BIOTIM

Exploitation de Gisements
Texte-Image en Biodiversité

http://www-rocq.inria.fr/imedia/biotim/

Contexte

- Grands volumes de fonds scientifiques accumulés en biodiversité
- Grands volumes de données produits par l'expérimentation à grande échelle
- Nature de ces contenus : textes + images

Rinorea breteleri Achoundong, **sp. nov.**

Rinoreae abbreviatae Achoundong & Bos *et* R. crassifoliae (Stapf) Chipp *affinis lamina inferne glandulosa ; differt a* R. abbreviata *inflorescentia thyrsiformi non paniculata, floribus minoribus, androecio minus zygomorpho, antheris subparallelis et connectivo lineari basi non dilatato ;* R. crassifoliae *affinis statura parva et inflorescentia thyrsiformi, sed differt ab lamina foliі persicue ellіptica non ovata et fructu ovoideo non fusiformi.*

TYPUS. — *F.J. & B.J.M. Breteler 15610,* Gabon, 35 km Nord-Est de Lastoursville, fl., fr., 14 juil. 2000 (holo-, WAG ; iso-, P, K).

Arbuste atteignant 2,5 m de hauteur, à appareil végétatif glabre. Stipules triangulaires, c. 5 mm de longueur, portant des nervures longitudinales saillantes, rapidement caduques, laissant sur la tige des cicatrices très nettes. Pétiole de (0,5-)1-2(-3) cm de longueur, ridé, canaliculé dessus, faiblement épaissi aux 2 extrémités ; limbe coriace, elliptique, généralement 1,8 à 2,8 fois plus long que large, de (11,5-)13-24(26) × (4-)5-9(-10) cm, cunéiforme à obtuse ± brusquement acuminé à l'apex; acumen de (0,6-)1-2(-3) cm de long, mucroné ; nervures c. 0,3 mm de long ; anthère c. 1,6-2 mm de longueur, appendice du connectif c. 1-1,5 mm de longueur, triangulaire à largement ovale, décurrent le long de l'anthère jusqu'au tiers inférieur, rouge foncé ; thèques à appendice apical commun, largement ovale, cochléariforme, rouge. Ovaire lagéniforme, c. 1 × 1 mm, ovules 6 ; style long c. 2,6 mm, avec un renflement annulaire vers le sommet. Capsule ovoïde-trilobée, c. 20-25 × 15 mm, coriace à subligneuse, verte à l'état frais ; graines anguleuses, subtriquètres, c. 5 × 5 mm, jaune paille, à fines marbrures plus foncées. — Fig. 1.

HABITAT ET DISTRIBUTION. — Forêt dense humide du centre du Gabon, jusqu'à 380 m d'altitude.

PARATYPES. — GABON : *Breteler 6683,* 8 km sur la route Lastoursville-Koula Moutou, 0°51'S, 12°39'E, fl., 28 sep. 1970 (WAG) ; *Breteler & J.J. de Wilde 461,* 10 km sur la route Lalara à Makokou, fl., 6 sep. 1978 (P, WAG) ; *Breteler & J.J. de Wilde 497,* 10 km sur la route Lalara à Makokou, fl., 8 sep. 1978 (K, P, WAG) ; *F.J. & B.J.M. Breteler 12306,* Ogooué-Lolo, forêt à environ 40 km E de Lastoursville, 0°50'S, 13°08'E, fr., 23 nov. 1993 (WAG) ; *Wilks 775,* Ogooué-Ivindo, forêt des Abeilles, 15 km SSE du con-

Contexte

- Utilisation souhaitée pour ces contenus :
  - Différents modes d'interrogation
  - Synthèses, comparaisons
  - Découverte de connaissances
- Pourquoi une telle exploitation est difficile ?
  - Absence de structuration des contenus dans chacune des modalités (textes, images)
  - Relations rarement explicitées entre les modalités (textes $\leftrightarrow$ images)
  - Volumes trop élevés pour un traitement manuel

