Automatisation du pipeline d'annotation de la plateforme MicroScope pour un passage à large échelle

Stéfan Engelen & David Vallenet

CEA/DSV/IG/Genoscope
CNRS-UMR 'Génomique Métabolique'
Laboratoire (Atelier)
de Génomique Comparative

Les composants de MicroScope

Système de production

Pipeline d'annotation automatique

Gestionnaire de données

Base de données relationnelle PkGDB

BD objet MicroCyc (Réseaux métaboliques)

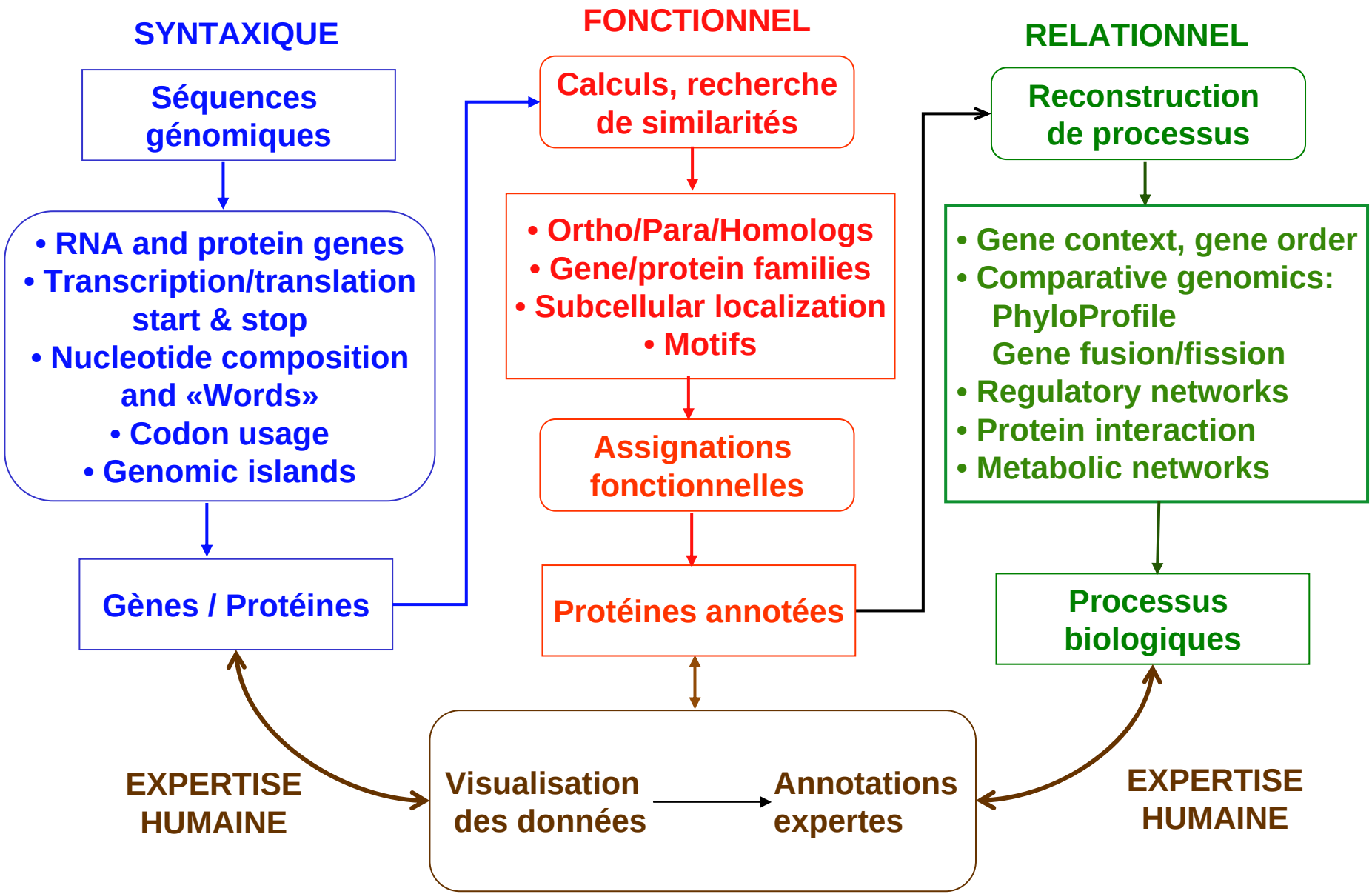
Système de visualisation

Interface graphique de consultation et d'édition MaGe

Explorer et éditer les connaissances contenues dans le gestionnaire de données :

- Outils de génomique comparative
- Annotation experte

Pipeline d'annotation : flux d'information



Pipeline d'annotation : calculs

- Plus de 20 workflows différents pour l'annotation syntaxique, fonctionnelle et relationnelle.

- Temps CPU (AMD optéron double-cœur) pour un génome (5000 gènes/protéines)

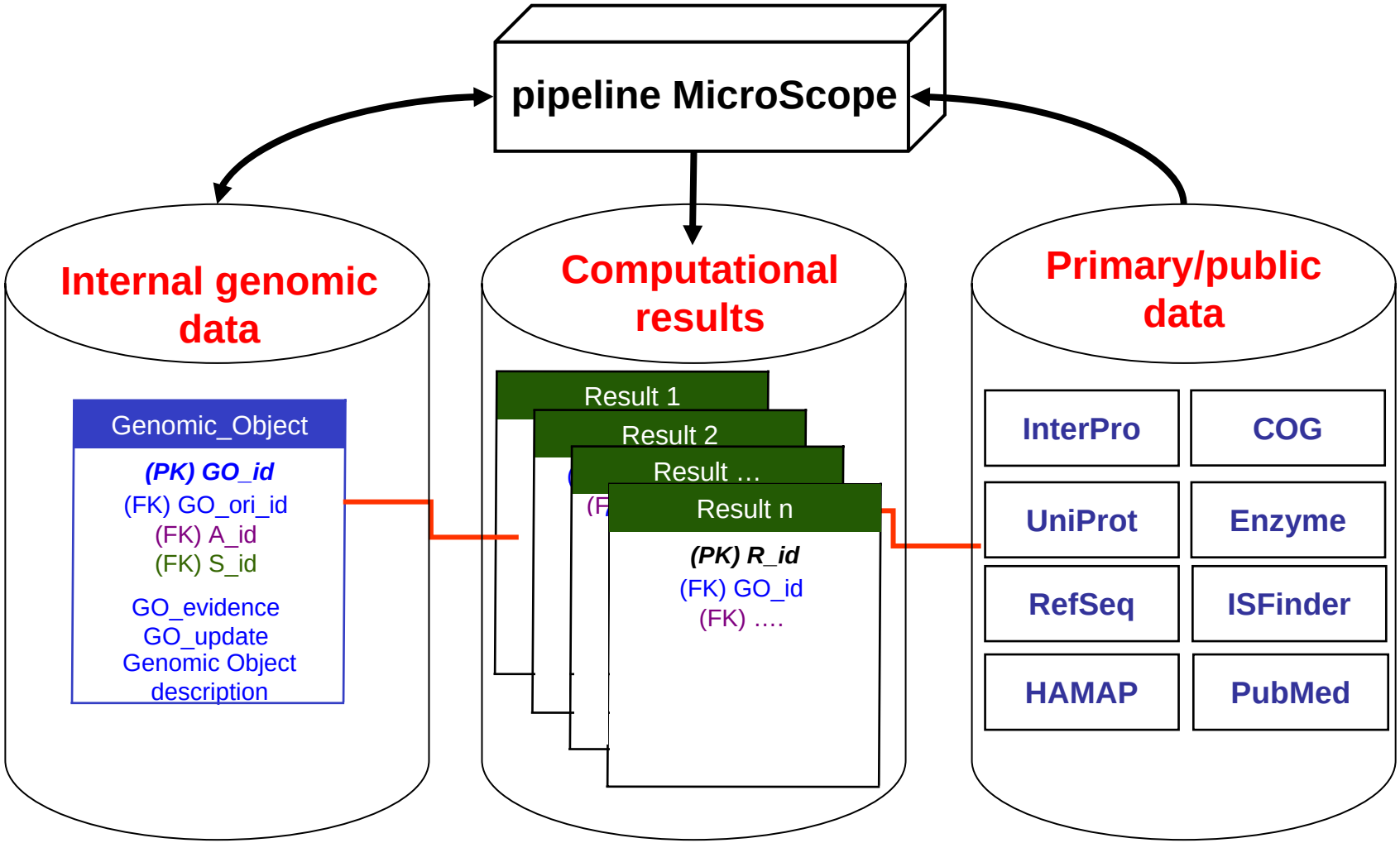
  - Blast sur Uniprot : 73 heures
  - Blast PkGDB + Syntenie : 16 heures
  - Blast RefSeq + Syntenie : 32 heures
  - InterPro : 305 heures
  - Autres analyses (COG, PRIAM, BioCyc, Rfam) : 10 heures

