# Outils d'exploration diachronique des collections ISTEX (ISTEX-R WP1)

*Jean-Charles LAMIREL (Synalp - LORIA)*

*Nicolas Dugué (Synalp - LORIA)*

*Pascal Cuxac (INIST – CNRS)*

**Financement : ANR-10-IDEX-0004-02**

**Séminaire CU-ISTEX-2016**

# Objectifs du WP1 ISTEX-R
## Outils et interface d'analyse diachronique

- Mettre en place des méthodes automatisées permettant de pister les sujets et leur changements dans un ensemble de publications extraites du réservoir ISTEX, à partir d'une recherche générale.
- Plusieurs approches à mettre en œuvre :
  - Analyses par pas de temps.
  - Analyse à grain fin.
- Proposer à l'utilisateur une interface conviviale et exploitable permettant de visualiser les évolutions de manière explicite.
- Un première étape du travail se focalise sur l'analyse par pas de temps.

# Challenges spécifiques à l'analyse diachronique et aux données ISTEX

- Les données ISTEX sont issues de différents éditeurs et il n'existe pas de standardisation des métadonnées (quand celles-ci sont disponibles …).
- Les méthodes utilisées doivent pouvoir fonctionner sur des collections de taille importante de manière non supervisée (contraintes de temps de calcul et limitation voire affranchissement des paramètres).
- Les taille des périodes qui englobent des sujets stables peut varier.
- Il reste des problèmes ouverts à résoudre :
  - Suivi de sujets.
  - Recherche de modèle optimal en apprentissage non supervisé.
  - Visualisation des changements diachroniques.

# Solutions apportées

- Adaptation de travaux sur l'analyse diachronique et le suivi de sujets (concurrents des méthodes de type LDA …) et exploitation de nouvelles métriques développés dans l'équipe Synalp.
- Mise en place de nouvelles méthodes fiables d'analyse de la qualité du clustering.
- Mise en place de nouveaux modèles de visualisation tels que les graphes de contrastes.
- Implantation et adaptation de méthodes de visualisation avancées pour la diachronie.
- Expérimentation de techniques d'extraction automatique de métadonnées (et de résumé automatique).
- Expérimentation-test menée sur premier ensemble de test de ~10000 publications du réservoir ISTEX relatives à la gérontologie (commun aux partenaires ISTEX-R).
- Mise en place de l'interface utilisateur opérationnelle Diachronic'Explorer.

# Vue générale de la méthodologie

Requête générale

Sous-collection

Indexation (Métadonnées *)

