




```
<token id="t20724" value="rachis">rachis</token>
<wordForm entry="rachis" lemma="rachis" tag="cat@noun␣type@common␣
  gender@masc" tokens="t20724"/>
<token id="t20725" value="portant">portant</token>
<wordForm entry="portant" lemma="porter" tag="cat@verb␣mode@part␣
  tense@pres" tokens="t20725"/>
<token id="t20726" value="quelques">quelques</token>
<wordForm entry="quelques" lemma="quelques" tag="cat@det␣num@pl" tokens
  ="t20726"/>
<token id="t20727" value="aiguillons">aiguillons</token>
<wordForm entry="aiguillons" lemma="aiguillon" tag="cat@noun␣
  type@common␣gender@masc␣num@pl" tokens="t20727"/>
```

Extraction terminologique – Résultats (Acabit)

Termes	Loglike
m de longueur	1734.856
De Wild	1731.003
paire de foliole	1668.948
nervure latérale	1661.068
paire de nervure	1660.56
forêt dense	1617.032
forêt dense humide	1598.193

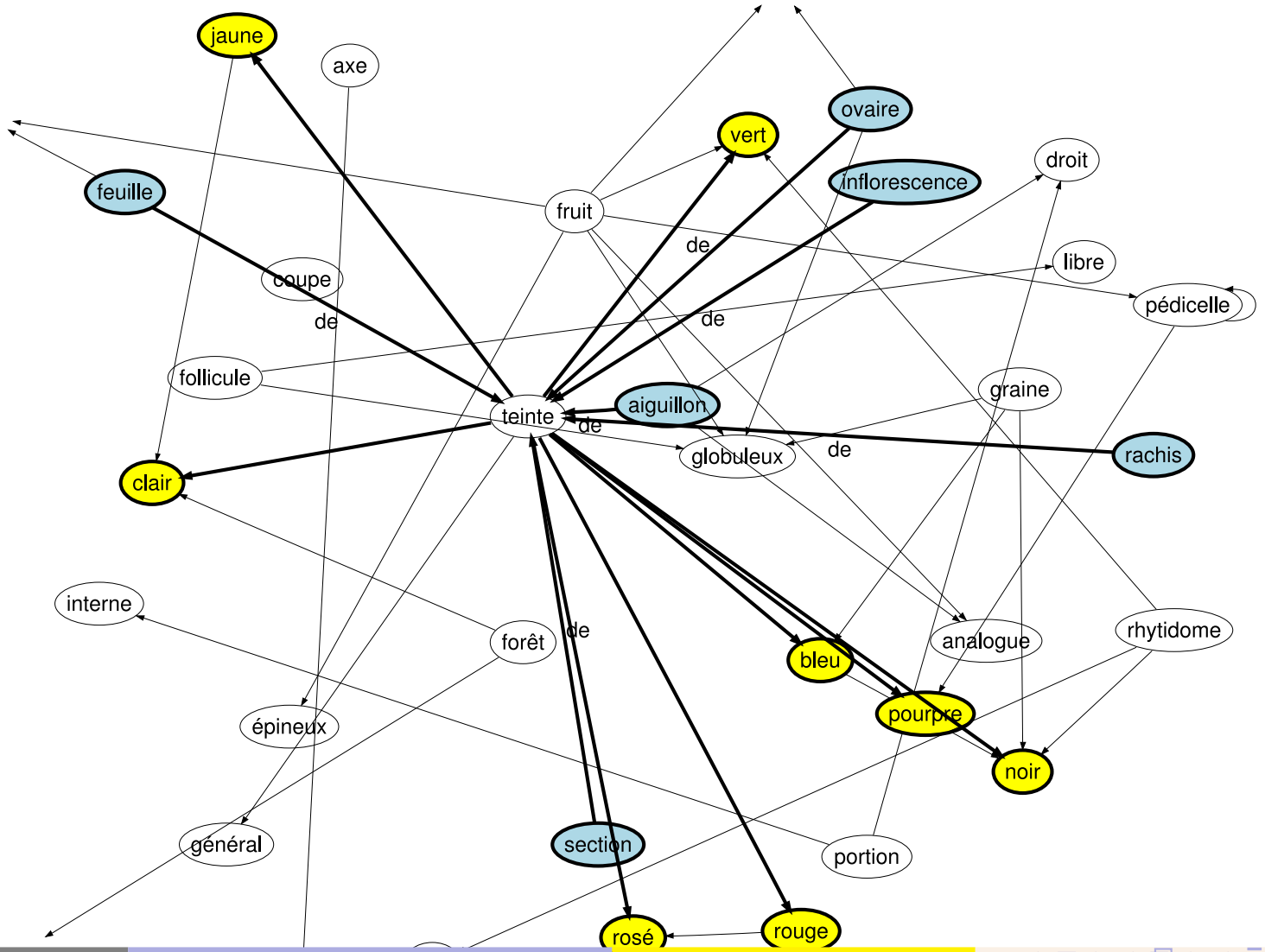
arbuste lianescent	arbuste iriermes	arbuste à feuille
arbuste inerme	arbuste épine	port d'arbuste
arbuste épineux	arbuste ornemental	premier arbuste
arbuste sarmenteux	arbuste lianescents	massif d'arbuste
arbuste de sous-bois	arbuste à jeune	espèce d'arbuste
arbuste à coloniser	arbuste à rameau	. . .