  => Total = 436 heures

- Parallélisation sur 36 CPU (72 cœurs)

  - pour **1 génome = 6 heures**
  - pour tous les génomes de PkGDB = 86 jours

La base PkGDB



5 Go
38 tables

600 Go
29 tables

6 Go
41 tables

Interface graphique : **Ma**gnifying **Ge**nomes

MaGe

MicroScope project

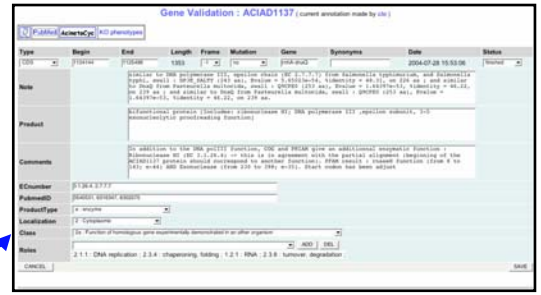
Authentification

Help(s)

Options

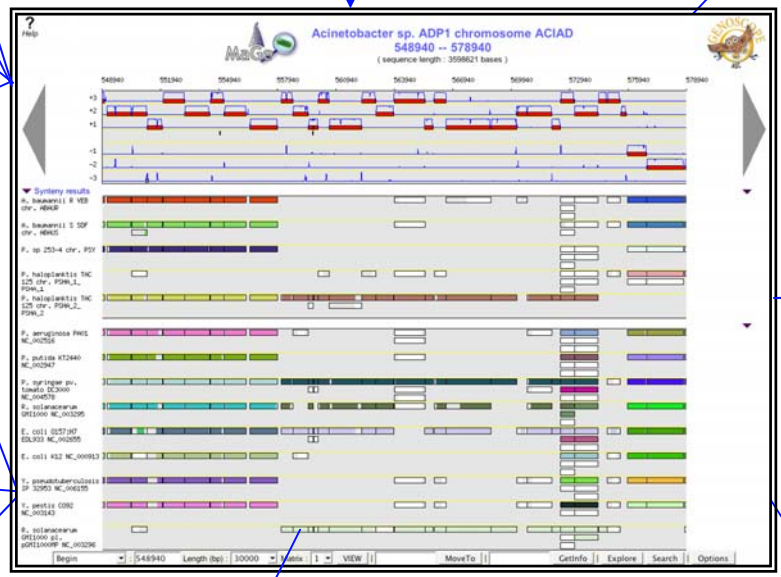
Genome Overview

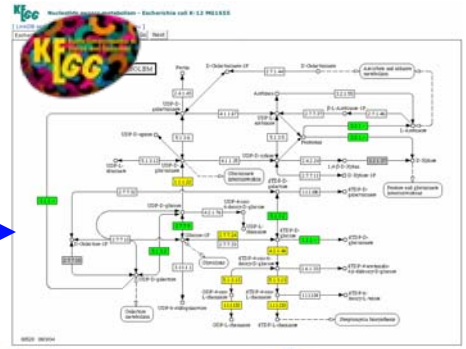
Export



Annotator editor

Artemis





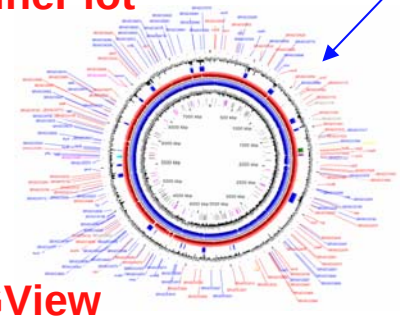
BioCyc Database Collection

Pathway Hunter Tool

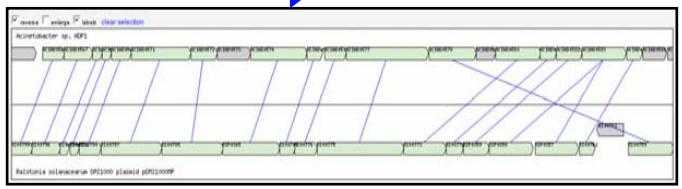
Metabolic pathways

LinePlot

Synteny map



CGView



Synton visualization

KeyWords
Blast / Motif
Phylogenetic profiles
Fusions / Fissions
Genomic islands
Metabolic profiles