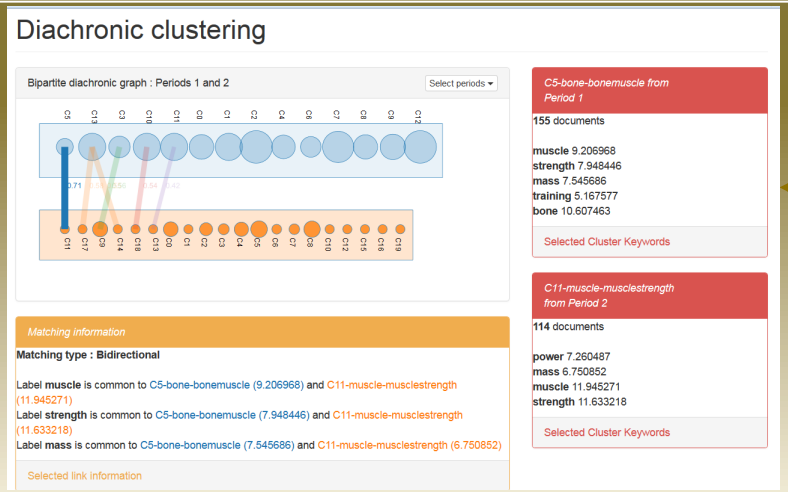
Détection automatique de méta-périodes

Clustering + Détection de modèle optimal

Sélection de traits saillants

Réservoir ISTEX

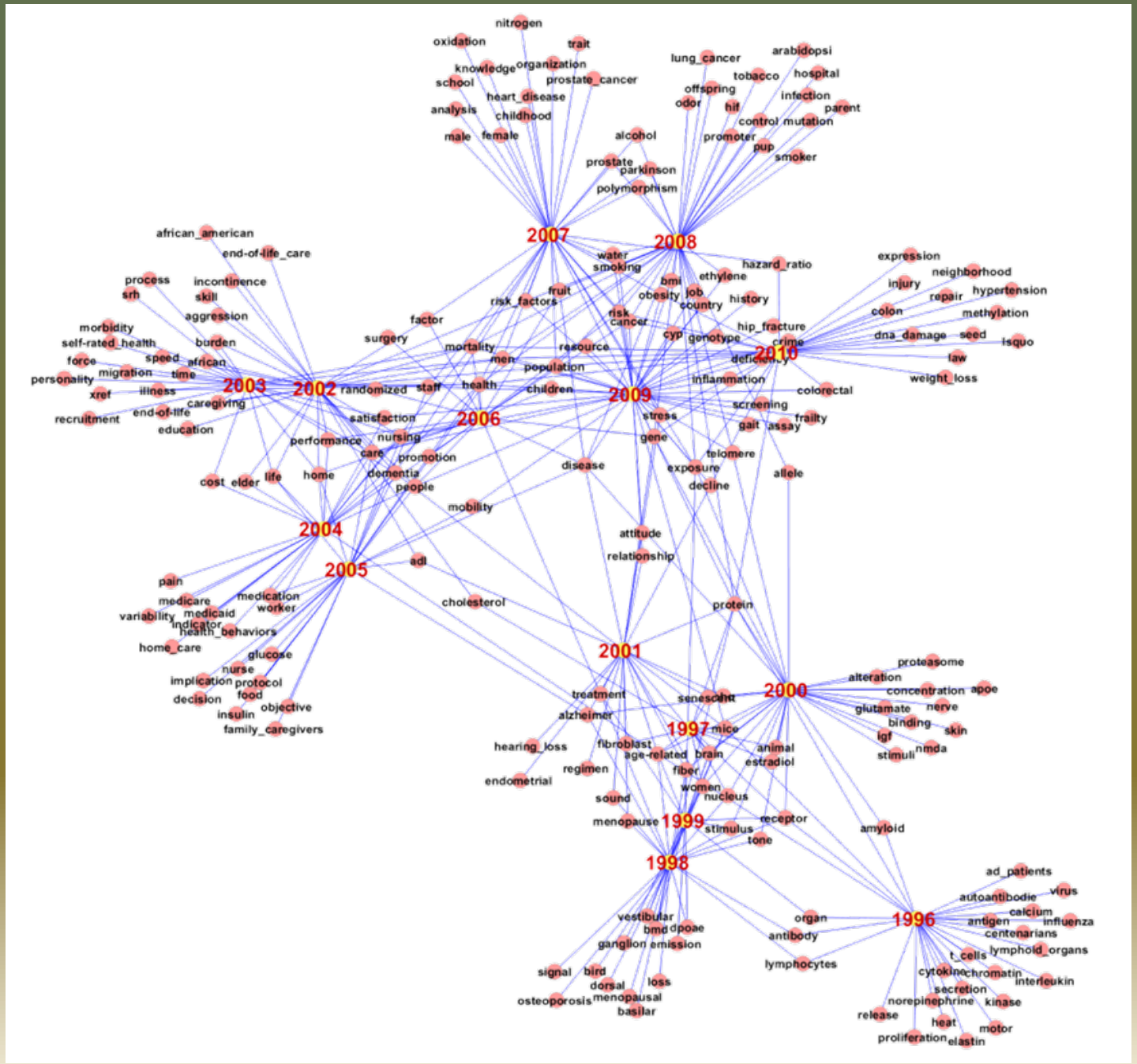
**Analyse diachronique et visualisations : Interface Diachronic'Explorer**




* = en cours

Graphe de contraste :

Graphe mots-années filtré fournissant une représentation lisible.



# Contacts et questions

En perspective : analyse à grain fin, extraction automatique de métadonnées, complément de méthodes en visualisation.

Email(s) :
lamirel@loria.fr, nicolas.dugue@loria.fr, Pascal.Cuxac@inist.fr

Rapport et 7 publications liés au WP1 disponibles sur demande.

Demo: http://www.github.com/nicolasdugue

## Quelques références :

1) Lamirel J.-C. : **A new diachronic methodology for automatizing the analysis of research topics dynamics** : an example of application on optoelectronics research, Scientometrics 93(1): 151-166 (2012).

2) Lamirel J.C., Cuxac P., Chivukula A.S., Hajlaoui K. : **Optimizing text classification through efficient feature selection based on quality metric**. *Journal of Intelligent Information Systems*, May 2014, p.1-18, Springer.

