GenoStats

- La plate-forme bio-informatique
- Pourquoi mesurer le niveau d'utilisation ?
- Les indicateurs

Jean-Marc Larré

- Pourquoi un nième outil ?
- Méthode
- L'architecture du logiciel
- Fonctionnement
- Exemple
- Bilan

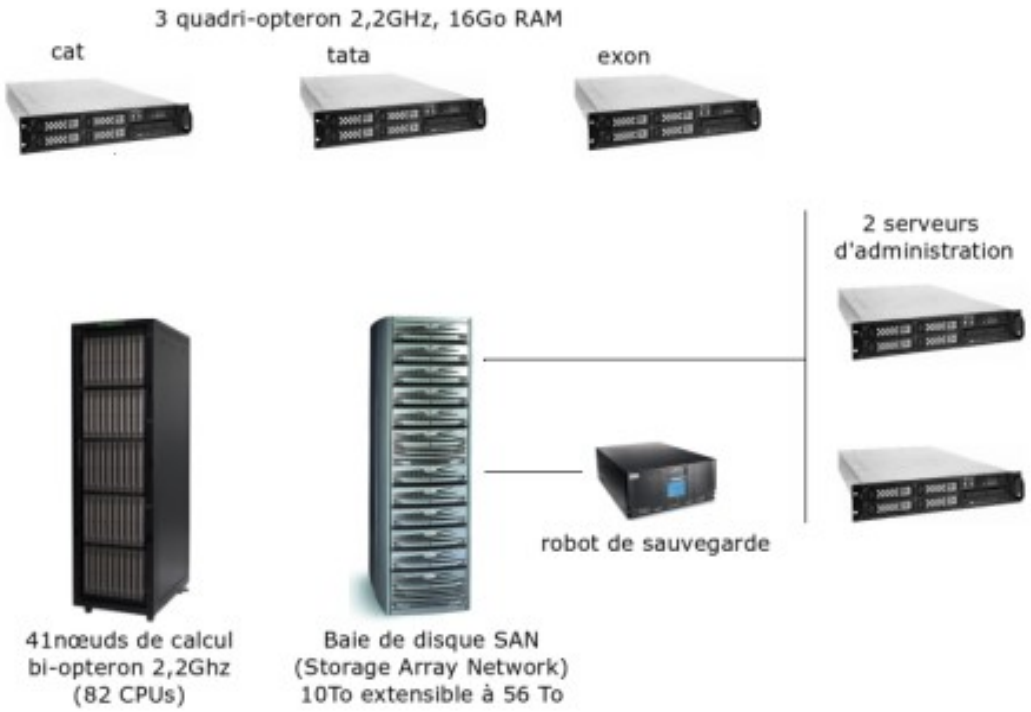
Cédric Chappert

La plate-forme bio-informatique

- inscrite dans la génopole Midi-Pyrénées (GIS) :
  - 5 plate-formes techniques
- localisée sur le campus de l'INRA à Castanet
- constituée de 5 permanents (1 DR, 1 IR et 3 IE) et 1 CDD (1 AI)
- ouverte aux bio-informaticiens et biologistes
- elle a 9 missions, dont :
  - proposer une infrastructure matérielle et logicielle
  - appui bio-informatique aux programmes scientifiques
  - développer des programmes en concertation
- **contacts** : http://bioinfo.genotoul.fr

L'infra-structure

- Jusqu'à aujourd'hui



- **Evolution 2008** : environ 300 coeurs, 50 TO, etc.

Métrologie : oui mais pourquoi ?

