Bases de Données & développement de produit technologique, GB4, année 2023-2024 Gilles Bernot

Directives importantes pour la conduite des projets

Les projets sont distribués et commentés durant le cours du 17 janvier et la constitution des groupes avec les choix de projet commencent à cette date. Les choix définitifs des groupes devront être rendus lors du TD du 22 janvier. N'oubliez pas d'indiquer aussi ***qui est le modérateur du groupe***.

Les projets seront à rendre par mail, *a priori* pour le 22 avril, à `Gilles.Bernot@univ-cotedazur.fr` sous forme d'un *fichier PDF*. Le sujet du mail devra être « `projet BD GB4` ».

Le cours du 17 janvier fournira les éléments de gestion de tels projets, ce sont des techniques standard de développement de produits technologiques, adaptées aux produits logiciels, mais qui s'appliquent en fait à toute activité d'ingénierie. Les groupes devront se réunir en dehors des horaires d'enseignement pour avancer le projet.

Les ambiguïtés sur les besoins des utilisateurs et sur les informations nécessaires dans la base de données devront être entièrement inventoriées pour le TD du 12 février. Le 12 février, l'enseignant répondra aux questions ***préalablement réfléchies dans chacun des groupes*** afin de trancher les choix possibles : il jouera donc le rôle du « ***client*** » qui précise ses besoins (mais seulement en réponse aux questions qu'on lui pose) et avec lequel les concepteurs de la base de données ***négocient*** d'éventuels simplifications et assouplissements (y compris du calendrier). Plus aucune négociation, ni de calendrier, ni sur les besoins des utilisateurs, ne sera possible ultérieurement.

Le 6 mars, chaque groupe aura conçu un ***schéma de base de données*** en prenant soin d'***éviter toutes redondances*** qui conduiraient à devoir gérer la cohérence des données lors des mises à jour. Le groupe aura également réfléchi aux ***tests fonctionnels*** qui seront utiles pour valider le projet. L'enseignant jouera alors le rôle de « ***conseiller technique*** » sur le schéma de la base.

Les cours et les TD vous donneront les bases pour écrire ensuite vos requêtes en SQL afin de réaliser les fonctionnalités demandées. N'hésitez pas à consulter internet pour vous inspirer mais il est très important que vous compreniez suffisamment ce que vous programmez pour expliquer en détail le fonctionnement de chaque requête dans le document rendu. Les sites `https://sql.sh` ou `https://www.w3schools.com/sql` sont des références très utiles.

Les TD des 12 février et 6 mars peuvent éventuellement se faire en distanciel, avec chacun des groupes successivement, si *tous les groupes* préfèrent cette organisation.

Il vous est demandé de rendre à la fin du projet les éléments suivants :

1. une reformulation ***complète*** de la description du projet qui supprime toutes les ambiguïtés du présent document et définit exhaustivement les services rendus par la base après négociations avec le client (≈ *spécification des besoins* complète),
2. le schéma de la base de données, exprimé d'une manière graphique de votre choix, et commenté. Le commentaire expliquera et justifiera les choix effectués (≈ *spécification détaillée*),
3. l'expression en SQL des requêtes sur la base de données, qui sont utiles pour rendre les services sus-mentionnés, avec des explications bien structurées de chaque requête (≈ description du *codage*),
   Important : chaque membre du groupe devra prendre en charge l'explication détaillée du

fonctionnement d'une des requêtes SQL et le document rendu devra indiquer ***qui*** a rédigé cette explication. S'il y a moins de requêtes que de membres du groupe, alors les requêtes les plus longues pourront être prises en charge par un binôme. S'il y a plus de requêtes que de membres du groupe alors les plus courtes seront prises en charge par la totalité du groupe, ou bien regroupées et prises en charge ensemble par un des membre du groupe.