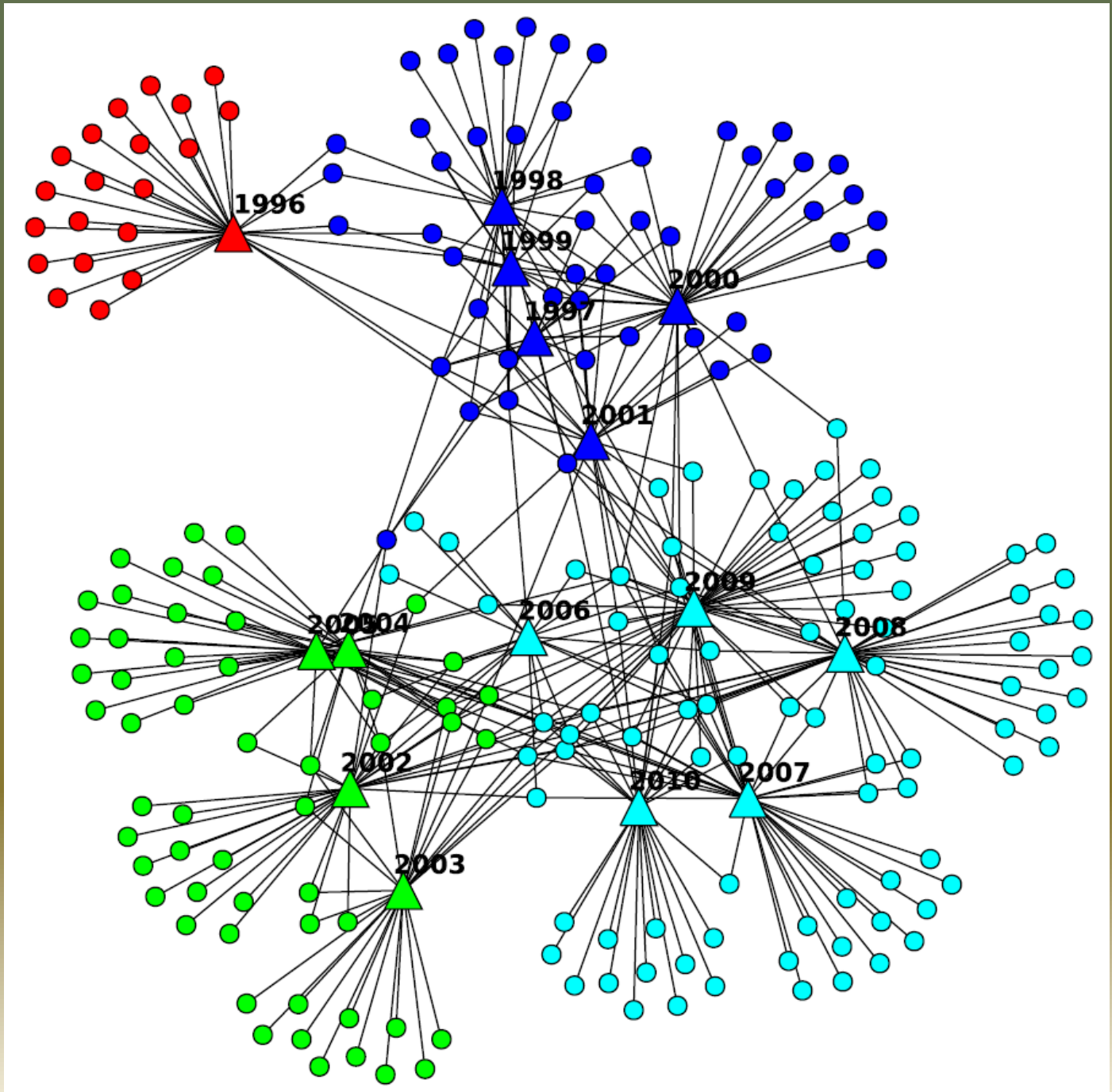
3) Lamirel J.-C., Cuxac P. : **New quality indexes for optimal clustering model identification with high dimensional data**, Proceedings of ICDM-HDM'15 - International Workshop on High Dimensional Data Mining, Atlantic City, USA, November 2015.

Place à la démo…

# Détection automatique de méta-périodes

- Depuis la sous-collection globale et sur la base d'une matrice (mots x années), la sélection de variables basées sur la maximisation des traits est opérée et un graphe de contraste est construit.
- Les méthodes de l'état de l'art en partionnement de graphe sont expérimentées sur ce dernier :
  - FastGreedy
  - SpinGlass
  - MCL
  - **Walktrap ****
  - …

Partionnement automatique en méta-périodes exploitant le principe de marche aléatoire sur le graphe de contraste.



