Le projet CADERIGE

Période 2000 - 2001

- **Objectifs du projet**

  Informatique développer de nouvelles techniques d'extraction de connaissances dans les bases documentaires scientifiques

  Biologique appliquer ces techniques au domaine de la génomique fonctionnelle notamment pour la modélisation des interactions géniques.

- **Laboratoires participants ...**

  Projet AÏDA de l'IRISA, UMR 6074, Rennes.
  Laboratoire LEIBNIZ de l'IMAG, Grenoble
  Laboratoire LIPN, UPRESA 7030, Université de Villetaneuse Paris 13.
  Laboratoire LRI, UMR 8623, Université d'Orsay Paris 11.
  Laboratoire MIG, UR INRA 1077, Versailles.
  Laboratoire GM, UR INRA 895, Jouy-en-Josas.
  Laboratoire LRV, UR INRA 309, Jouy-en-Josas.

Le problème

- **Fragment** : “ Transcriptional studies showed that nadE is strongly induced in response to heat, ethanol and salt stress or after starvation for glucose in a sigma B-dependent manner. Two promoters are involved during his transcriptional initiation, the sigma A-dependent upstream promoter contributes to the basal level during growth, whereas the sigma B dependent downstream promoter is induced after different stress conditions. ”

- **Comment extraire automatiquement les informations sur la régulation ?**

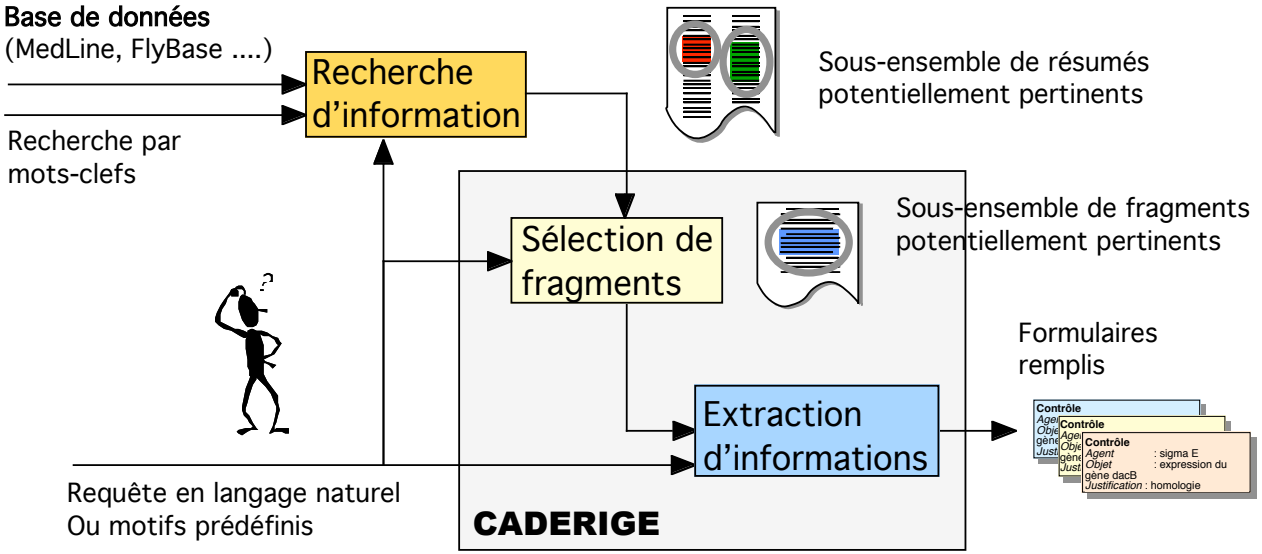
  - Recherche de « pattern » spécifiques (perl) : Précision ++ ; Rappel --
  - Critère statistiques généraux : Rappel ++ ; Précision --