Notes Résulats obtenus sans structuration logique et faible détection des entités nommées

Vers un réseau de mot (fragment)

Exploitation des relations *gouverneurs/gouverné* présentes dans les termes multi-mots.

$\Rightarrow$ premiers pas vers des classes sémantiques



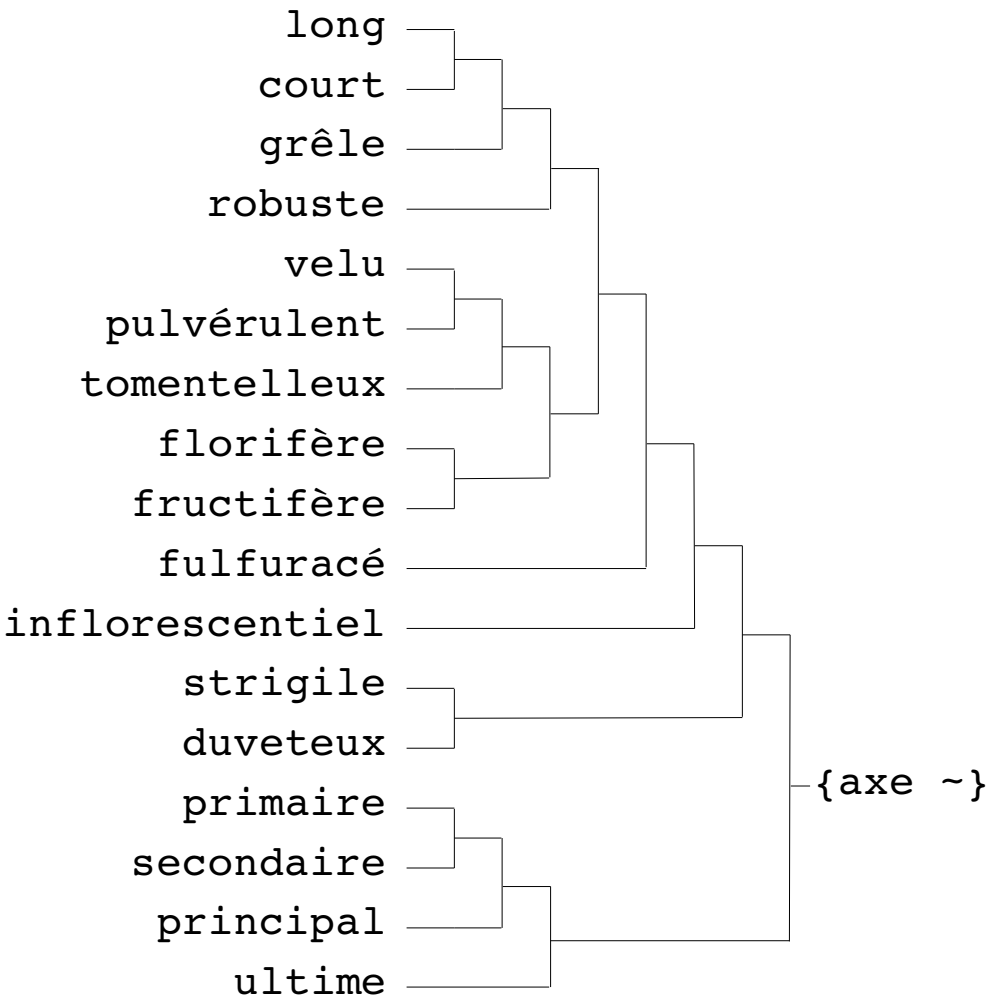
Vers des classes sémantiques

- Préfiguration des expériences avec analyse syntaxique
- Utilisation des informations morpho-syntaxique pour obtenir des **contextes** (locaux) d'occurences
  Nom-(Prep.(Det))-□ Adj-□ □-(Prep.(Det))-Nom
- **Hyp distributionelle d'Harris** : des mots partageant souvent les mêmes contextes sont sémantiquement proches
- Proximité (asymétrique) entre mots fonction de la similarité et productivité des contextes

$$\alpha(m_i, m_j) = \frac{\sum_{c \in C_i \cap C_j} \frac{1}{\sqrt{prod(c)}}}{\sum_{c \in C_i} \frac{1}{\sqrt{prod(c)}}}$$

$$p(m_i, m_j) = \alpha(m_i, m_j).\sum_{c \in C_i \cap C_j} \frac{1}{\sqrt{prod(c)}}$$