# Clustering et détection des changements

- Beaucoup de méthodes de clustering ont été testées :
  - K-means
  - SOM
  - **GNG ****
  - IGNGF
  - …..
- Le modèle optimal de chaque période est extrait en exploitant la métrique de maximisation des traits.
- La détection des changements est opérée en utilisant le raisonnement bayésien non supervisé basé sur les traits et les mesures de contraste.

# Clustering et détection des changements

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | *1.610* | -1.521- | 75 | 37.50 | 4.992 | 75 | 37.50 | -2.008- |
| 3 | 3 | *2.046* | -1.943- | 292 | 97.33 | 19.311 | 446 | 148.67 | -8.549- |
| 4 | 4 | *2.366* | -2.121- | 361 | 90.25 | 18.449 | 675 | 168.75 | -9.754- |
| 5 | 5 | *2.649* | -2.232- | 416 | 83.20 | 17.549 | 879 | 175.80 | -10.246- |
| 6 | 6 | *2.851* | -2.326- | 492 | 82.00 | 14.304 | 1087 | 181.17 | -10.572- |
| 7 | 7 | *3.123* | -2.408- | 548 | 78.29 | 14.463 | 1302 | 186.00 | -10.892- |
| 8 | 8 | *3.404* | -2.476- | 620 | 77.50 | 14.587 | 1517 | 189.62 | -11.073- |
| 9 | 9 | *3.542* | -2.502- | 685 | 76.11 | 16.266 | 1708 | 189.78 | -12.326- |
| **10** | **10** | ***3.749*** | **-2.548-** | **756** | **75.60** | **17.705** | **1905** | **190.50** | **-13.399-** |
| 11 | 11 | *3.884* | -2.509- | 837 | 76.09 | 16.236 | 2119 | 192.64 | -12.349- |
| 12 | 12 | *4.117* | -2.577- | 881 | 73.42 | 15.855 | 2348 | 195.67 | -12.232- |
| 13 | 13 | *4.187* | -2.573- | 952 | 73.23 | 13.088 | 2536 | 195.08 | -10.218- |
| 14 | 14 | *4.425* | -2.592- | 1000 | 71.43 | 14.289 | 2741 | 195.79 | -11.162- |
| 15 | 15 | *4.537* | -2.630- | 1068 | 71.20 | 14.988 | 2917 | 194.47 | -11.676- |
| 16 | 16 | *4.740* | -2.668- | 1090 | 68.12 | 14.085 | 3112 | 194.50 | -11.124- |
| 17 | 17 | *5.010* | -2.707- | 1171 | 68.88 | 13.397 | 3439 | 202.29 | -10.682- |
| 18 | 18 | *5.124* | -2.694- | 1248 | 69.33 | 12.076 | 3623 | 201.28 | -9.672- |
| 19 | 19 | *5.223* | -2.726- | 1282 | 67.47 | 12.251 | 3783 | 199.11 | -9.840- |
| 20 | 20 | *5.240* | -2.694- | 1383 | 69.15 | 12.453 | 3937 | 196.85 | -9.916- |
| 21 | 21 | *5.509* | -2.710- | 1450 | 69.05 | 12.232 | 4186 | 199.33 | -9.782- |
| 22 | 22 | *5.582* | -2.739- | 1489 | 67.68 | 12.871 | 4327 | 196.68 | -10.277- |
| 23 | 23 | *5.683* | -2.754- | 1519 | 66.04 | 12.318 | 4507 | 195.96 | -9.907- |
| 24 | 24 | *5.685* | -2.708- | 1600 | 66.67 | 12.697 | 4690 | 195.42 | -10.156- |
| 25 | 25 | *5.825* | -2.744- | 1671 | 66.84 | 12.597 | 4866 | 194.64 | -10.079- |

Le modèle optimal de chaque période est caractérisé et le traits dont le contraste est supérieur à la moyenne dans les clusters sont isolés.