Objectifs

- Conception exploratoire de méthodes pour
  1. L'analyse et la structuration de masses de textes et de masses d'images

     ⇒ importance du choix de domaines de connaissance systématique
  2. L'acquisition d'une sur-couche sémantique commune pour l'intégration entre modalités
  3. L'interrogation pluri-modale des contenus ainsi structurés, exploitant aussi la complémentarité entre textes et images

Partenaires

- Institut de Recherche pour le Développement (IRD Orléans)
- Unité Mixte de Recherche en Génomique Végétale (INRA Évry)
- Équipe « Contraintes et Apprentissage » (LIFO – Université d'Orléans)
- Projet ATOLL (INRIA Rocquencourt)
- Projet IMEDIA (INRIA Rocquencourt) – coordination
- Équipe VERTIGO (CEDRIC – CNAM)

Composantes du travail

- Préparation et mise en forme des corpus
- Acquisition de l'ontologie textuelle
- Construction d’une base de connaissances
- Mise au point de descripteurs visuels spécifiques
- Méthodes d'association texte – image
- Navigation et interrogation bi-modale

Texte : avancement

- Phases préparatoires en cours d'achèvement
  - Examen des corpus
  - Correction des erreurs OCR (automates finis) : taux d'erreur important dans les documents numérisés
  - Structuration logique des corpus (~ « chunking ») : retrouver la structure logique des documents
  - Chaîne morpho-syntaxique (pipeline XML) : segmentation, consultation de lexique, étiquetage, entités nommées
  - Extraction terminologique (ACABIT et FASTR) : obtenir termes du domaine et quelques informations sur les liens entre mots
- Analyse syntaxique prochaine sur les corpus (DyALog) – phase de départ pour
  1. L'extraction d'ontologie
  2. L'extraction de connaissances par taxon

Texte : résultats des phases préparatoires

Erreurs	Correction
pubescences	pubescentes
généralemint	généralement
lerbier	herbier
nigerian	Nigerian
lordu	tordu

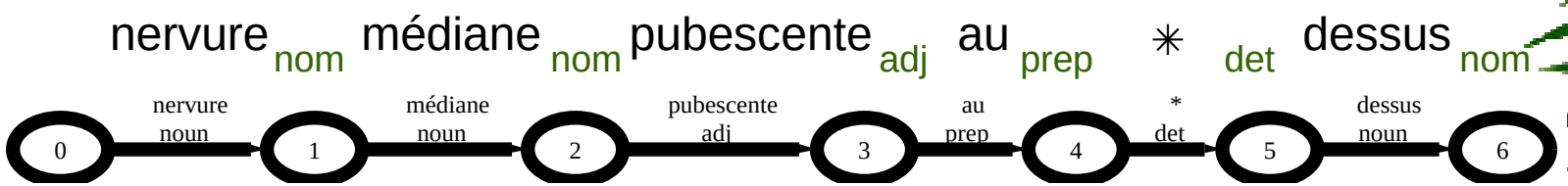
(a) Correction OCR

Err.	Corr.
ß	B
6	é
i	l
l	t

(b) Table de correction

Term	Loglike
m de longueur	1735
De Wild	1731
paire de folioles	1669
nervure latérale	1661
paire de nervures	1661
forêt dense	1617
forêt dense humide	1598