- **Utilisation d’outils linguistiques pour une analyse plus profonde**

  - Absence de nomenclature des gènes, ni de règles de nommage
  - Compréhension profonde du texte basée sur un étiquetage sémantique
  - Présence d’anaphores (le sujet est énoncé dans une phrase précédente)

Extraction d'informations

- **Mettre en place un mécanisme d'extraction d'informations**



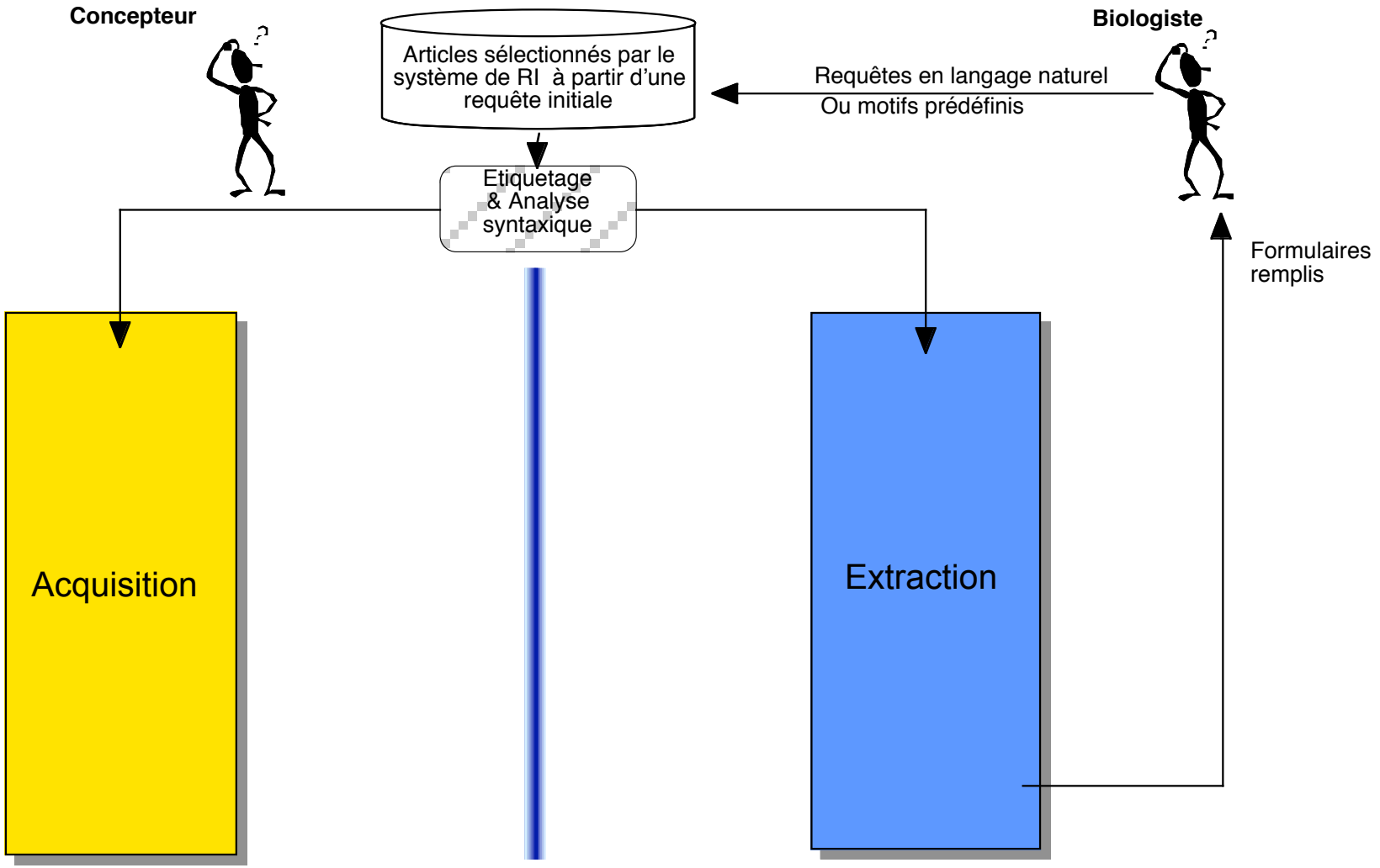
- **Acquérir des connaissances structurées**

Previously, it was shown that **the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK**.