4. la description du jeu de données utilisé pour tester les différentes requêtes sous des conditions typiques. La taille de ce jeu de données peut être faible mais on justifiera soigneusement le choix de ces données pour la ***complétude*** des tests (= *validation*),
5. une analyse critique du déroulement de votre projet : découpage et répartition du travail en équipe, difficultés rencontrées et réflexion sur des améliorations éventuelles de l'organisation des tâches au sein du groupe de travail... mais aussi les aspects bien réussis et les éléments clefs de cette réussite. Bref, comment vous y prendriez-vous « *si c'était à refaire* » ? Il *ne* s'agit *pas* de proposer des améliorations du produit logiciel en soi.

Vous effectuerez une implémentation de ce schéma de base de données dans un environnement de bases de données relationnelles de votre choix mais il est *inutile* d'en rendre une sauvegarde : indiquez seulement quels sont les tests qui ont éventuellement échoué... par conséquent la justification sus-mentionnée de *complétude des tests* est cruciale.

La note finale sera principalement fondée sur la qualité globale du dossier rendu, donc par défaut commune à tous les membres du groupe, l'objectif étant de préserver la solidarité du groupe. Une variation individuelle mineure de la note sera fondée sur la pédagogie des explications des requêtes SQL faites par chaque membre du groupe.

Tous les projets ont été calibrés pour être *a priori* de difficulté identique. Par conséquent le choix du projet n'a pas d'influence sur la complexité de réalisation. En revanche, une négociation bien menée avec le client peut avoir une *influence notable sur la simplicité du projet*, de même qu'un bon usage du conseiller technique en anticipant les questions, courtes et précises, qu'on souhaite lui poser sur le schéma de la base.

Description des projets

Projet n°1

Base d'interactions protéiques et voies de signalisation

Un consortium européen regroupant biologie structurale et étude des voies de signalisation souhaite se munir d'une base de données sur la structure des protéines qui l'intéressent et sur leurs interactions. Il s'agit de faciliter la réponse à des questions fréquentes sur ses affinités protéiques d'intérêt pour la signalisation. La base doit pouvoir mémoriser en particulier :

- Pour chaque protéine on doit connaître le nom courant de la protéine pour ce consortium.
- On doit aussi connaître la liste de ses domaines d'interaction répertoriés par les laboratoires de biologie structurale appartenant au consortium et pour chacun d'eux on devra disposer d'une liste de domaines interagissant avec ce domaine.
- Enfin on doit mémoriser toutes les interactions protéiques de signalisation, constatées par les recherches sur les voies de signalisation au sein du consortium. Toutes les protéines mises en jeu doivent figurer dans la base et on ne s'intéresse qu'à des interactions qui mettent en jeu des domaines protéiques. Malheureusement, en revanche, la base ne connaît pas tous les domaines d'intérêt de chaque protéine (compléter la base est en fait l'un des objectifs de recherche du consortium).

Le but de ce projet est donc de construire une base de données apte à stocker au minimum les informations précédentes et qui soit utilisable pour traiter les questions suivantes :

1. Quels sont les complexes de protéines de la base contenant une protéine donnée ?
2. Quels sont les complexes de protéines qui ne s'expliquent pas *via* les domaines protéiques actuellement connus ?
3. La collaboration entre un labo de bio structurale et un labo sur la signalisation a permis de mettre en lumière deux domaines protéiques induisant une affinité entre deux protéines. Comment compléter la base ?
4. Deux protéines données (dont l'interaction n'est pas nécessairement déjà observée dans une voie de signalisation) ont-elles des chances d'interagir ?
5. Un laboratoire découvre une nouvelle affinité protéique participant à une voie de signalisation, sans pour autant connaître les domaines impliqués. Comment l'ajouter à la base ?
6. Quelles voies de signalisation raisonnablement courtes permettent d'aller d'une protéine à une autre ?
7. Étant donné un complexe de protéines, quelles sont les protéines susceptibles d'interagir avec ce complexe ?

On notera que pour les questions 3 et 5, tant les domaines que les protéines impliquées peuvent ou non être déjà connues du consortium, donc préalablement présents ou non dans la base.