Regroupement hiérarchique – Dendogrammes

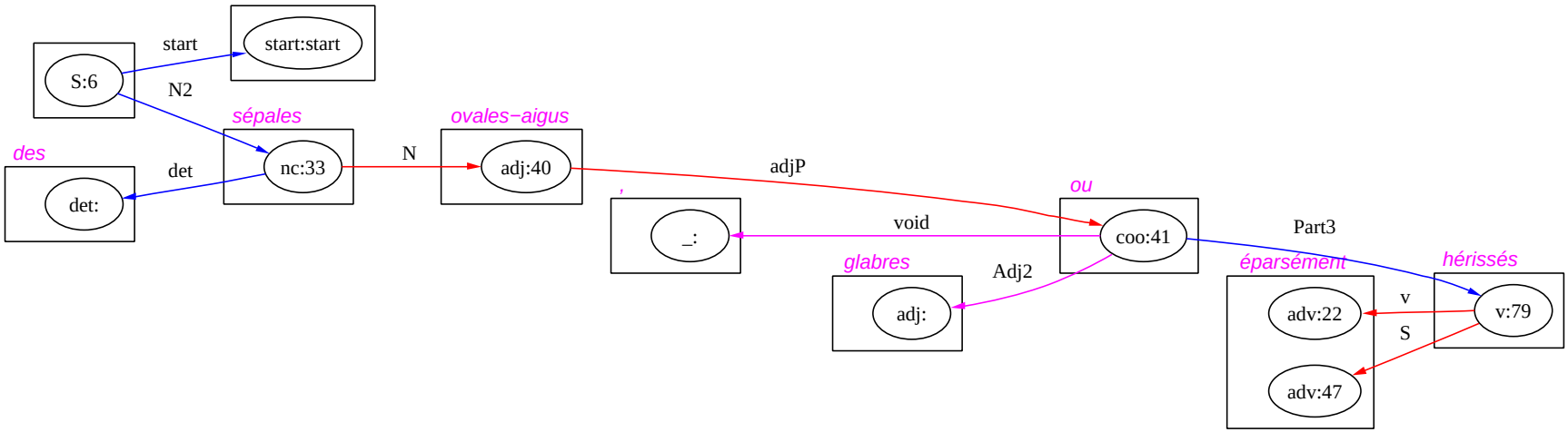


Dépendances syntaxiques

Sorties d'analyse comme un graphe (ambiguë) de dépendance entre mots utiles pour :

- acquisition d'ontologie :
  - contextes syntaxiques similaires $\Rightarrow$ identification classes sémantiques
    classes sémantiques $\Rightarrow$ classes de contextes syntactiques
  - constructions syntactiques : (coord) "X et X", (intervalle) "de X à X"
  - marqueur linguistique explicite : "en forme de X","de couleur Y"
  - ontologie germe
- fouille de textes (avec désambiguation grâce à l'ontologie)

des sépales [part] ovales-aigus [shape], glabres [texture] ou éparsement hérissés [texture]