Exploration

Taux d'utilisation de la plateforme

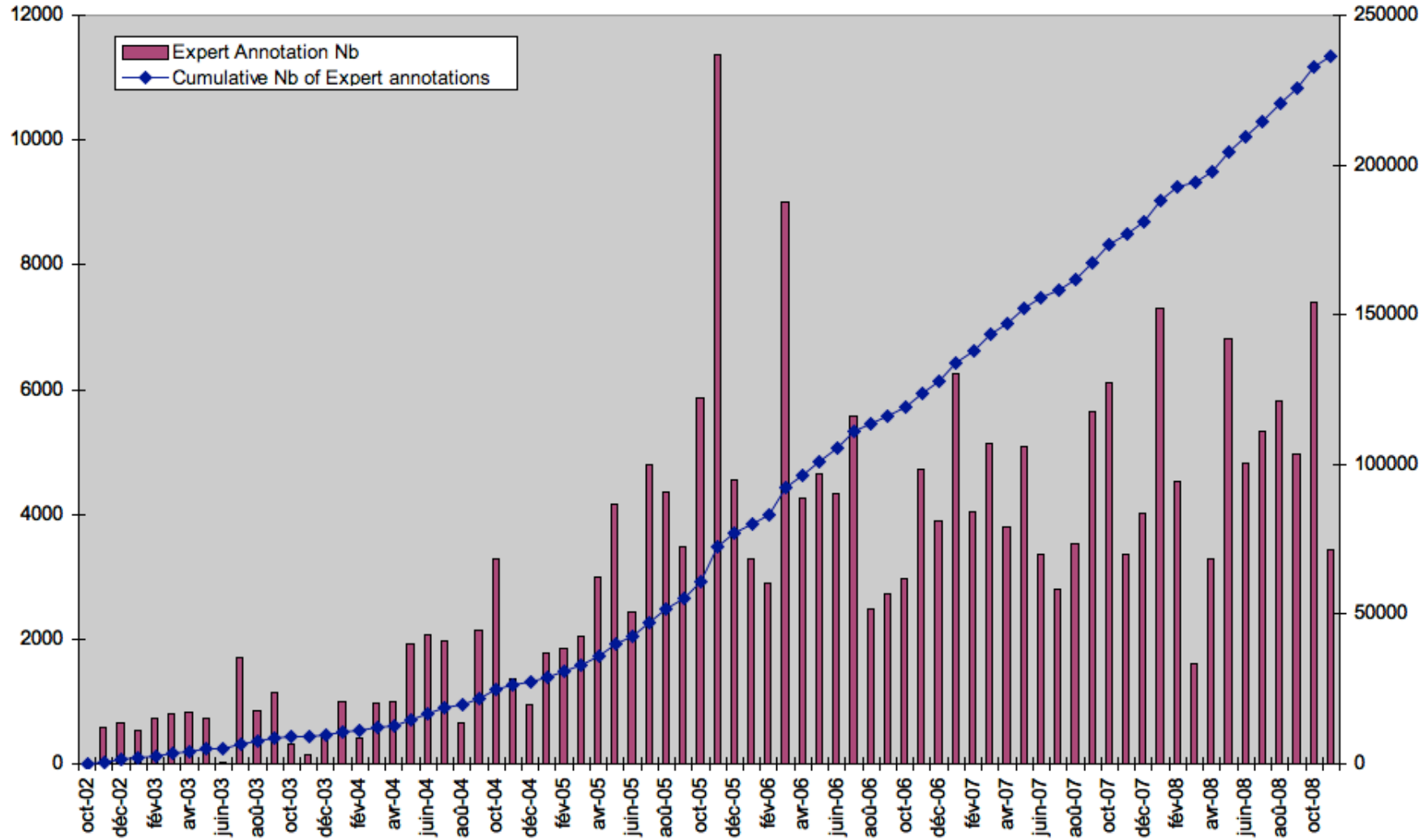
- **563 comptes** personnels { **380 en France**; **76 en Europe**; **45 Etats-Unis + 62 autres pays**

- Projets collaboratifs avec une **trentaine de laboratoires en France** : Universités Paris, Strasbourg, Lyon, Marseille - CNRS - Institut Pasteur - CIRAD Montpellier – INRA - CEA.
- Projets collaboratifs avec **16 laboratoires hors France** : USA, Chine, Corée, Allemagne, Belgique, Portugal, …
- **Partenaire(s) industriel(s)** : Sanofi-Aventis

- **70 authentifications par jour (env. 150 pages consultées par session)**

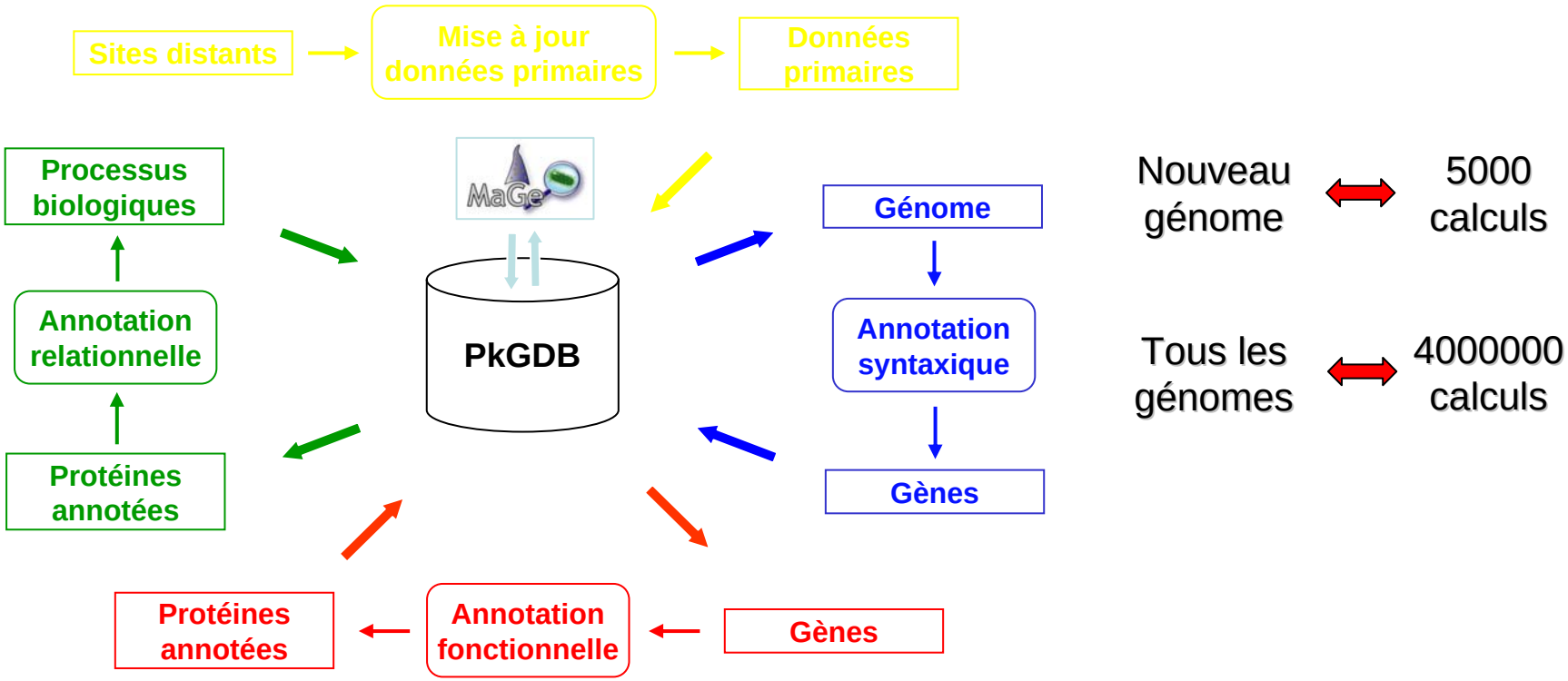
⇒ Efficacité du serveur Web MaGe (interface graphique)

Evolution du nombre d'annotations expertes



- **236 233 annotations expertes sont à ce jour enregistrées dans la base PkGDB**

Pipeline d'annotation : objectifs ?