D'autres questions pourront éventuellement compléter ce panel.

Projet n°2

OGM et pesticides

Il s'agit de mettre en place une base de donnée pour un laboratoire d'un organisme de recherche qui souhaite évaluer l'effet cancérigène éventuel des OGM et des pesticides sur les mammifères.

Cet organisme de recherche dispose de plusieurs champs de culture ayant des surfaces et des conditions environnementales similaires et y cultive diverses céréales :
- chaque champ ne contient qu'un seul type de céréale, ayant éventuellement subi une modification génétique donnée,
- ces cultures peuvent être faites avec ou sans divers pesticides mais il y a au plus un seul pesticide par champ,
- pour chaque céréale, il y a un champ « bio » de référence.

Par ailleurs, les récoltes servent à alimenter des populations de mammifères de différentes espèces herbivores. Chaque espèce peut être alimentée par certaines des céréales sus-mentionnées et pour chacune d'elles il y a un sous-groupe alimenté uniquement « bio ». Pour chacun des sous-groupes d'alimentation on connaît :
- quelle est la proportion de chaque céréale constituant son alimentation (toujours identique pour une même espèce) ; une seule d'entre elles peut être OGM et une seule d'entre elles peut être avec pesticides
- quels sont les champs d'où provient l'alimentation sachant que pour une céréale donnée, chaque sous-groupe ne puise que dans un seul champ
- le nombre de cancers observés à l'issue de l'expérience.

Le but de ce projet est de construire une base de données relationnelle prenant en compte au minimum les informations précédentes. La base de données devra être utilisable pour traiter en particulier les questions suivantes :

1. Quelle est l'espèce la plus exposée aux effets d'un OGM donné ? ou d'un pesticide donné ?
2. Quelle est la combinaison la plus dangereuse pour les mammifères en général ?
3. Même question pour une espèce donnée,
4. Étant donnée une combinaison OGM/pesticide, quelles espèces sont les plus vulnérables (par ordre décroissant de gravité).
5. A l'issue d'une expérience, et après avoir nourri un certain temps en « bio » l'espèce considérée, le laboratoire met en place un nouveau régime pour cette espèce :

   comment mémoriser les résultats de l'ancienne alimentation ? et passer l'espèce à l'alimentation « bio » ?
6. Comment initialiser par la suite les données de la nouvelle expérience ?
7. Comment introduire un changement de culture dans un champ donné ?
8. Même question pour l'arrivée d'une nouvelle espèce de mammifères.

D'autres questions pourront éventuellement compléter ce panel.

Projet n°3

Filière biologique expérimentale d'épuration des eaux

Une station d'épuration des eaux a développé un circuit de bassins expérimentaux dans lesquels une évolution rapide des bactéries est favorisée par divers procédés. Elle vous demande de mettre en place une base de données lui permettant de surveiller l'évolution et l'efficacité de ce circuit de bassins pour la dépollution des eaux. On ne gère pas ici tous les détails des techniques « d'évolution dirigée » que peuvent employer les chercheurs pour améliorer l'efficacité du circuit de bassins.

La base de données doit pouvoir stocker les données issues de :

- Une analyse mensuelle du métagénome présent dans ses bassins. Pour simplifier, on supposera qu'il s'agit de la liste des gènes présents, avec une mesure normalisée de nombre d'occurrences. On n'a aucune connaissance de l'espèce de bactérie dont ils proviennent mais on dispose de deux classes de métagénomes, ceux provenant du milieu aérobie et ceux provenant du milieu anaérobie.
- Une description textuelle courte, mensuelle, des procédés d'évolution dirigée appliqués durant le mois.
- Un inventaire des espèces pathogènes connues avec, pour faire simple, la liste des gènes « marqueurs » de chacune de ces espèces.
- Une analyse chimique mensuelle de l'eau sortant de ce circuit de bassins, qui mesure les polluants résiduels (divers composés carbonés, perturbateurs endocriniens, *etc*)

On suppose que les conditions expérimentales assurent un flux d'entrée d'eaux polluées constant et invariable.

