

Cours d'Informatique de Base
2ème Candidature en Sciences Economiques et
Sociales – Option Informatique
Travaux Pratiques:
Système de Gestion de Bases de Données
Bibliographiques Hétérogènes
Année Académique 2000-2001

Michaël Petit

20 février 2001

1 Préambule

Ce document décrit le travail à réaliser dans le cadre des travaux pratiques du cours d'Informatique de Base de seconde candidature en Sciences Economiques et Sociales, option Informatique.

La section 2 décrit le programme qui devra être réalisé. La section 3 décrit les modalités pratiques du TP telles que le planning et les rapports qui seront à remettre.

2 Enoncé de l'Exercice

Le programme à réaliser est un programme de gestion de bases de références bibliographiques hétérogènes. Après avoir défini ce que sont des références bibliographiques (section 2.1), des bases de références bibliographiques hétérogènes (section 2.2) et les particularités du système de gestion à réaliser dans ce TP (section 2.3), nous décrirons les problèmes posés par ce genre de bases de références (section 2.4). Le problème principal de l'hétérogénéité est la difficulté de comparer les éléments de ces bases. Pour contourner ce problème, une notion de similitude entre éléments d'une base de références bibliographiques sera définie dans la section 2.5. Finalement, nous décrirons les fonctionnalités que le programme devra réaliser (section 2.7).

Kathleen Jensen, Niklaus Wirth, *Pascal : User manual and report, second edition*, Springer-Verlag, 1978, ISBN 0-387-90144-2.

FIG. 1: Exemple de référence bibliographique.

2.1 Références Bibliographiques

Une référence bibliographique est un ensemble d'informations qui définissent un document (le plus souvent écrit) et qui permettent de référencer (citer) ce document dans un autre document. Des exemples de document sont un livre, un article publié dans une revue scientifique, un recueil d'articles (livre regroupant plusieurs articles), etc. Des exemples d'informations contenues dans une référence bibliographique sont les noms des auteurs, l'année de publication, le titre, l'éditeur, etc. Un exemple de référence bibliographique est donné à la figure 1.

Une référence bibliographique définit le document de la façon la plus complète possible. Les informations contenues dans une référence dépendent du type de document. Par exemple, un recueil d'articles scientifiques sera principalement défini par un titre, un ou des éditeurs (personnes qui recueillent et assemblent les contributions des différents auteurs), une année de publication, une maison d'édition, etc. Un article publié dans un recueil d'articles sera lui non seulement défini par les informations propres au recueil dans lequel il est publié mais également par une série d'informations propres à l'article lui-même (un titre, un ou des auteurs, les numéros de pages auxquels on peut trouver l'article dans le recueil, etc.

Certains formats "standards" de références bibliographiques identifient un certain nombre de types de documents fréquemment rencontrés et définissent les types d'informations qui doivent (et qui peuvent) être présentes pour chaque type de document afin que la référence soit correctement définie. La description d'un tel format pour des documents scientifiques (le format BibTeX) est donnée en annexe à titre d'illustration.

2.2 Bases de Références Bibliographiques Hétérogènes

Pour écrire des documents, un auteur utilise de manière répétitive un certain nombre de références bibliographiques. Afin de faciliter les citations, les références sont souvent collectées dans une *base de références bibliographiques (BRB)* afin d'être facilement réutilisées d'un document à l'autre. La base de références peut également être utilisée à d'autres fins telles que la gestion d'une bibliothèque.

Chaque propriétaire d'une ou plusieurs bases de références bibliographiques peut utiliser un format qui lui est propre. Ce format peut être un fichier de texte dans le cas le plus simple, ou des fichiers propres à un langage (par exemple des fichiers de record Pascal) ou propres à un système particulier (par exemple une base de données Access) dans les cas plus complexes.

Un format peut permettre de représenter des informations que les autres ne permettent pas de représenter (comme par exemple la localisation d'un exem-