Objectifs

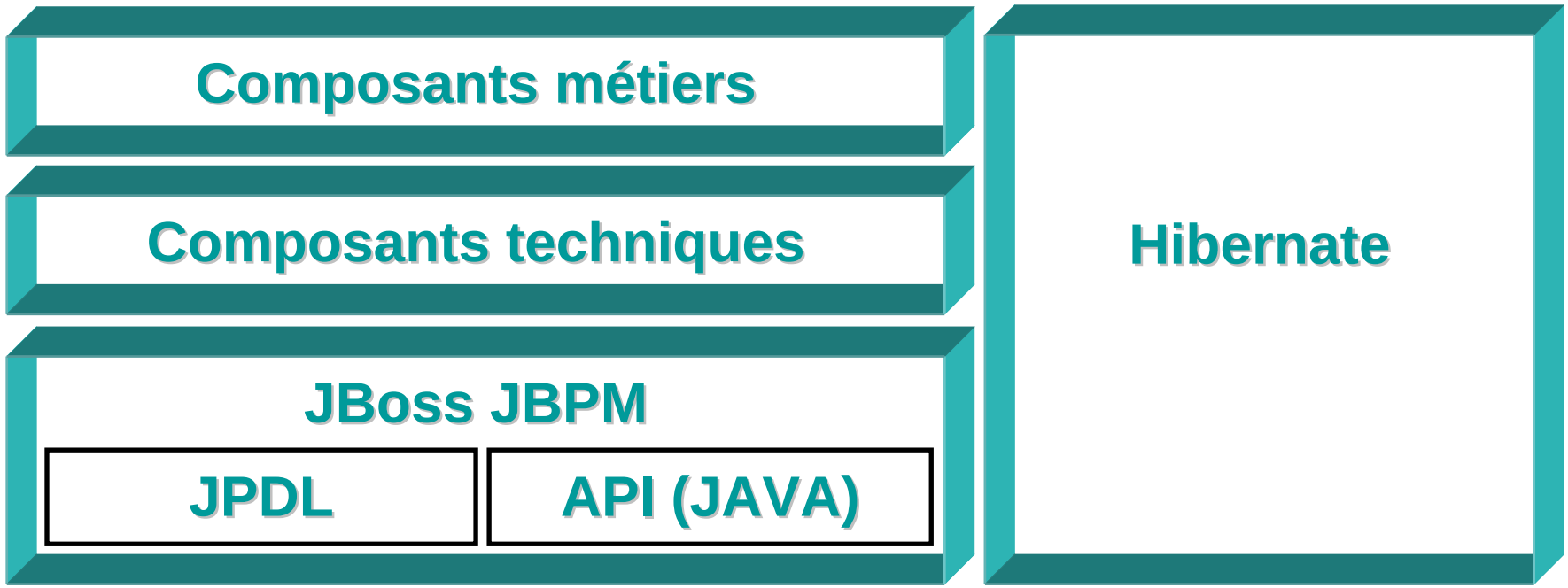
- ❑ Automatisation pour un passage à large échelle
- ❑ Garantir les performances d'accès (concurrence requêtes utilisateurs / chargement calculs)
- ❑ Maintenir les données primaires et calculs à jour

Identification des besoins

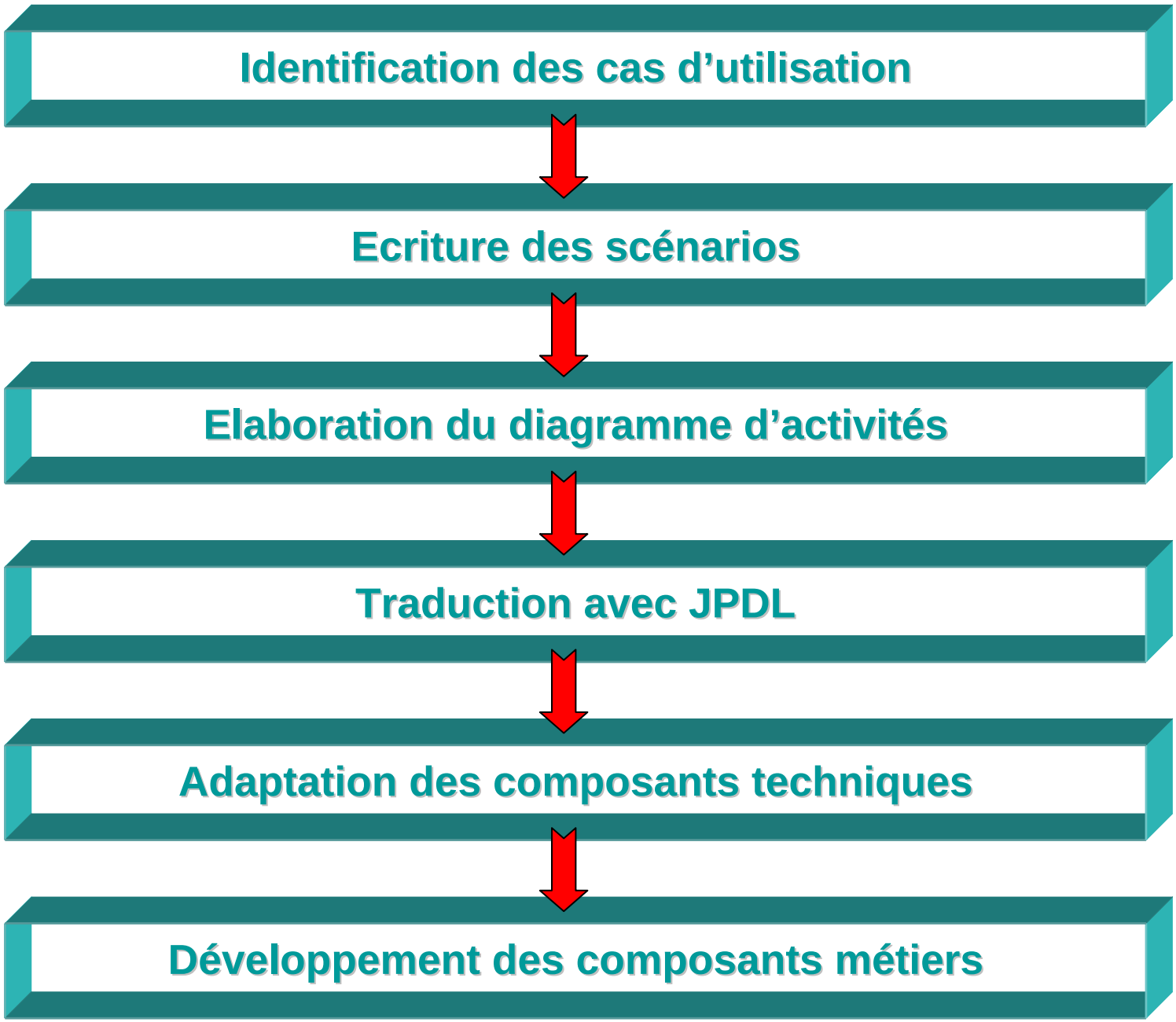
- **Définir / décrire** rigoureusement les processus métiers : diagramme d'activités
- **Orchestrer / synchroniser** les activités humaines, calculs et processus systèmes l'aide d'un langage de description
- **Robustesse :** reprendre sur échec une activité
- **Contrôler et suivre** l'évolution des calculs en temps réel
- **Garantir la traçabilité :** sauvegarde de l'historique