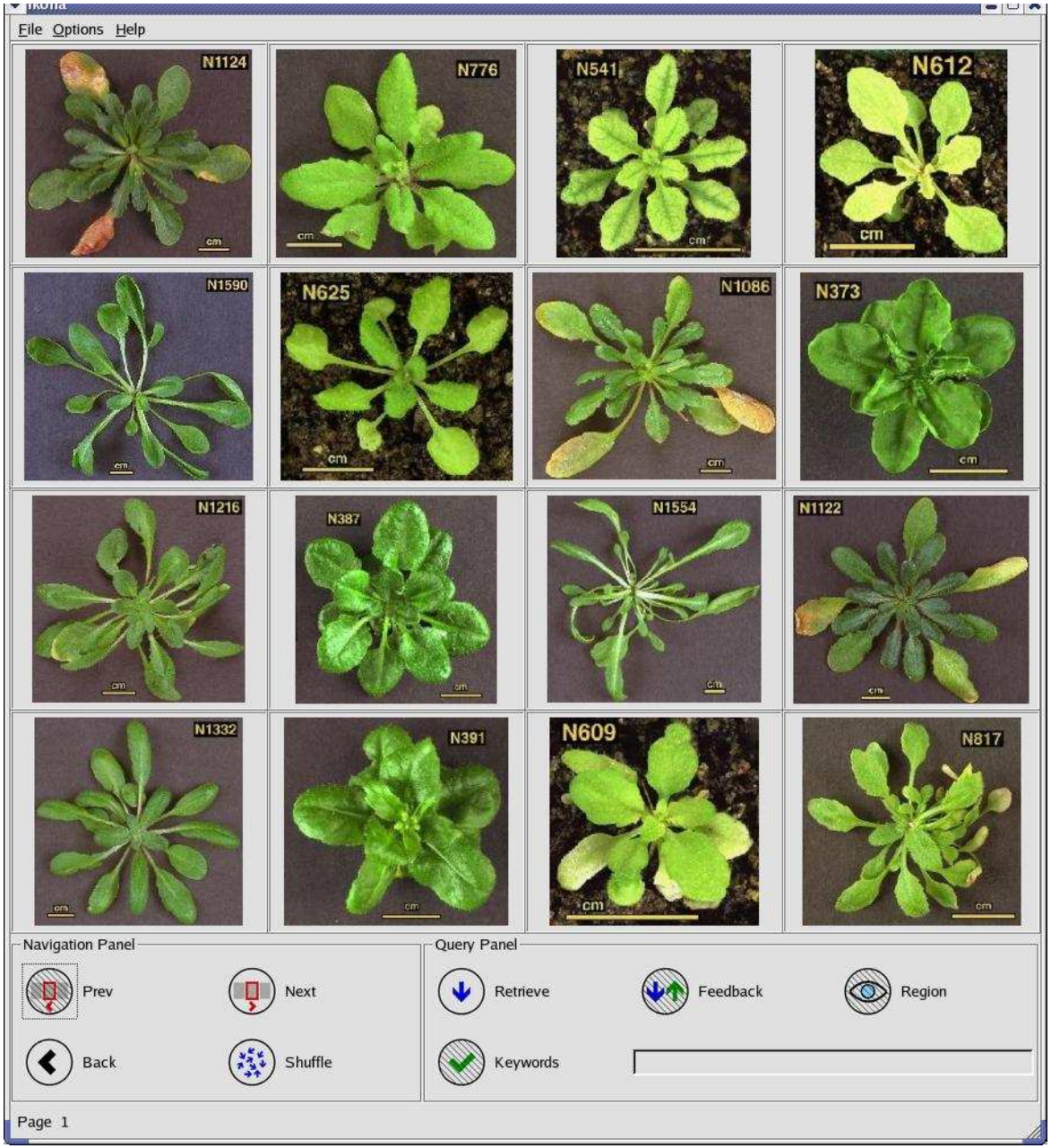
Traitement des images

- Indexation Offline des images :
  - calculs de signatures (couleurs, textures, . . .)
  - identification des points caractéristiques
  - identification des régions homogènes
- Recherche
  - recherche par similarité sur les signatures et motifs de points caractéristiques
- Recherche de signatures adapté
  (reconnaissance de plantes $\neq$ reconnaissance de visages)
- Apprentissage/identification de (certains) des descripteurs visuels de l'ontologie (lancéolé/forme/feuille, glabre/texture/feuille)
- Ikona demo : `http://www-rocq.inria.fr/cgi-bin/imedia/ikona`