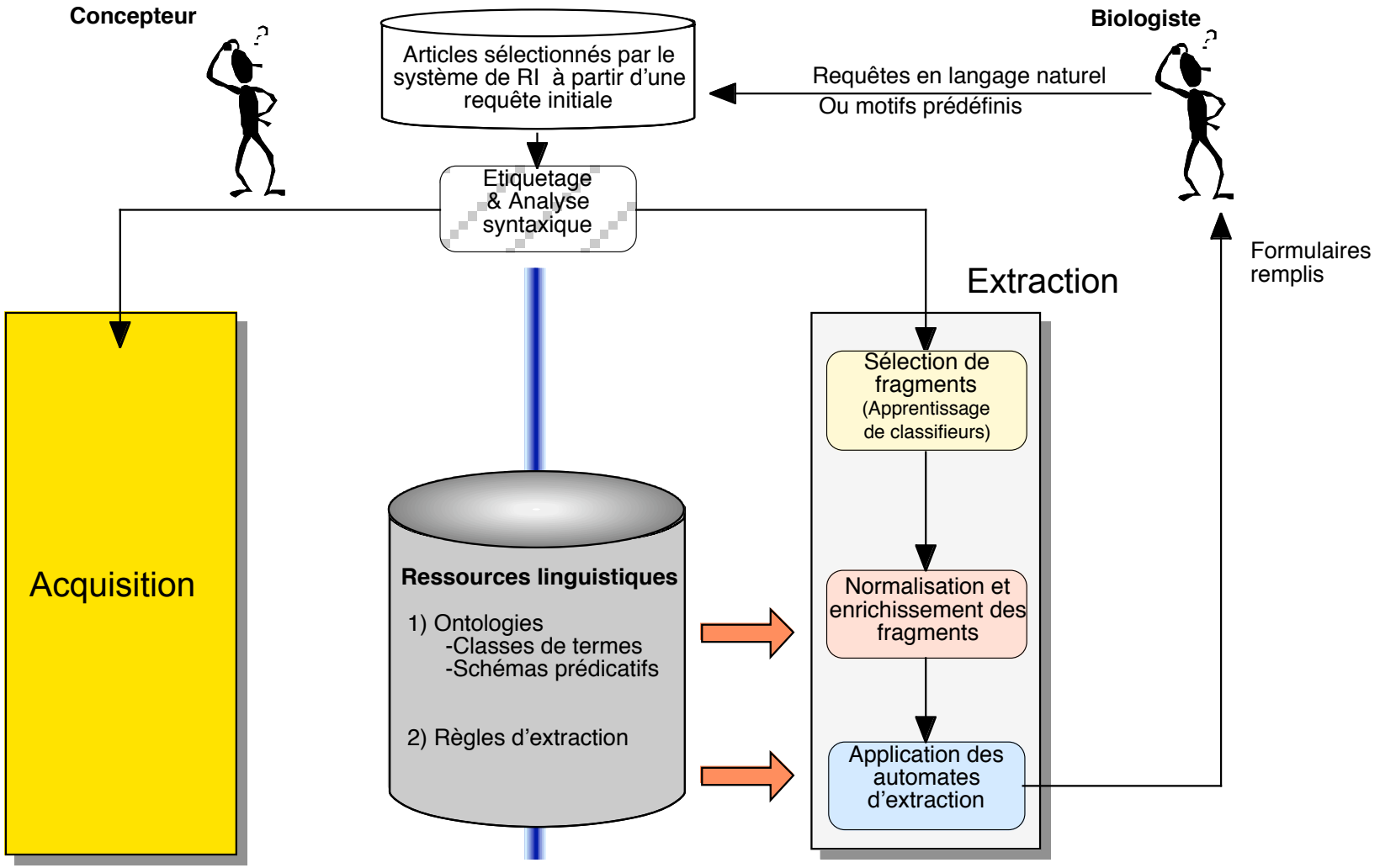
→

Interaction
Type: negative
Agent : GerE protein
Target :
Expression
Source : gene sigK
Product: protein sigmaK