JBPM - **J**ava **B**usiness **P**rocess **M**anagement : solution technique à la mise en œuvre des processus métiers

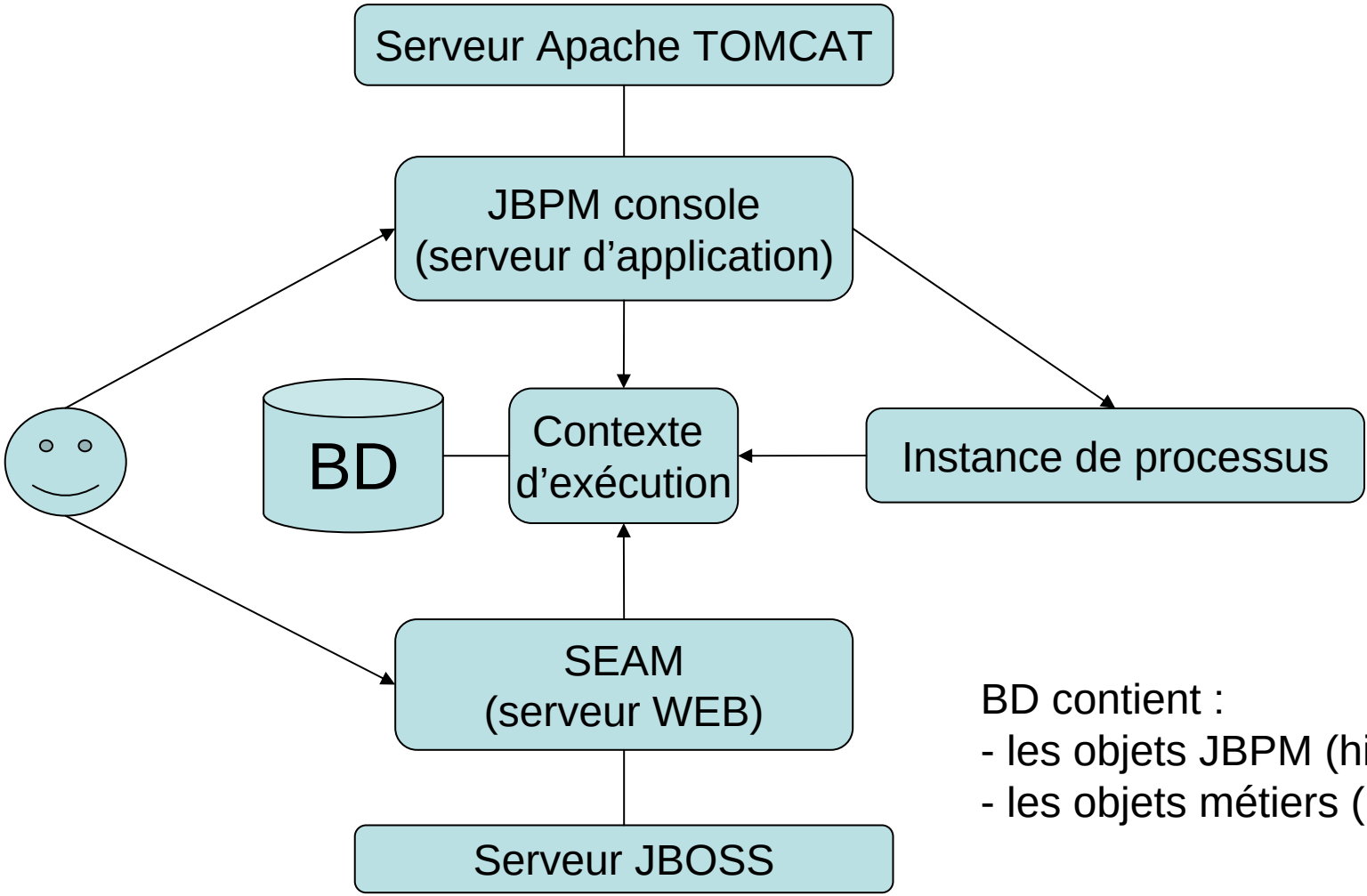
- Solution open-source, souple, capable d'orchestrer un ensemble de programmes, quelque soit le langage d'implémentation
- JPDL : langage de description des activités (**Orchestrer / synchroniser**)
- API (JAVA) : développement de nouveaux composants techniques et métiers (**suivi / contrôle**)
- Hibernate : Persistance dans une BD (**robustesse / traçabilité**)



JBPM : cycle de développement



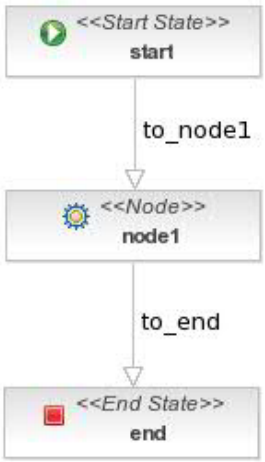
JBPM : architecture logicielle



BD contient :
- les objets JBPM (hibernate)
- les objets métiers (PkGDB)

Définition du processus

- On décrit le processus avec le langage **JPDL** (Java Process Definition Language) au format XML
- JPDL décrit un graphe et des propriétés dans chaque noeud et transition du graphe