(c) Extraction terminologique

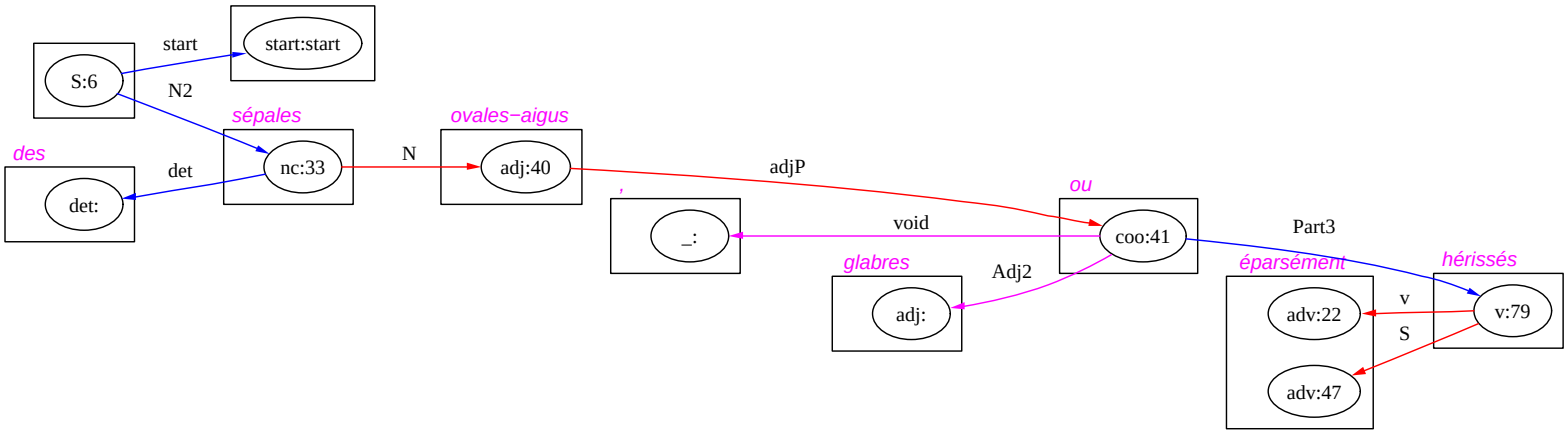


(d) Traitement morpho-syntaxique

Traitements syntaxiques

Motivation : fournir des dépendances syntaxiques entre mots traduisant des relations sémantiques ⇒ base pour

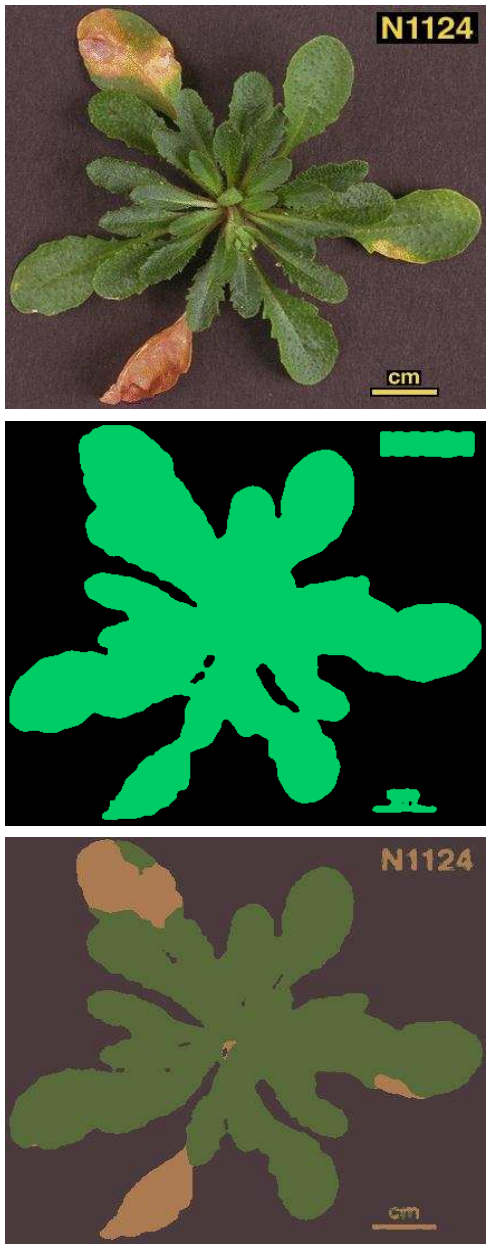
1. L'extraction d'une ontologie
   [adj lancéolé : forme caractérisant les feuilles]
2. Pour chaque taxon, l'extraction de ses propriétés
   [forme(taxon_X, feuilles) = lancéolée]