La base de données devra être utilisable pour traiter en particulier les questions suivantes :

1. Quelles sont les espèces pathogènes présentes à un mois donné ? (celles dont un nombre suffisant de gènes marqueurs de l'espèce sont présents)
2. Quelle est la liste des pathogènes qui augmentent en nombre d'un mois donné au mois suivant ?
3. Même question avec chacune des molécules polluantes répertoriées dans la base
4. Pour un polluant donné, quelle est la liste des gènes (non issus de pathogènes) qui covarient avec ce polluant ?
5. Quels sont les gènes, s'il y en a, qui participent potentiellement à une diminution des perturbateurs endocriniens ?
6. Par ailleurs, bien sûr, il faut pouvoir en fin de mois :
   ajouter une nouvelle analyse du métagénome
7. et mettre à jour l'application d'un procédé d'évolution dans un bassin donné
8. et ajouter une nouvelle analyse de l'eau sortant du circuit de bassins

D'autres questions pourront éventuellement compléter ce panel.

Projet n°4

Cahier de laboratoire

Un assez grand service de R&D, qui utilise actuellement des cahiers de laboratoire « papiers », se trouve handicapé par la difficulté à retrouver des protocoles expérimentaux déjà validés, à leur associer les données expérimentales obtenues ou encore les conclusions auxquelles elles ont permis d'arriver. En fait retrouver des informations dans ces cahiers repose sur le souvenir des membres du service de R&D, et devient de plus en plus difficile avec les mouvements de personnel.

Ce service de R&D vous demande donc de créer une base de donnée apte à mémoriser le contenu d'un cahier de laboratoire partagé, et capable de retrouver les informations utiles par interrogations de cette base de données.

Ce « cahier électronique » doit répertorier des *expériences*, qui participent à des *projets* du service de R&D (identifiés par leur nom). Une même expérience peut éventuellement participer à plusieurs projets, et par ailleurs elle peut dépendre d'autres expériences qui doivent être menées avant de la commencer. Chaque expérience contient une description d'un ou plusieurs *protocoles*, elle fournit des *données expérimentales* et, à la suite de l'expérience, des *données traitées* peuvent être produites, et dans tous les cas une *analyse des résultats* est consignée. Une même expérience peut être réitérée à différentes dates, donc avec des données et des analyses distinctes. Les données expérimentales et les données traitées sont mémorisées dans des fichiers à part et la base de données mémorise seulement l'adresse de ces fichiers. Un protocole est décrit sous forme de plusieurs textes successifs mémorisés dans la base, qui décrivent les étapes du protocole, à l'exception du premier texte, qui décrit les objectifs de l'expérience. Enfin une analyse de résultats est un simple texte mémorisé dans la base de données.

Le but de ce projet est de construire une base de données prenant en compte les informations précédentes. Afin de faciliter la recherche d'informations au sein de ce cahier de laboratoire informatisé, la base de données devra être utilisable pour traiter en particulier les questions suivantes :

1. Étant donnée une expérience, à quels projets a-t-elle participé ?
2. Quelles sont toutes les expériences qui participent à un projet donné ?
3. Étant donné un mot clef, quelles sont toutes les expériences dont les objectifs contiennent ce mot ?
4. Même question avec l'analyse des résultats, mais dans ce cas on veut aussi connaître la date de l'expérience.
5. Ajouter un projet dans le cahier, et de même ajouter une expérience rattachée à un projet donné.
6. Étant donnée une expérience, la rattacher à un projet supplémentaire.
7. Étant donné un projet, trouver le ou les projets les plus similaires, c'est-à-dire qui partagent le plus grand nombre d'expériences en commun
8. Modifier le contenu d'une expérience (ses objectifs, son protocole, ajouter des données expérimentales ou traitées, ajouter une analyse de résultats, ajouter une expérience préalable nécessaire, *etc*).

D'autres questions pourront éventuellement compléter ce panel.