- **définition de la métrologie de la plate-forme** : mesurer le niveau d'utilisation de notre système
- **comment** : en relevant la valeur d'indicateurs de manière récurrente
- **pourquoi** : pour élaborer des statistiques dans le but de :
  - maîtriser l'utilisation faite de la ressource afin d'améliorer le service rendu
  - de justifier le niveau d'utilisation de la ressource auprès des instances de tutelles pour obtenir des financements, ou autres ressources
  - mettre en place à terme un système de facturation
  - obtenir la certification ISO 9001 pour mars 2010

Les indicateurs

- comptes ouverts
- connexions ssh
- connexions services web
- support utilisateur
- volumes disques
- sauvegarde
- consommation des ressources de calcul
- indisponibilités des systèmes

- banques de données
- logiciels de bio-informatique

Environ 100 types de données à recueillir

Au moins autant de statistiques à extraire et à calculer

Pourquoi un n^{ième} outils ?

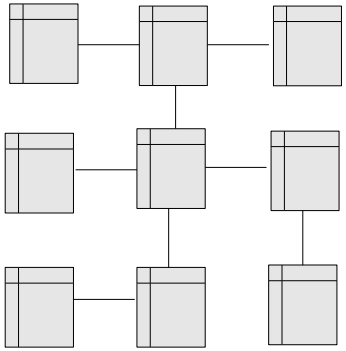
- Nagios
  - Réseau, hôtes, service.
  - Ressources serveurs (CPU, HDD, RAM, SWAP).
- Cacti
  - Monitoring réseau
  - Base RRDTool.
- RRDTool
  - Base de données à taille fixe.
  - Données temporelles (2 dimensions)
- Ganglia
  Monitoring cluster
  Base RRDTool.

Concepts de Génostats

- Ne s'applique pas seulement à du monitoring de systèmes.
- Données multi dimensionnelles.
- Modèle plus souple qu'une base RRD.
- Facilement utilisable :
  - L'utilisateur remplit des fichiers de configuration.
  - Il ne devrait taper aucune requête manuellement.
- Pré requis, la commande existe
- Évolutif et Générique.

Méthodes

- Au départ, un modèle relationnel.



Difficile à maintenir,
Rigide et peu évolutif,
Des requêtes complexes,
...