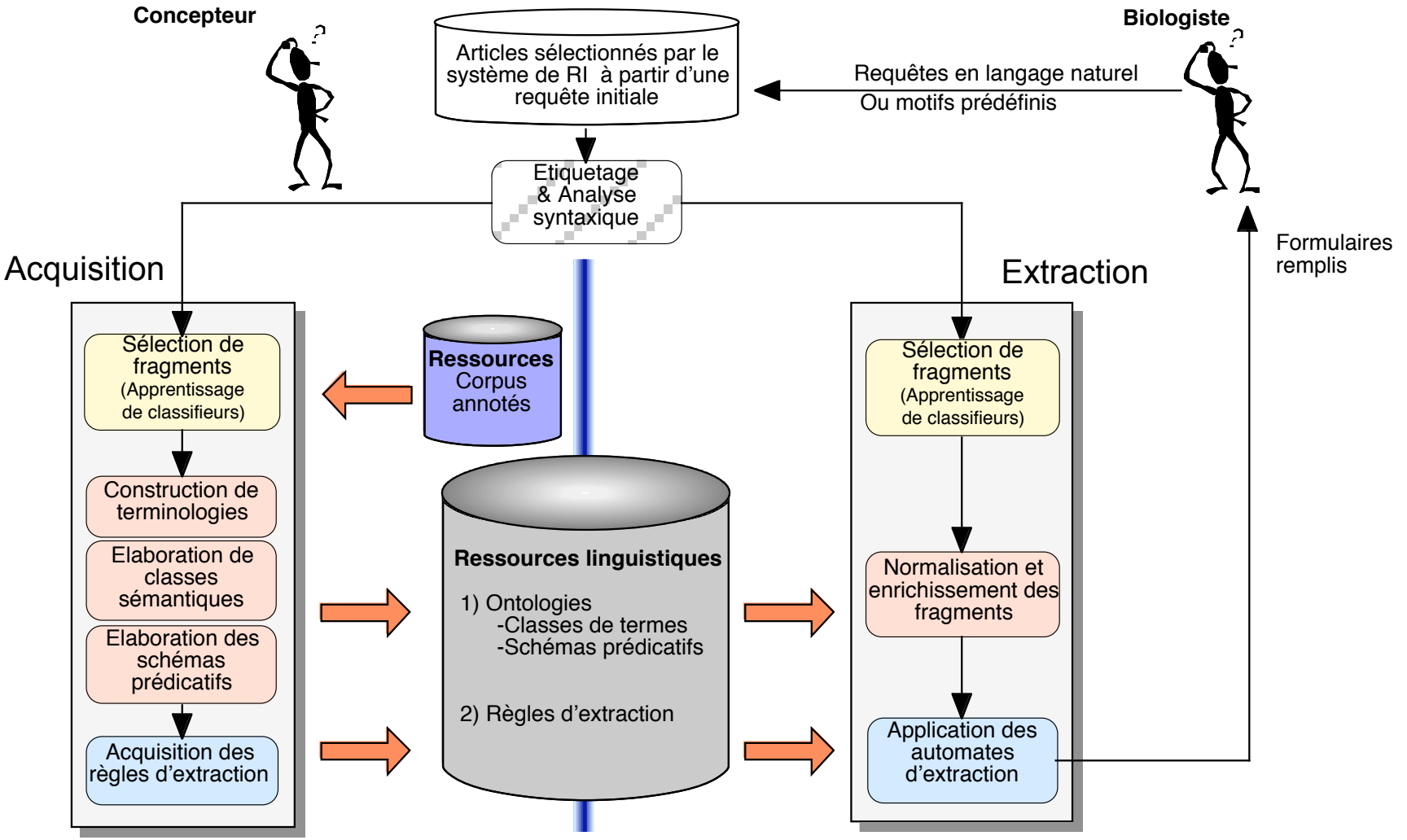
Architecture générale



Architecture générale Extraction



Architecture générale Acquisition



Bilan 2000-2001

- **Sélection des fragments pertinents**
  - Filtrage automatique de phrases sur 3 bases de test (IVI, Bayes, ID3)
  - Apprentissage à partir de corpus annotés (manuellement) de phrases

		Rappel	Précision
• Bs (P. Bessieres, MIG) *dispo*	:	91%	82%
• Dro (B. Jacq, LGPD-IBDM)	:	90%	85%
• HM (C. Brun, LGPD-IBDM & Valigen)	:	97%	89%

- **Construction de classes sémantiques**
  - Etude des paramètres d'apprentissage
  - Apprentissage semi-automatique de hiérarchies «clustering»

- **Annotation des textes pour l'apprentissage de règles**
  - Définition d'un langage d'annotation XML

<GENIC-INTERACTION id = "1" assertion = exist …

<CF> <C> Previous studies showed </C> </CF> that <AF1> <A1 type=gene role=activate direct=undefined > spoIIID </A1> </AF1> <IF> <I> is needed to produce </I> </IF> <TF1> <T1 type=protein> sigma K</T1> </TF1>, but suggested that spoIIID represses sigma K directed transcription of genes encoding spore coat proteins.

</GENIC-INTERACTION >

Publications

- BESSIERES P., NAZARENKO A., NEDELLEC C. (2001). Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques. A paraitre dans les actes de CIDE 2001, 4e Colloque International sur le Document Electronique. Toulouse 24-26 oct. 2001, France.
- BISSON G., NEDELLEC C., CAÑAMERO L. (2000). Designing clustering methods for ontology building: The Mo'K workbench. Ontology Learning workshop (ECAI 2000), Berlin, 22 août 2000.