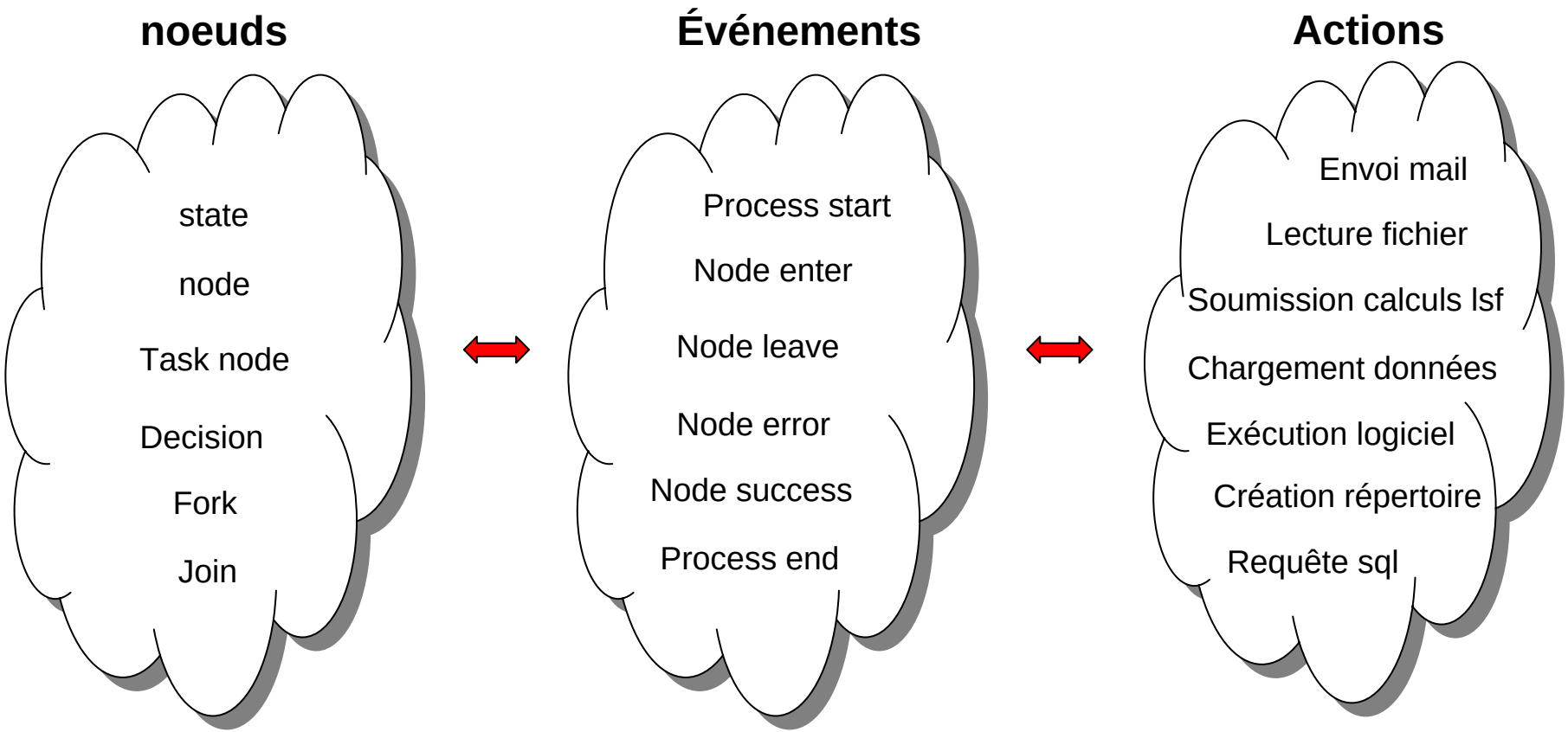

```
<?xml version="1.0" encoding="UTF-8"?>

<process-definition xmlns=""  name="Demo">
   <start-state name="start">
      <transition name="to_node1" to="node1"></transition>
   </start-state>
   <end-state name="end"></end-state>
   <node name="node1">
      <transition name="to_end" to="end"></transition>
   </node>
</process-definition>
```

Définition du processus

- Plusieurs type de nœuds dont l'activité dépend :
  - d'un comportement initial du nœud
    - **State** (état d'attente)
    - **Node** (état à comportement personnalisable)
    - **TaskNode** (assigne et crée des tâches)
    - **Decision** (dirige le flux en fonction d'un critère)
    - **Fork** (sépare le chemin d'exécution entrant)
    - **Join** (état bloquant/passant)
  - de l'exécution d'un certain nombre d'actions implémentées via l'API JBPM
  - de la synchronisation des actions par des évènements
- Un seul type de transition

Définition du processus

- L'activité d'un nœud correspond à l'exécution d'actions associées à des types d'évènements
- Action : composant métier reposant sur des composants techniques



Cas d'utilisation : workflow de calculs des synténies

Comparaison blast

Tache métier 1 : whichGenome
- ❑ génomes non à jour

Tache métier 2 : executeLSFscript
- ❑ comparer 2 à 2 n génomes avec blast
- ❑ 2^n calculs parallélisés sur un cluster
- ❑ calculs terminés ou échoués

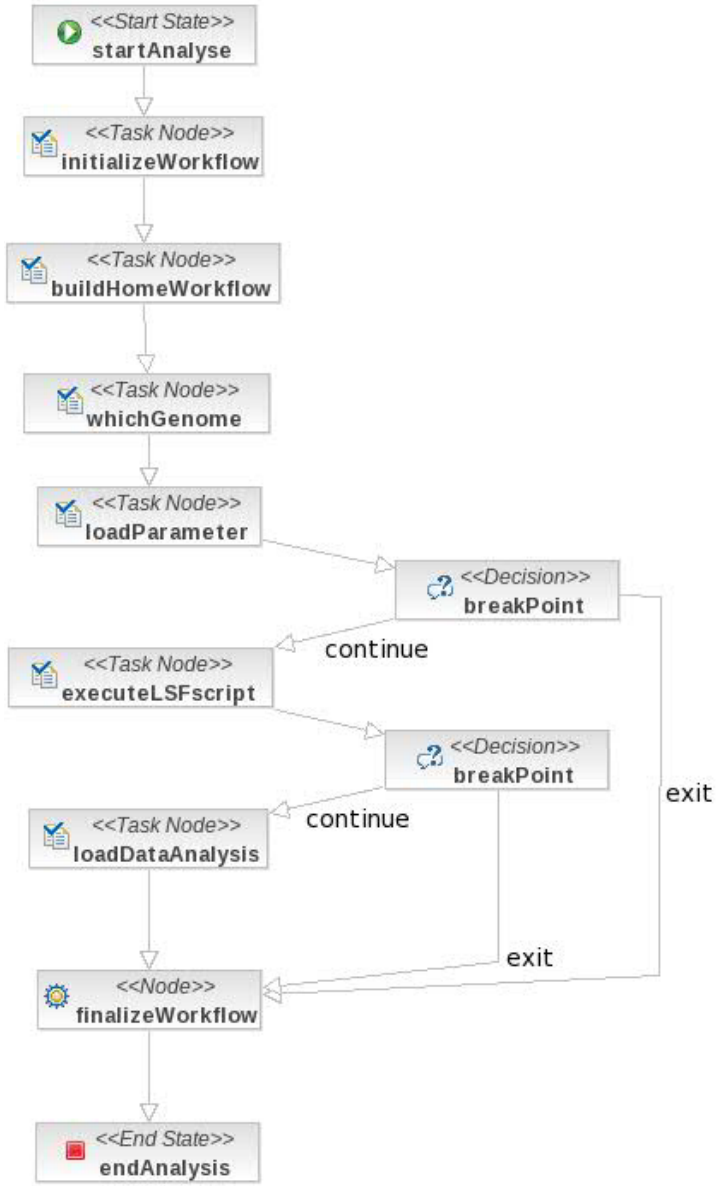
Tache métier 3 : loadDataAnalysis
- ❑ chargement dans PkGDB des calculs terminés à la tache 2