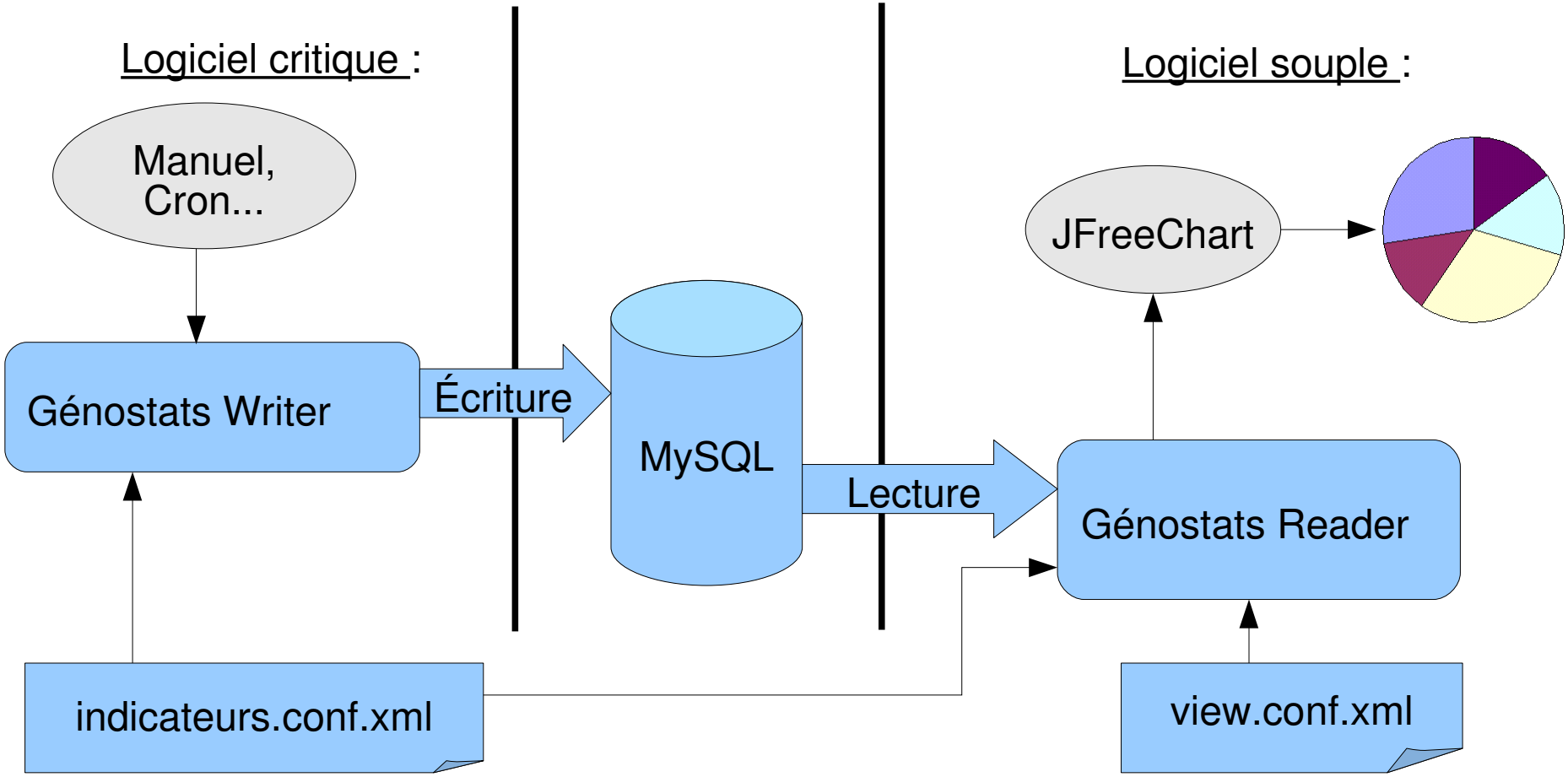
Métamodélisation

Au départ, les tables n'existent pas.
La création des tables, l'insertion, la mise à jour et la lecture de données sont dynamiques.

Base SQL
(Vide)

Architecture logicielle

- 2 parties bien distinctes :



L'indicateur 'Banks'

- Le résultat du script biomaj2genostat.pl :

indexation des banques génomiques

Banques	Release	Taille	Blast	SRS
Anopheles_gambiae	2008-01-15	672	0	0
Apis_mellifera	2007-11-15	393	1	0
Arabidopsis_thaliana	2007-04-28	1260	0	0
unigene	2008-03-03	35393	0	1

Exemple de configuration

```
<indicator name="banks">

     <indicator_properties
          completeName="Les banques génomiques mises à jour sur la plate-forme"
          comment="utilise un script PERL biomaj2metastat.pl" />

     <measureFrequency type="month" historic="true" />
     <attributes key="bankName"> <!-- Primary key = historic + bankName -->
          <attribute name="bankName" type="String" associated="false" />
          <attribute name="releaseDate" type="Date" format="ISO_8601" associated="false" />
          <attribute name="bankSize" type="Int" associated="false" />
          <attribute name="isBlastIndexed" type="Boolean" associated="false" />
          <attribute name="isSRSIndexed" type="Boolean" associated="false" />
     </attributes>
     <indicatorInfo>
          <command>/home/genostat/scripts/biomaj2genostat.pl</command>

          <parsers></parsers>

          <result>
               <column name="bankName" />
               <column name="releaseDate" />
               <column name="bankSize" />
               <column name="isBlastIndexed" />
               <column name="isSRSIndexed" />
          </result>
     </indicatorInfo>
</indicator>
```