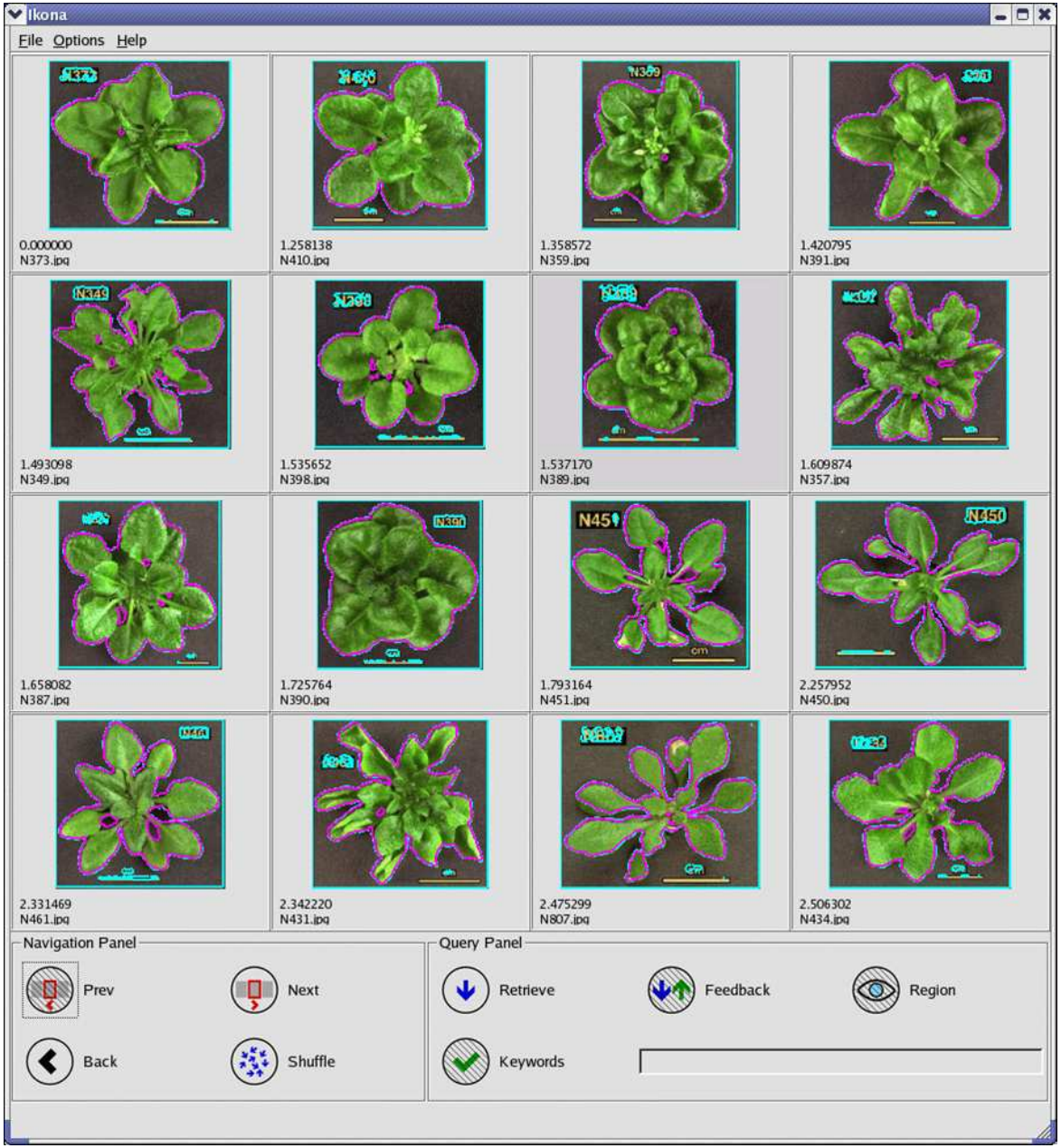
Image Biotim – recherche par similarité

Image Biotim – recherche par similarité

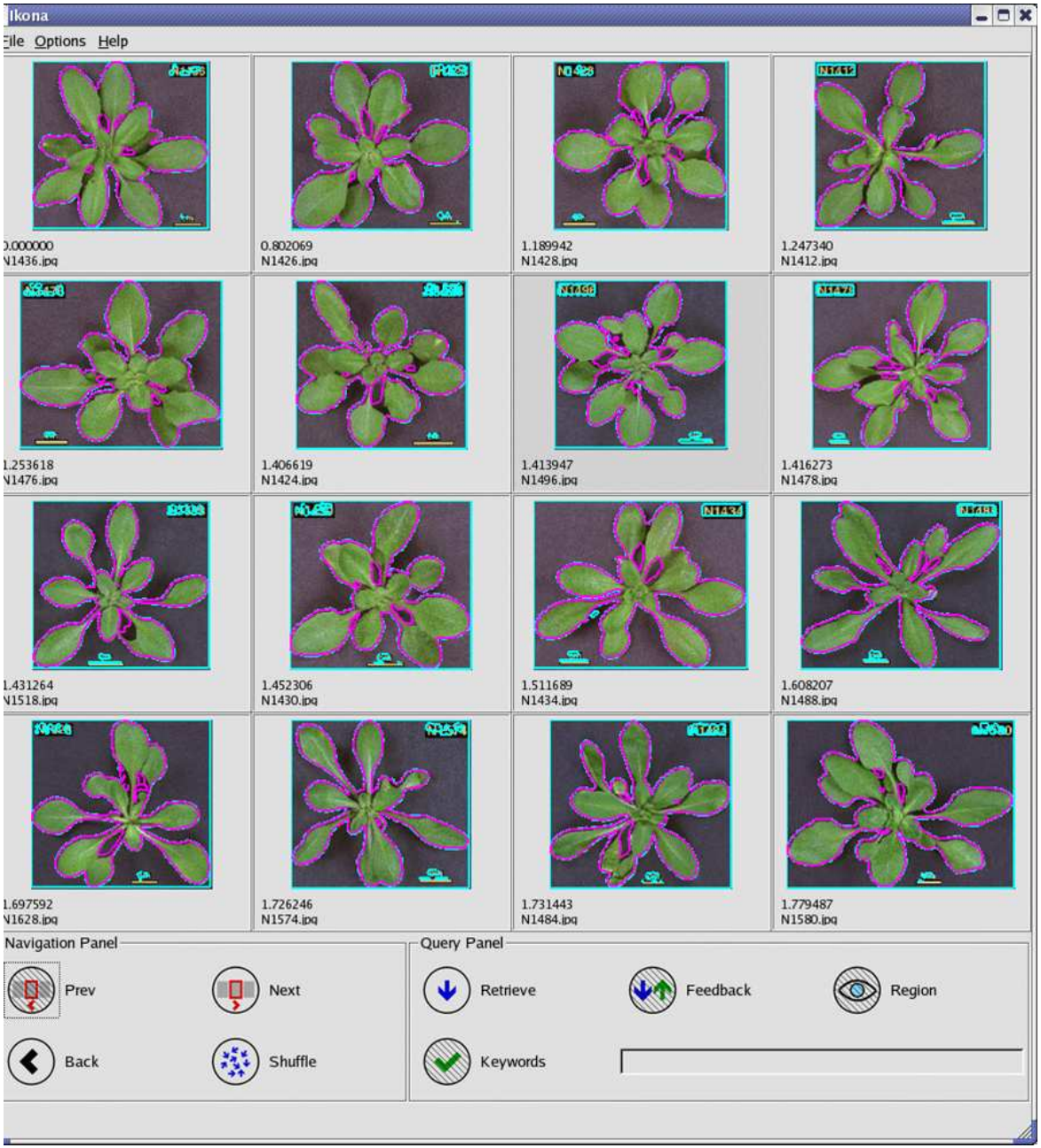
Expression génique – tirage aléatoire