Calcul de synténies

Tache métier 1 : whichGenome
- ❑ génomes avec blast et sans synténies

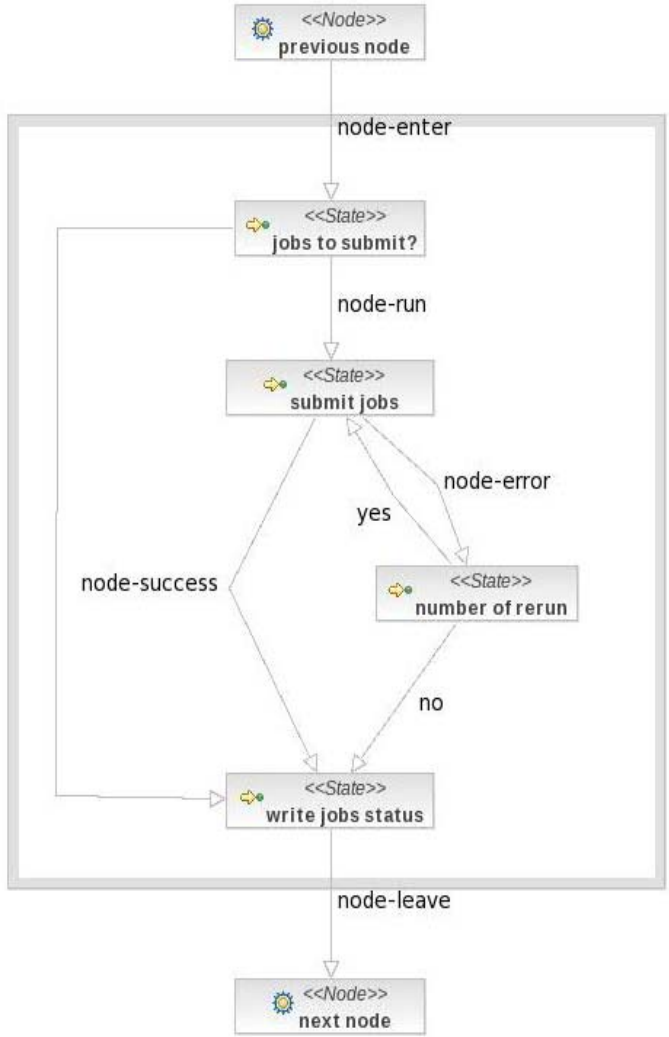
Tache métier 2 : executeLSFscript
- ❑ calculs de synténies
- ❑ 2n calculs parallélisés sur un cluster
- ❑ calculs terminés ou échoués

Tache métier 3 : loadDataAnalysis
- ❑ chargement des synténies terminés à la tache 2

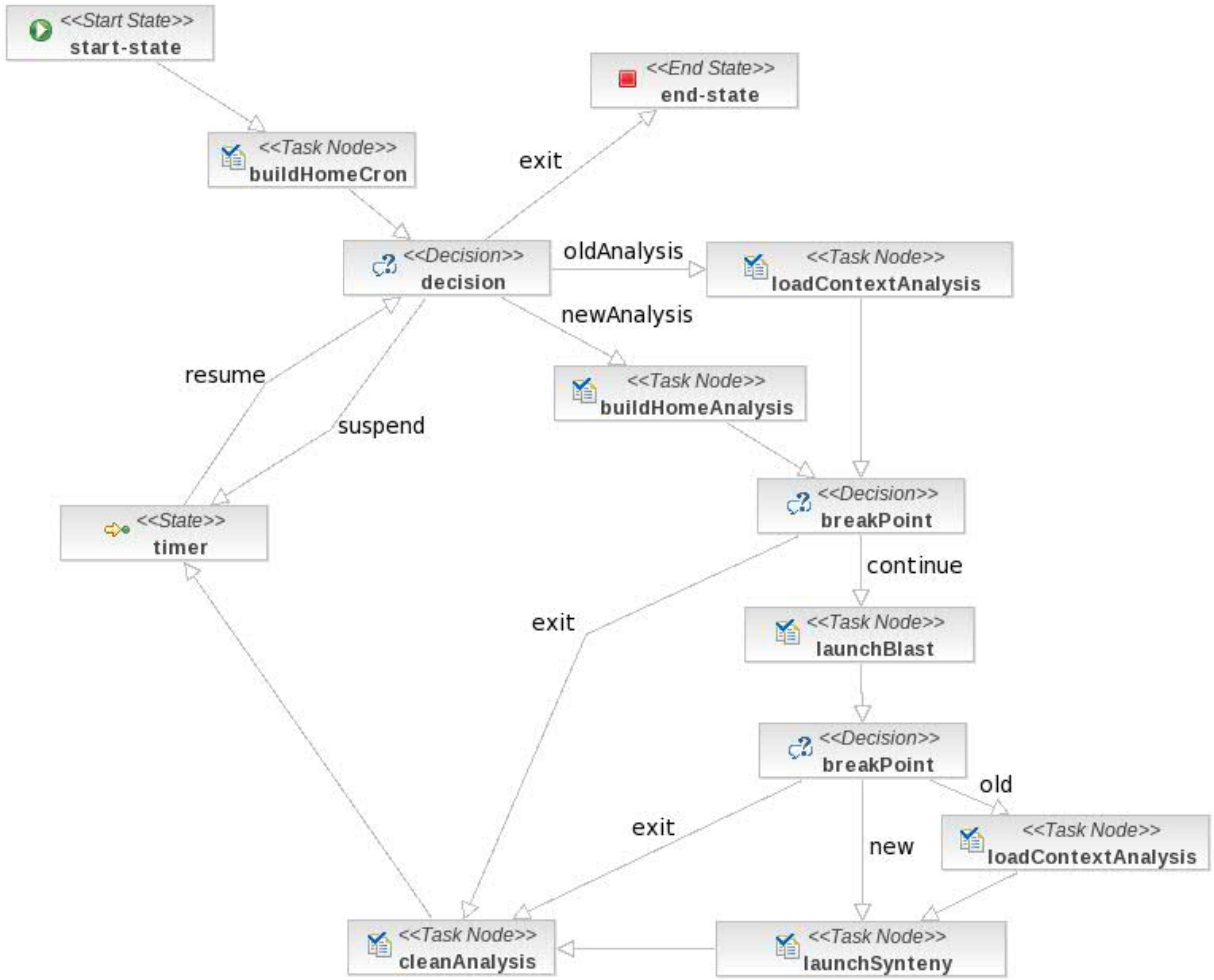


Tache métier : soumission de calculs blast

- Entrée : n génomes
- Formatage : 2^n lignes de commandes
  - blast -d genome_1 -i seq_1_1.fasta
  - blast -d genome_n -i seq_m_n.fasta
- Soumissions de 2^n calculs répartis en p paquets : LSF
  - bsub -o seq1.out -q big ‘blast -d genome_1 -i seq_1_1.fasta’
  - bsub -o seqN.out -q big ‘blast -d genome_n -i seq_m_n.fasta’
- Attente des résultats
- Parçage des p fichiers de résultats
- Resoumissions des calculs en erreurs
  - nombre de resoumissions fixé
- Persistance des statuts des calculs
  - N_e calculs en erreur
  - N_s calculs en succès
- Sortie : (2^n - N_e) alignements
- Reprise sur erreur
  - base de donnée
  - id workflows non terminés
  - id taches échouées
  - ligne de commande échouées
  - données utilisées