Génération des requêtes

- La création de la table :

```
CREATE TABLE IF NOT EXISTS ms_banks (
     ms_historicDate TIMESTAMP NOT NULL,
     ms_bankName TEXT NOT NULL,
     ms_bankSize INT NOT NULL,
     ms_releaseDate TIMESTAMP NOT NULL,
     ms_isSRSIndexed BOOLEAN NOT NULL,
     ms_isBlastIndexed BOOLEAN NOT NULL,
     ms_uniqId INT NOT NULL AUTO_INCREMENT,
     KEY (ms_uniqId),
     PRIMARY KEY (ms_historicDate , ms_bankName(32))
)
```

- L'insertion des données :

```
INSERT INTO ms_banks (ms_bankName, ms_bankSize, ms_releaseDate, ms_isSRSIndexed,
ms_isBlastIndexed, ms_historicDate )VALUES ('Anopheles_gambiae', '672', '2008-01-15 00:00:00.523',
'0', '0', '2008-03-28 13:38:16.632')
INSERT INTO ms_banks (ms_bankName, ms_bankSize, ms_releaseDate, ms_isSRSIndexed,
ms_isBlastIndexed, ms_historicDate )VALUES ('Apis_mellifera', '393', '2007-11-15 00:00:00.527',
'0', '1', '2008-03-28 13:38:16.632')
INSERT INTO ms_banks (ms_bankName, ms_bankSize, ms_releaseDate, ms_isSRSIndexed,
ms_isBlastIndexed, ms_historicDate )VALUES ('Arabidopsis_thaliana', '1260', '2007-04-28
00:00:00.528', '0', '0', '2008-03-28 13:38:16.632')
INSERT INTO ms_banks (ms_bankName, ms_bankSize, ms_releaseDate, ms_isSRSIndexed,
ms_isBlastIndexed, ms_historicDate )VALUES ('unigene', '35393', '2008-03-03 00:00:00.598', '1',
'0', '2008-03-28 13:38:16.632')
```

Exemple de vue

```
<variables>
          <var name="fileMask">%s-%e-%n</var> <!-- dateDebut-fin-nomVue -->
</variables>

<view name="indexedBanks" title="Indexation des banques" type="pie">
     <option legend="false">
          <values viewValue="true" valueFormat="0"
               viewPercent="false" percentFormat="0.00%" />
          <style>
          </style>
     </option>
     <select>
          <indicator name="banks" option="count" as="nombre" />
     </select>
     <filter>
          <group by="isSRSindexed" />
          <group by="isBlastIndexed" />
          <lines>
               <line label="Non indexées" />
               <line label="Indexées pour Blast" />
               <line label="Indexées pour SRS" />
          </lines>
     </filter>
     <when type="MONTH" Period="CURRENT" on="banks.historic" />
     <file>
          <html name="\$fileMask" />
          <img href="\$fileMask" height="800px" width="600px" />
     </file>
</view>
```