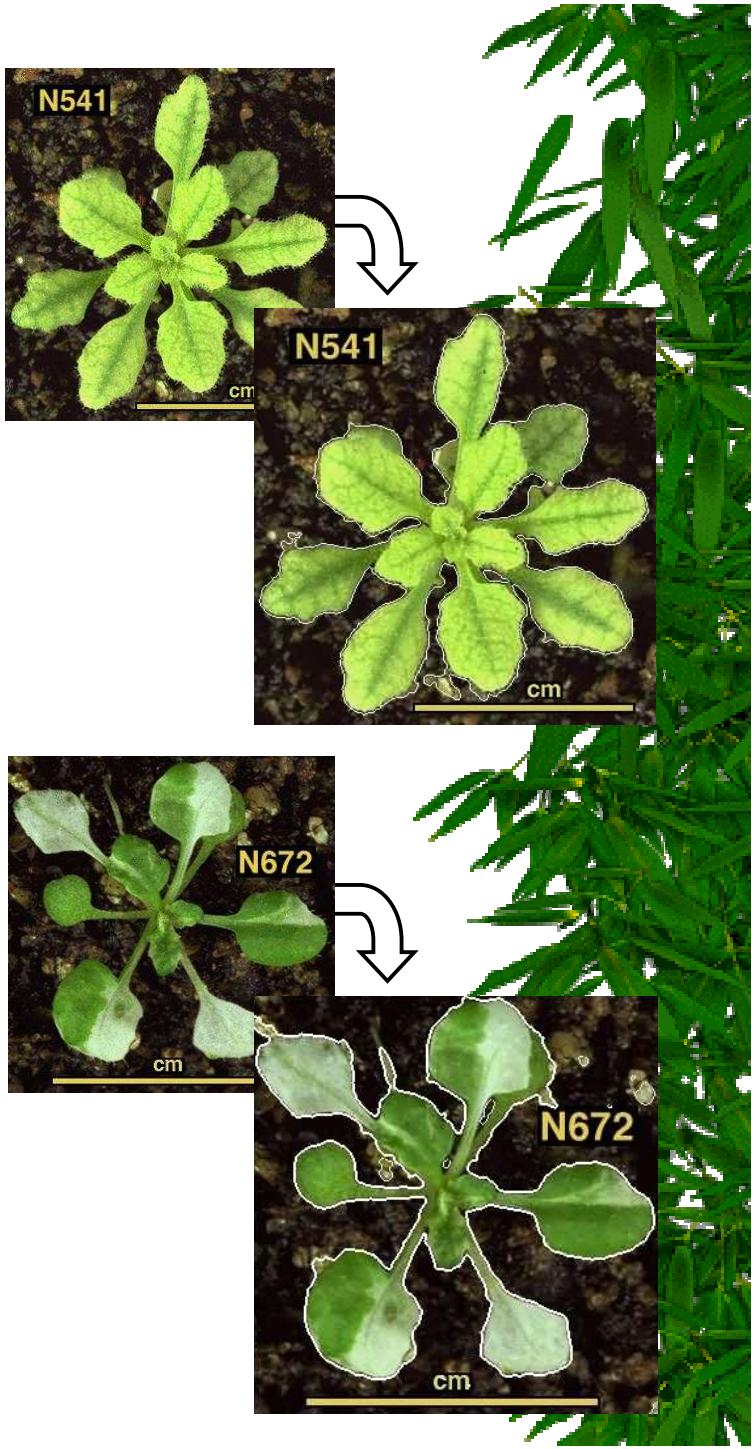


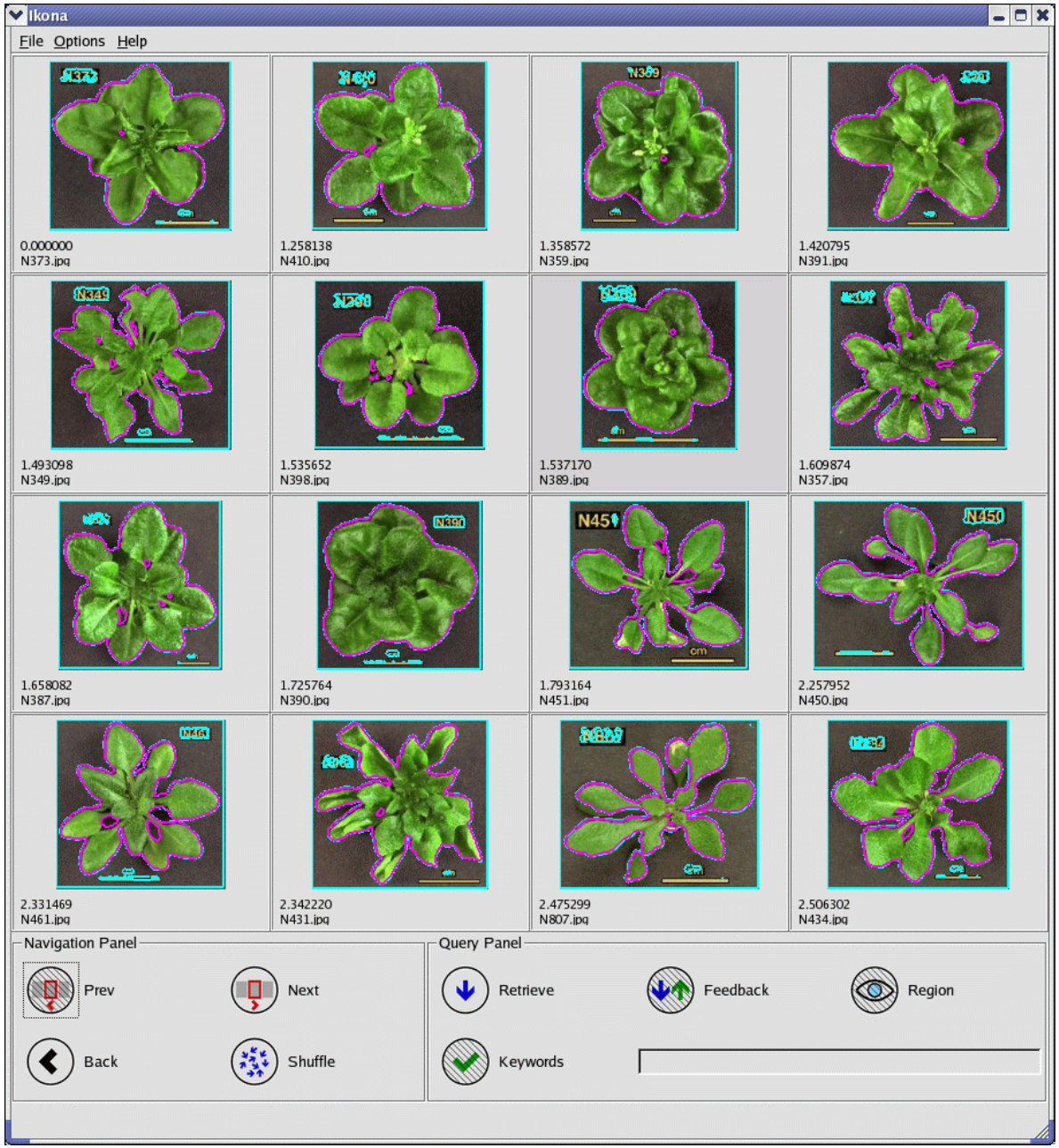
Dépendances pour « des sépales ovales-aigus, glabres ou éparsément hérissés »