- BISSON G. ET NEDELLEC C. (2001). Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K. In H. Brian Eds. Actes des Journées Francophones d'Extraction et de Gestion des Connaissances (EGC 2001), Hermès (Pub.) Nantes.
- NEDELLEC C., OULD ABDEL VETAH M., BESSIERES P., BRUN C., JACQ J. Reconnaître les fragments de phrases pertinents pour l'extraction d'information dans les textes de génomique, un problème de classification. Actes de la Conférence francophone d'Apprentissage (CAP 2001), PUG (eds).
- NEDELLEC C. ET NAZARENKO A. (2001). Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet CADERIGE : identification d'interactions géniques. In Actes de la Journée thématique Exploration de données issues d'Internet organisée le 2 mars 2001 au LIPN. Bennani Y., et al. (Eds).
- NEDELLEC C. ET OULD ABDEL VETAH M. (2001). Modélisation des interactions géniques à partir de textes. Journée Post-Génomique de la Doua (JPGD), Lyon.
- NEDELLEC C. ET OULD ABDEL VETAH M., BESSIERES P. (2001). Sentence Filtering for Information Extraction in Genomics, a Classification Problem. To appear in proceeding of 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 01). September 3-7, 2001, Freiburg, Germany

CADERIGE en 2002

- **Nouveau projet 2001-2003**
  - Mêmes partenaires – (GM et LRV)
  - + Laboratoire INRA-ENSAR de Génétique Animale.

- **Deux outils bientôt disponibles pour la communauté**
  - Filtreur de phrases (biblio + nom de genes)
  - Editeur d'annotations (JAVA)

- **Travaux en cours**
  - Comparaison des analyseurs syntaxiques «□bre» & «□ouvert□» (LinkParser)
  - Extraction terminologique en cours …
  - Apprentissage des règles d'extraction (Amilcar)

Rendez-vous en 2003 et sur http://caderige.imag.fr/