Le résultat des vues

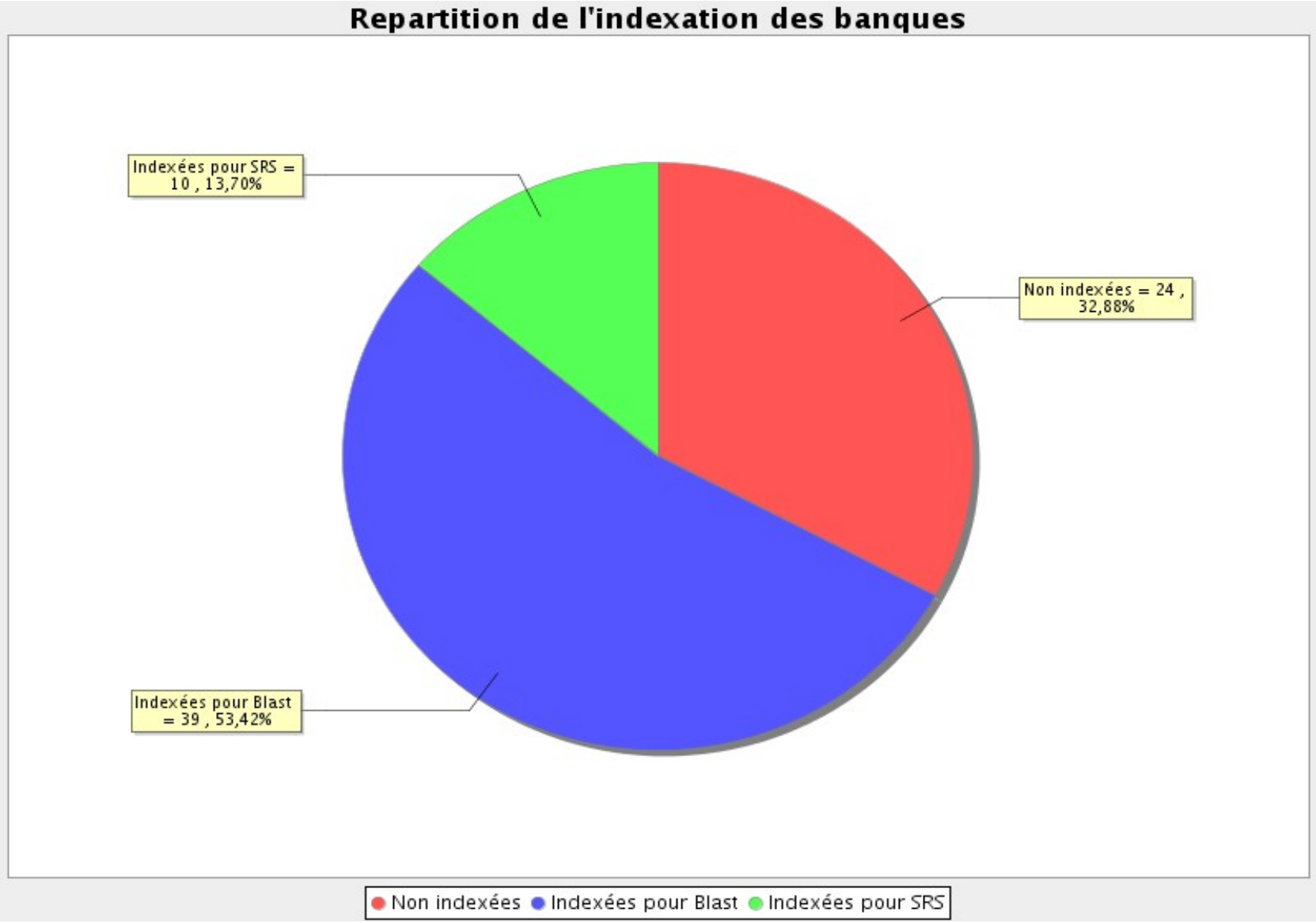
- La requêtes générée

```
select count(*) as nombre from ms_banks
WHERE DATE_FORMAT(ms_historicDate ,'%Y %m') =
DATE_FORMAT('2008-03-28 13:51:12.424','%Y %m')
group by ms_isSRSindexed, ms_isBlastIndexed ;
```

- Le résultat de la requête

```
+--------+
| nombre |
+--------+
|     24 | <line label="Non indexées" />
|     39 | <line label="Indexées pour Blast" />
|     10 | <line label="Indexées pour SRS" />
+--------+
3 rows in set (0.00 sec)
```