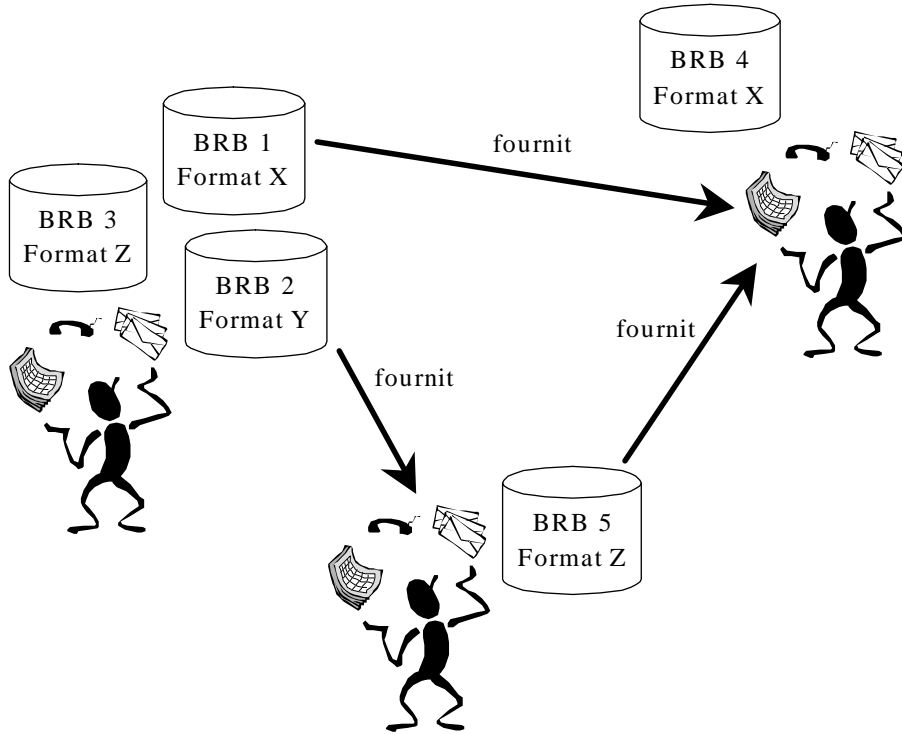


FIG. 2: Trois propriétaires et leurs BRB hétérogènes.

plaire d'un document dans une bibliothèque). On parlera donc de *bases de références bibliographiques hétérogènes*.

La constitution de bases de références bibliographiques est une activité coûteuse en temps (encodage, gestion, etc). En conséquence, les propriétaires de BRB désirent profiter de l'existence de BRB créées par d'autres propriétaires et que ceux-ci mettent à leur disposition, et ce malgré cette hétérogénéité.

La figure 2 illustre l'existence de BRB hétérogènes et la mise à disposition de certains propriétaires d'une ou plusieurs BRB par d'autres propriétaires.

2.3 Un Système de Gestion de Bases de Références Bibliographiques Hétérogènes

On désire implémenter un **système de gestion de bases de références bibliographiques hétérogènes (SGBRB)**. Idéalement, ce système devrait permettre de gérer et d'utiliser des BRB dans un grand nombre de formats différents, sans que l'utilisateur ait à se soucier du format dans lequel est représentée un BRB. Cependant, vu le temps limité dont nous disposons dans le cadre de ce TP, on ne considèrera qu'un seul format en faisant l'hypothèse simplificatrice suivante :

Chaque propriétaire de BRB qui désire faire profiter les autres d'une de ses BRB est capable de transformer celle-ci dans un format textuel équivalent. Ce format sera le seul géré par notre système.

Le format retenu devra être suffisamment expressif pour permettre à n'importe quel système de gestion de BRB de transformer le contenu de sa ou ses BRB dans ce format sans perdre trop d'information. Le format retenu sera donc le suivant.

Comme indiqué précédemment, une référence bibliographique est un ensemble d'informations. Chaque information est représentée par *un champ* de référence bibliographique. Chaque champ est composé de deux parties : *un nom* et un *contenu*. La valeur du nom de champ détermine le type d'information présente dans le contenu. Par exemple, si la valeur du nom de champ est "auteur", ceci indique que la valeur du contenu du champ est un nom d'auteur (par exemple "Baudouin Le Charlier").

Chaque BRB est contenue dans un fichier de texte. Un champ de référence bibliographique est représenté par une ligne du fichier. Dans une ligne du fichier, le nom de champ et son contenu sont séparés par un signe =. Une ligne de champ contient toujours un signe =. Tout ce qui précède le premier signe = de la ligne est considéré comme le nom de champ et tout ce qui suit est considéré comme le contenu du champ. Chacun de ceux-ci peut être vide mais les deux ne peuvent être vides simultanément.

Les champs d'une même référence bibliographique sont des lignes contiguës de la BRB. Deux références différentes sont séparées par une ou plusieurs lignes vides.

La figure 3 donne un exemple de contenu d'une BRB contenant trois références.

Notons que dans la plupart des SGBRB, un champ particulier sera généralement associé à chaque référence et sera utilisé pour identifier et citer le document. Ce champ est souvent appelé "clé" ou "identifiant", mais son nom peut varier d'un système à l'autre. La clé est normalement unique pour chaque référence d'une BRB (il n'y a pas deux références avec la même clé dans une BRB).

2.4 Conséquences de l'Utilisation de BRB Hétérogènes

Bien qu'un seul format soit considéré pour toutes les BRB, on peut considérer qu'il s'agit bien de bases hétérogènes. En effet, le type de contenu est différent d'une BRB à l'autre. Une BRB peut contenir des champs décrivant par exemple la localisation physique (dans une bibliothèque) d'exemplaires de documents référencé alors que dans une autre BRB, les référence aux mêmes documents ne contiendront pas nécessairement cette information, mais en contiendra peut-être d'autres. L'ordre des champs d'une référence peut également être différent d'une BRB à l'autre.