**5.417783 muscle**
**5.327384 strength**
**5.217635 exercise**
**4.168652 power**
**3.841000 mass**
**3.776102 training**
**2.814531 body**
**2.502115 performance**
**********
**********
**2.466132 week**
**2.371438 weight**
**1.897246 participant**
**1.857157 gerontol**
**1.775870 month**
**1.752784 woman**
**1.688477 measurement**
**1.677586 baseline**
**1.648806 biol**
**1.637920 function**
**1.625222 test**
**1.599973 decrease**
**1.586033 animal**
**1.529848 colleague**
**1.528399 age-related**
**1.489449 disability**
**1.482194 group**
**1.458563 adult**
**1.434127 change**
**1.403684 increase**

# Clustering et détection des changements

→ ?

| | |
|---|---|
| 2 | 2.103293 |
| 3 | 4.894217 |
| 4 | 4.661046 |
| 5 | 4.283290 |
| 6 | 4.116885 |
| 7 | 4.485136 |
| 8 | 4.229456 |
| 9 | 4.093945 |
| 10 | 4.053818 |
| 11 | 4.025391 |
| 12 | 4.059147 |
| 13 | 3.866609 |
| 14 | 3.577087 |
| 15 | 3.776318 |
| 16 | 3.537132 |
| 17 | 3.608063 |
| 18 | 3.642788 |
| 19 | 3.514834 |
| 20 | 3.430515 |
| 21 | 3.327520 |
| 22 | 3.456252 |
| 23 | 3.239336 |
| 24 | 3.192178 |
| 25 | 3.271146 |

→ ?

**Indice de Davis-Bouldin** (se comporte de manière convexe sur l'intervalle d'analyse….)

?

**Termsuite** (fournit principalement des termes généraux ….)

```
4.724185 americans
4.350677 gap
4.307484 cell
4.244061 circumstances
4.232458 health_condition
4.044825 instrument
3.659395 future
3.653448 threshold
3.189270 orientation
2.608994 problem
2.606182 frame
2.476067 protocol
2.188425 quantitative
**********
**********
2.042368 code
1.885890 absence
1.830976 admission
1.793681 term
1.759934 allele
1.686220 flexor
1.550164 mellitus
1.520950 fast
1.503231 flow
1.472898 medium
1.436958 scores
1.430196 editor
1.419317 paper
1.403544 notion
1.368368 understanding
1.295318 compute
1.262428 glucose
1.193671 church
1.159139 home_resident
1.138192 sign
1.091506 technology
1.076663 cope
1.036130 transportation
1.032563 hospitalize
1.013571 writing
1.001360 public_health
```

Une mauvaise indexation ainsi que de mauvais indices de qualité produisent des résultats ingérables.

# Perspectives : extraction des métadonnées et production de résumés automatiques

Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD, Radvan Urbanek, MD, and Johannes Forster, MD
From University Children's Hospitals, Freiburg, Germany, and Vienna, Austria

**To investigate the risk of life-threatening reactions to future stings, we sequentially challenged 113 children (aged 2 to 17 years) allergic to insect stings with a sting by the relevant insect. The time interval between the challenges varied from 2 to 6 weeks. The history of the index stings was a large local reaction (LR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal LR, 11% a large LR, and 13% an SR. On the second challenge, 78% of the children had a normal LR, 5% a large LR, and 17% an SR. Thirty-nine of the untreated children were exposed to a field sting during the subsequent 3-year follow-up period. In comparison with other diagnostic evaluations such as skin-prick tests, determinations of specific IgE and IgG antibodies, and single-sting exposure, the dual sting challenge scheme appears to be the best predictor of reactions to subsequent stings. It also appears to be helpful in selecting patients with an uncertain sensitization status for venom immunotherapy. (J Pediatr 1995;126:185-90)**

In childhood, allergy to Hymenoptera venom is mainly caused by stings of honeybees and wasps. In Europe, yellow jackets are known as "wasps," whereas in the United States, Polistes wasps are known as "wasps."[1] Between 0.4% and 4% of the population have systemic allergic reactions to insect stings.[2-4] The incidence of systemic reactions to subsequent stings is lower in children and adolescents than in adults.[5-8] Prospective observations of the natural course of insect allergy show that adults have a risk of 27% to 57%[3, 9-11] of having repeated systemic allergic reactions, in comparison with a risk of 10% to 20% in children.[4-6, 8] Therefore venom immunotherapy should be indicated less frequently in children.[8] In vitro assays and risk scores provide only limited help in identifying those patients at risk of having further life-threatening allergic reactions. Numerous studies[12-15] have been unsuccessful in showing a correlation between the standard diagnostic methods—mainly skin-prick tests and measurements of specific IgE and IgG

Submitted for publication April 15, 1994; accepted Aug. 10, 1994.
Reprint requests: Johannes Forster, MD, University Children's Hospital, Mathildenstr. 1, D-79106 Freiburg, Germany.

0022-3476/95/$3.00 + 0 9/20/59779

Le pouvoir de synthèse de la métrique F-max est très important car elle permet de mettre en évidence la structure d'un texte (et ses métadonnées descriptives) par un mécanisme simple de compétition de blocs.