Le graphique



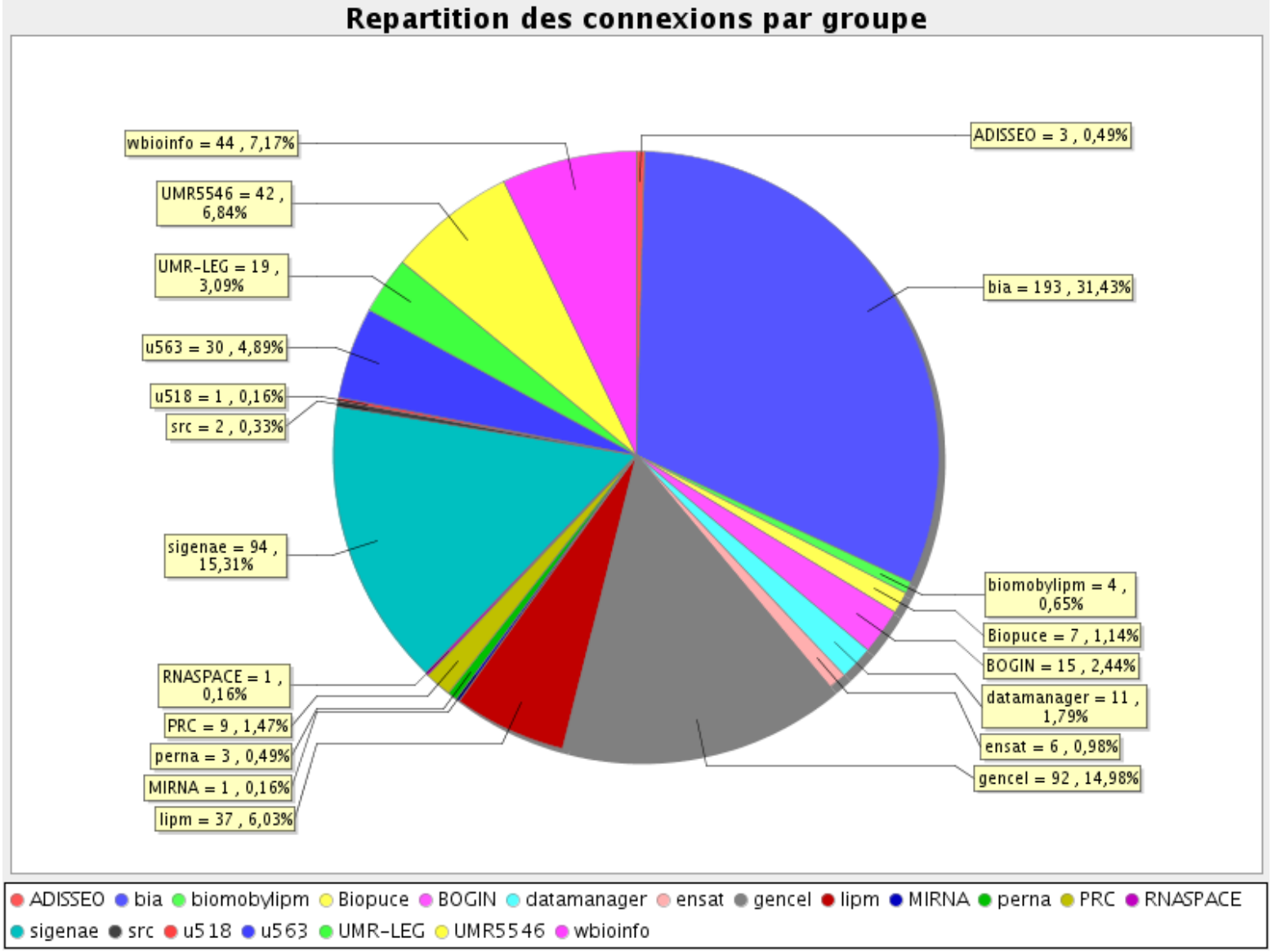
Bilan et avancement

- Avancement du projet :
  - 80% du 'writer' :
    - Des cas particuliers à développer.
  - 40% du 'reader' :
- Déploiement :
  - Mise en test de GenoStats writer : Avril 2008

Bilans et perspectives

- Licence :
  - CeCILL-C, compatible GNU GPL
- Outils connexes :
  - API Shell : CeCILL-C, par adiGuba
  - API JFreeChart : GNU LGPL
- MySQL : version 5
- Première Release : Juin 2008
- Disponibilité : Forge INRA
- Evolution de Genostats 2 : IHM de configurations

Autres résultats



Encore un exemple