GE – recherche par similarité sur la forme

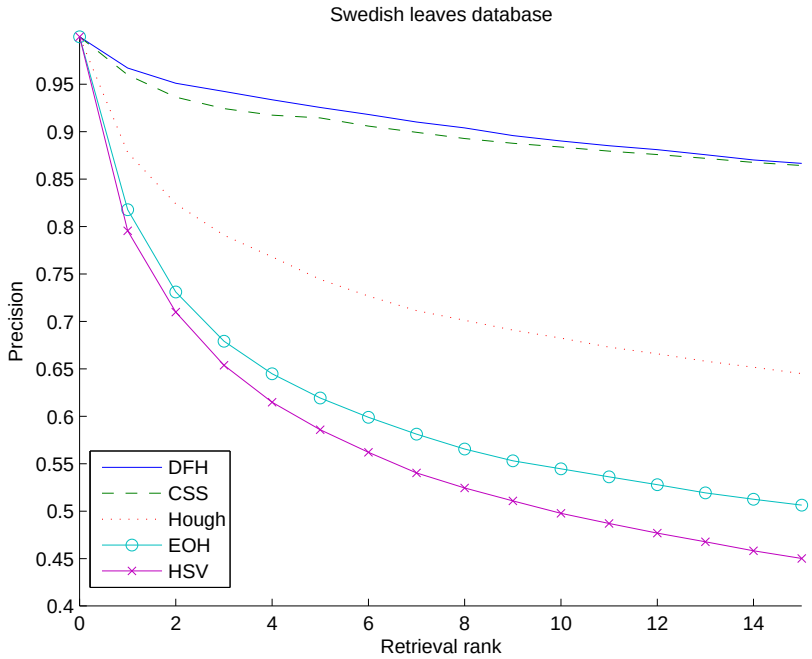


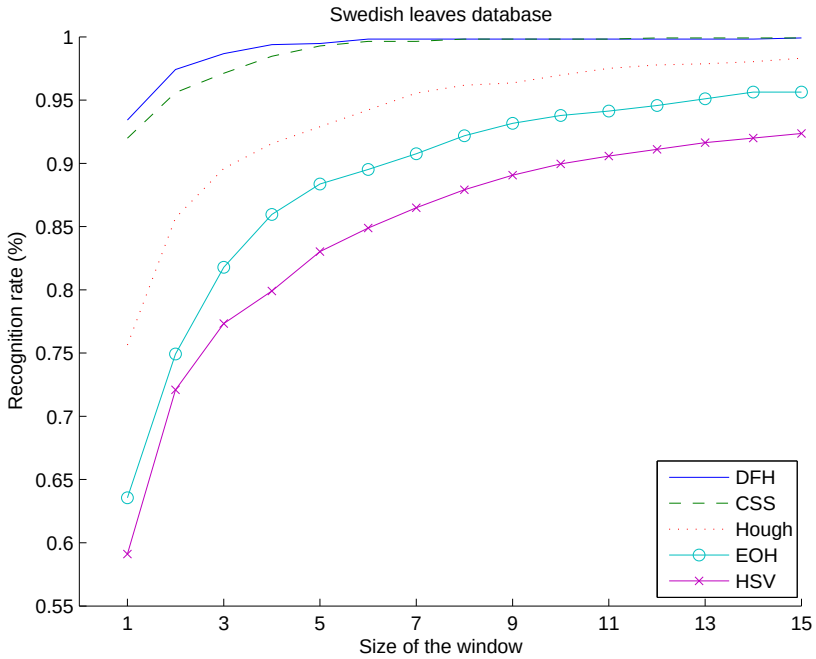
GE – recherche par similarité sur forme et couleur