Description apparence visuelle : avancement

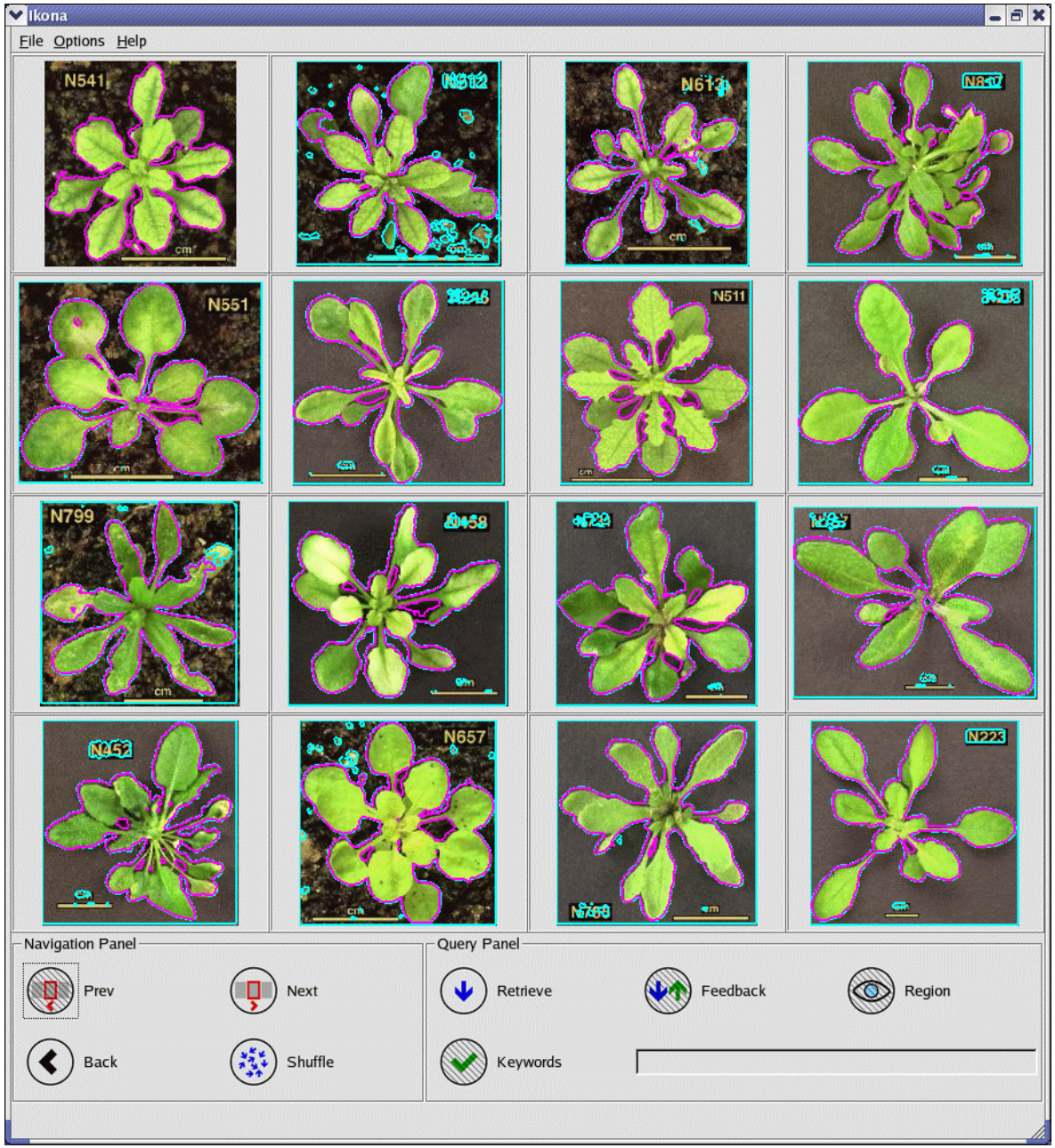
- Nouvelles méthodes de segmentation par classification
  1. Extraction du masque de la plante avec résolution des problèmes : fond hétérogène, ombre, nuances multiples, unicité
  2. Mesures quantitatives : par exemple, étendue relative du changement de couleur
- Nouveau descripteur de forme : caractériser le contour extérieur de la plante
  - Invariance à : rotation, changement échelle, translation