CRON : Synchronisation de workflows métiers



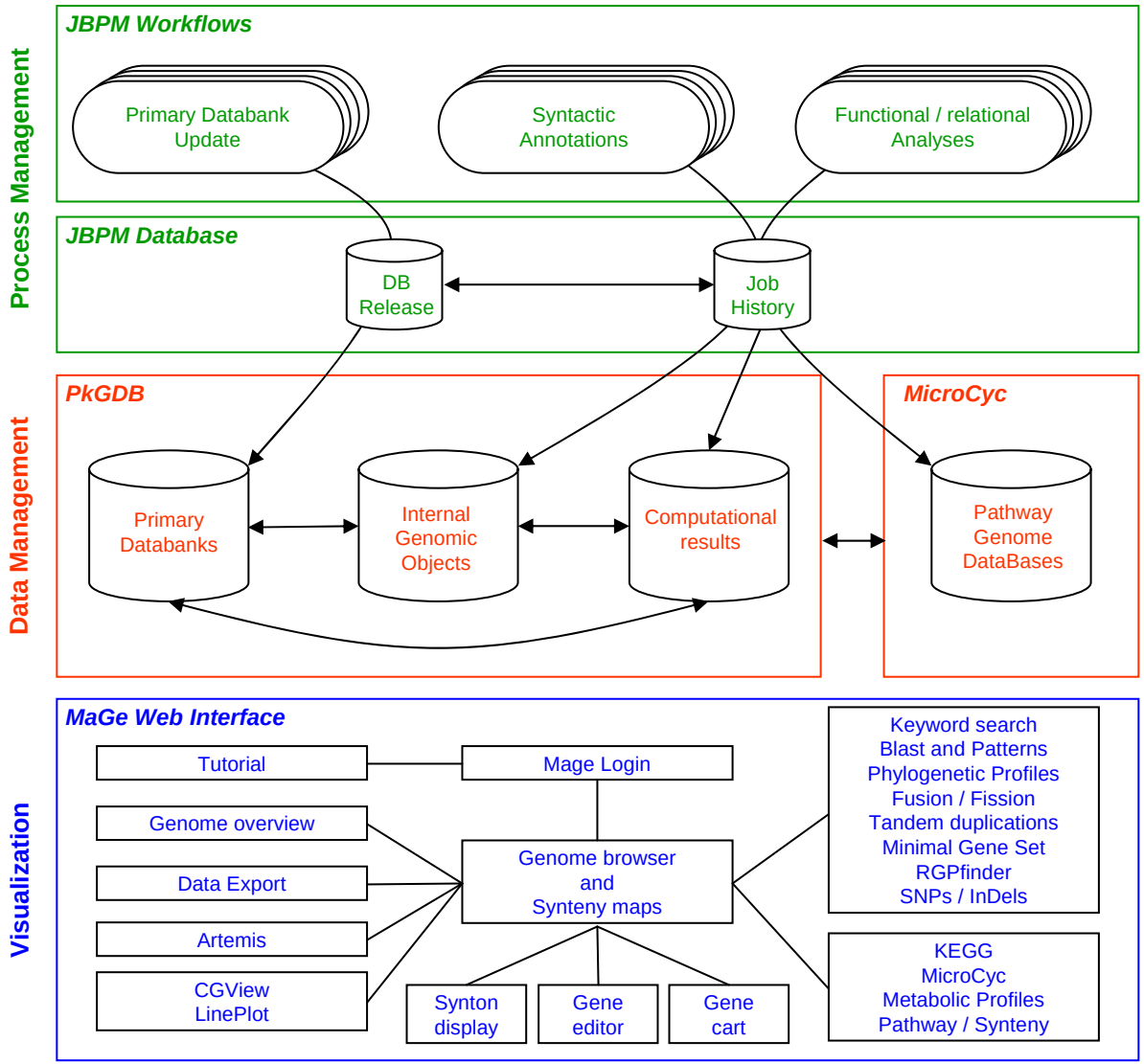
CRON

- ❑ Start
- ❑ Suspend
- ❑ Resume
- ❑ Stop
- ❑ new analysis
  - ➢ input sequences
  - ➢ update all sequences
- ❑ old analysis
  - ➢ id of process to rerun

Workflow métier

- ❑ Suspend jobs (bstop)
- ❑ Resume jobs (bresume)
- ❑ Stop jobs (bkill)
  - ➢ with load
  - ➢ without load
- ❑ Exit at breakpoint n

Architecture plateforme



Conclusion

- Automatisation :
  - 15 workflows de calculs de fréquence journalière
  - 9 workflows de mise à jour des données primaires
- Synchronisation :
  - Taches séquentielles, branchements conditionnels, états d'attente
  - Taches parallèles de chargement dans la base, backup des BD
  - Mise à jour données primaire / calculs
- Robustesse :
  - Reprise automatique en cas de panne d'un composant de la chaîne (base de données, cluster de calculs, panne électrique,…)
  - Mise en évidence d'erreurs et reprise des traitements sur erreurs
- Interaction en ligne de commande :
  - Suivi : avancement et historique des calcul, calculs en erreurs …
  - Contrôle : priorité, reprise et arrêt de calculs, fréquence de mise à jour
- Volumétrie :
  - 2004-2009 : 2276387 calculs lancés et 200 GO chargés
  - depuis 6 mois : 4330795 calculs lancés et 400 GO chargés

Perspectives

- Court terme:
  - o Intégration des méthodes d'analyses syntaxiques et relationnelles de la plateforme
  - o Intégration de nouveaux workflows automatiques de mise à jour des données primaires différentielles ou complète
- Moyen terme:
  - o Mise en place d'un service web « MaGe Light » pour la soumission anonyme de projets d'annotation de génomes
  - o Mise en place et intégration de nouvelles méthodes d'analyses
- Long terme:
  - o Utilisation de grilles pour les calculs lourds de la plateforme

Remerciements

- ANR PFTV MicroScope
- Equipe Informatique : Claude Scarpelli
  - Support informatique
  - Ludovic Fleury
- Equipe LGC : Claudine Medigue
  - Alexandra Calteau
  - Stéphane Cruveiller
  - David Vallenet
  - Zoé Rouy
  - Aurélie Lajus
  - Grégory Salvignol
  - David Roche
  - Alexander Smith
  - Damien Mornico