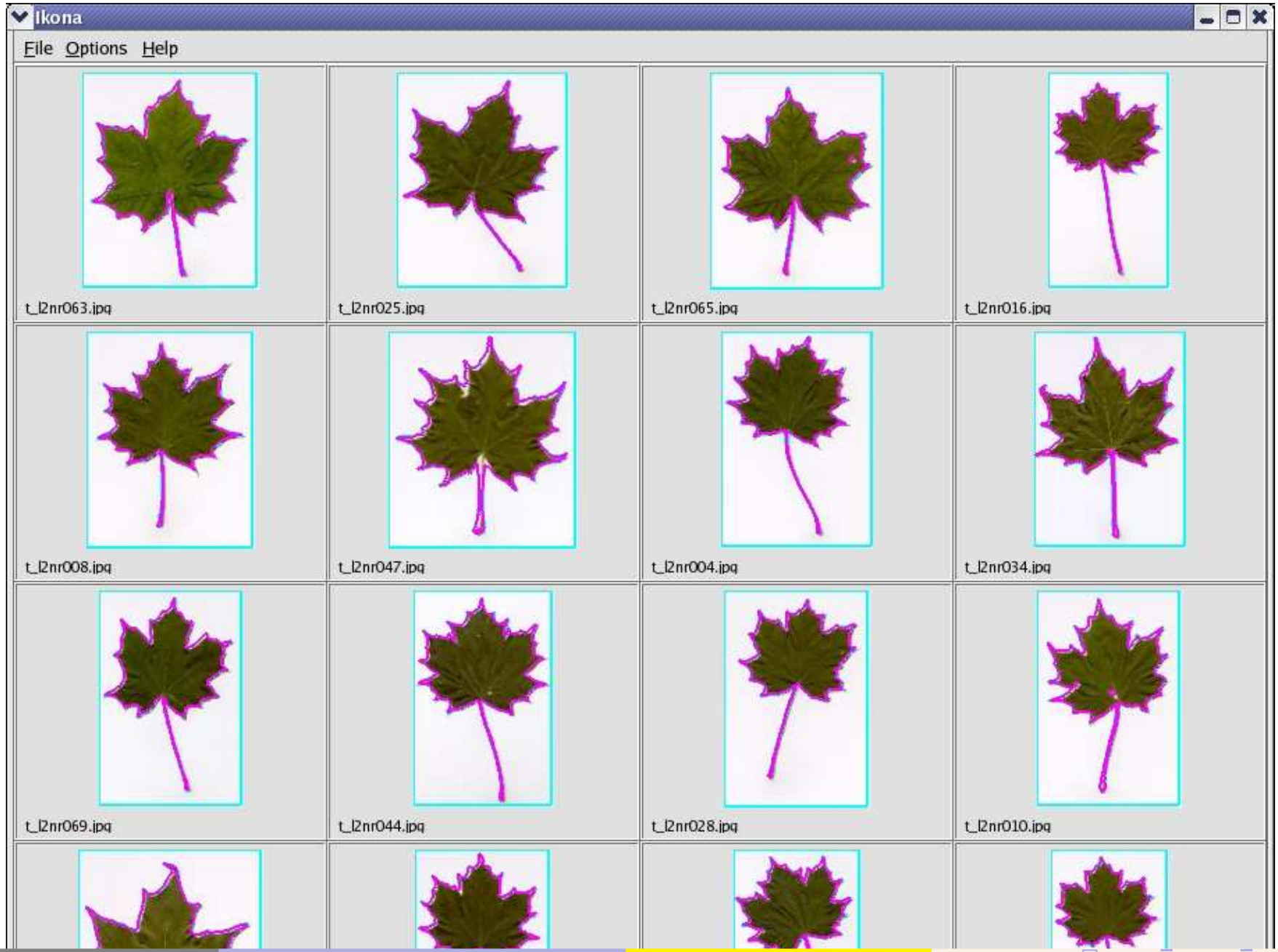
Reconnaissance de formes

Utilisation **DFH** d'un nouveau descripteur de forme, meilleur et à coût réduit



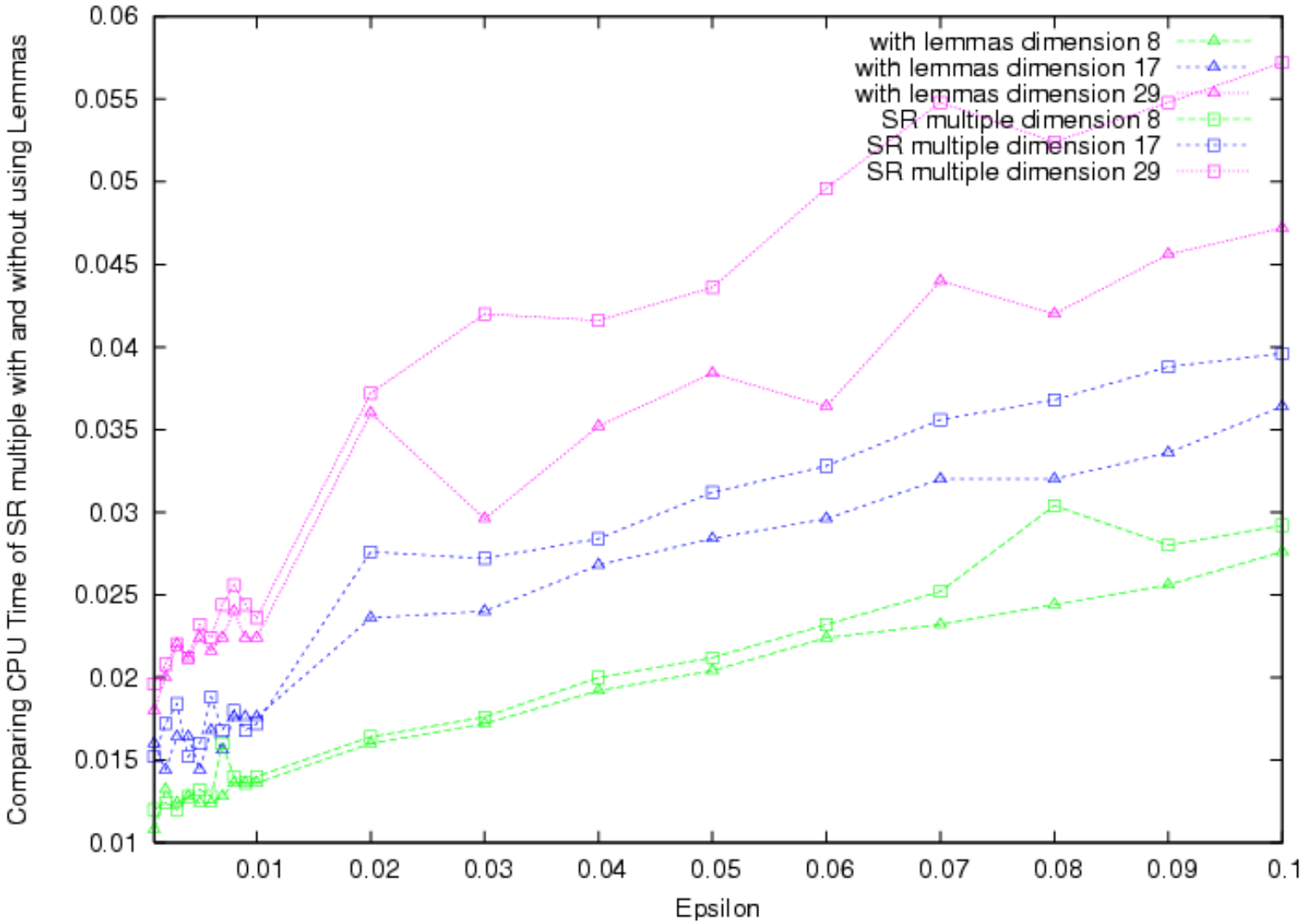


Reconnaissance de formes



Images : passage à l'échelle

Nouvelle méthode pour la recherche par point d'intérêt : minimisation E/S et calcul de distances pour SR-Tree



Visibilité

Contacts

- Muséum National d'Histoire Naturelle (MNHN, Paris) ; CIRAD (Montpellier)
- Association ENDEMIA (Nouvelle Calédonie)
- Projet Electronic Field Guide (NSF, États-Unis)
- Australian National Botanical Garden

Au moins 9 publications

Quelques URLs

- Site Biotim `http://www-rocq.inria.fr/imedia/biotim/`
- Site de travail Biotim `http://graves.inria.fr/biotim/`
- Demo Ikona `http://www-rocq.inria.fr/cgi-bin/imedia/ikona`
- Demo Parseur du Français `http://atoll.inria.fr/parserdemo`

Conclusions

- Projet représentatif de nombreux projets similaires à venir
- Difficulté d'obtenir des corpus d'images de grande taille relatif au texte
  Nécessaire pour l'apprentissage
- Difficulté de l'apprentissage de propriétés visuelles fines
  images non homogènes, qualité variable, . . .
- Difficulté des évaluations :
  - besoin d'experts (occupés)
  - problème de présentation des résultats