N1124
cm
N813
cm
N1124
cm
N813
cm


N541
cm
N541
cm
N672
cm
N672
cm



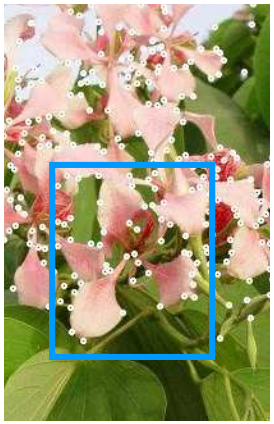
Exemple de requête partielle

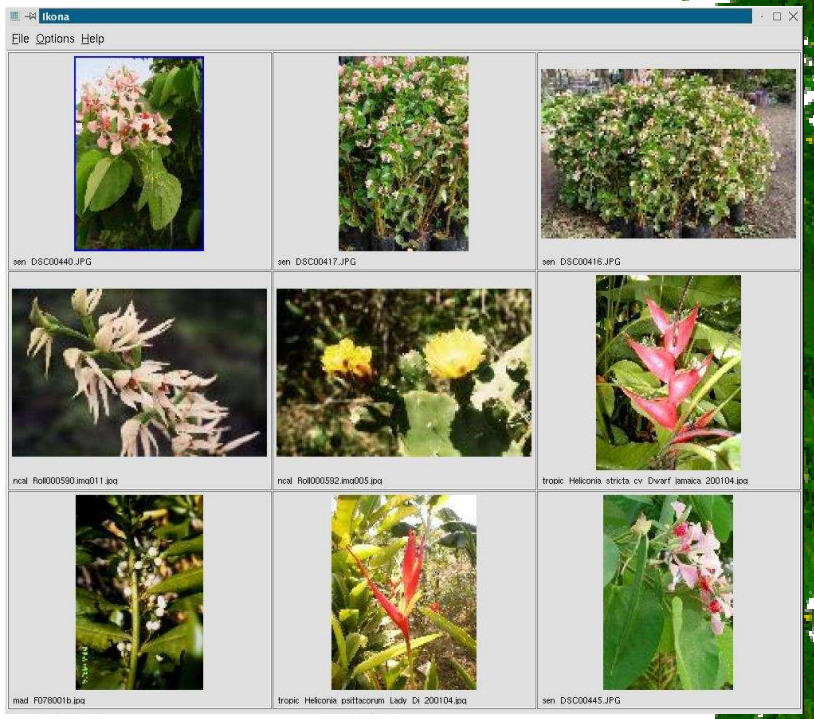


Exemple de requête partielle

Description locale

Exemple de requête





Configuration de points d'intérêt

Optimisation des accès

Indexation par points d'intérêt :

60 000 images $\rightarrow$ 15 millions points

$\Rightarrow$ Nécessité d'accélérer la recherche par similarité

- Techniques évaluées : SR-tree, VA-file
- Résultats obtenus (SR-Tree) : gain moyen de 10 fois en dimension 8 et 17, temps de réponse moyen de 3 secondes
- Recherche en cours : traiter simultanément (et non séquentiellement) tous les points de la requête (gain escompté : 3 à 5 fois)

Agenda

- Analyse et structuration
  - Texte :
    - Finalisation structuration logique et traitement entités nommées, démarrage analyse syntaxique (2004)
    - Apprentissage sur ensembles de dépendances, extraction d'ontologie (2005)
    - Extraction de base de connaissances (2005-2006)
  - Image :
    - Finalisation définitions nouveaux descripteurs (2004)
    - Évaluation des ensembles de descripteurs (2005)
    - Poursuite de l'optimisation des accès (2005-2006)

Agenda

- Acquisition de sur-couche sémantique commune
  - Apprentissage relations texte – images (2005)
- Navigation et interrogation pluri-modale
  - Recherche avec retour de pertinence, résumés visuels (2004-2005)
  - Langage de requête hybride similarité visuelle + configuration spatiale 2D (2005-2006)*

* Travail conditionné par l'obtention de ressources supplémentaires, hors BIOTIM