De plus, les noms de champs utilisés dans différentes BRB peuvent être différents. Ceci peut-être dû à l'emploi de synonymes (par exemple "auteur"

```
type-ref = livre
titre = Pascal: User manual and report
édition=second edition
auteur= Kathleen Jensen, Niklaus Wirth
éditeur = Springer-Verlag
année = 1978
isbn = 0-387-90144-2
isbn=3-540-90144-2
résumé =

= Germinal, Zola
= très bon livre!
localisation = 3ème travée, rayon "classiques"

type-ref = BD
titre = J'étais infirme hier
collection = Les femmes en blanc
numéro =5
dessinateur = Bercovici
scénariste=Cauvin
édition= Dupuis
style=humour
```

FIG. 3: Exemple de contenu d'une BRB.

```

type-ref = livre
titre = Pascal: User manual and report
  édition=second edition
auteur= Kathleen Jensen, Niklaus Wirth
éditeur= Springer-Verlag
  année = 1978
mots-clé = Pascal fichiers records programmation
  numéro-bump = #I4228/002D
identifiant=Pascal-manual

key=JENSEN78
title = Pascal: User manual and report
  edition=second edition
author= Kathleen Jensen, Niklaus Wirth
  isbn = 0-387-90144-2
isbn=3-540-90144-2
abstract = Everything you need to know about Pascal.
keywords= Pascal syntax files data-structures

= Pascal: User manual and report, Kathleen Jensen, Niklaus Wirth
=Springer-Verlag, 1978, isbn 0-387-90144-2
résumé =

```

FIG. 4: Trois références vers le même document

et "écrivain") ou à l'utilisation de langues différentes (par exemple, "auteur" et "author"). Certains champs peuvent également contenir des informations exprimées sous forme de synonymes (par exemple, deux références différentes peuvent utiliser dans le contenu de leur champ "mots-clés", les mots "vélo" et "bicyclette" qui sont synonymes).

En outre, il est possible que dans certaines BRB, certaines informations soient absentes (par exemple, les noms de champs peuvent ne pas être disponibles) ou structurées de façon différente (par exemple, le titre d'un recueil d'articles et le nom de son éditeur pourrait être groupé en un seul champ nommé "titre-recueil" dans une BRB et être décomposés en plusieurs champs dans une autre BRB).

Finalement, on considèrera également que le format des différentes références d'une même BRB pourra être variable. Il est en effet possible que des BRB aient été créées en concaténant les fichiers de plusieurs autres BRB qui ont des formats différents.

La figure 4 donne un exemple où trois références différentes mais similaires décrivent le même document.

Vu l'utilisation de plusieurs BRB de provenance différentes, il devient difficile d'éviter une certaine redondance entre les différentes BRB : des doubles (références différentes vers un même document) peuvent apparaître dans ces

différentes BRB, voir même à l'intérieur d'une seule BRB. Tous les éléments d'hétérogénéité décrits ci-dessus rendent la comparaison de références (et donc la recherche de doubles) difficile. Cette difficulté apparaît encore accrue si l'on considère les erreurs possibles suivantes :

- de fautes de frappe ou d'orthographe peuvent se glisser dans les références des différentes BRB. Ces fautes peuvent être des omissions ou ajouts d'un ou plusieurs caractères à un mot, des inversions de l'ordre des caractères dans un mot, l'utilisation d'un caractère incorrect en lieu et place d'un autre, l'utilisation de majuscules à la place de minuscules ou inversement et l'utilisation d'un caractères accentués en lieu et place du même caractère non accentué ou inversement.
- des mots peuvent être inversés par erreur dans le contenu des champs. Par exemple, "un TP amusant et enrichissant" à la place de "un TP enrichissant et amusant".
- des mots peuvent être omis ou ajoutés par erreur dans le contenu des champs. Par exemple, "vivre sur terre" à la place de "vivre sur la terre".

Le SGBRB à implémenter devra donc être capable de comparer des références en ne se basant pas seulement sur une stricte égalité champ à champ des références mais également en tenant compte de la similitude entre les références. La notion de similitude devra permettre une certaine tolérance aux erreurs et hétérogénéités décrites ci-dessus. Une telle notion sera utile par exemple pour découvrir des doubles dans une ou plusieurs BRB ou pour faire des recherches dans une BRB.

A cette fin, une méthode particulière de comparaison visant à déterminer efficacement la similitude entre éléments d'une BRB est décrite dans la section suivante. Cette méthode se base principalement sur des mesures quantitatives dérivées des éléments de la BRB. Elle permet de définir différents niveaux de similitude qui peuvent être choisis par l'utilisateur en fonction de l'objectif qu'il poursuit en comparant des éléments de la BRB.

2.5 Définition de la Notion de Similitude entre les Éléments d'une BRB

En fonction des objectifs poursuivis lors d'une comparaison d'éléments d'une ou plusieurs BRB, différentes notions de similitudes peuvent être utiles. Par exemple, pour comparer des clés attachées à des références, une égalité stricte sera nécessaire alors que pour rechercher des références sur base d'un titre de livre, plus de tolérance au fautes de frappe et à l'inversion de mots sera la bienvenue.

Il nous est donc parru judicieux de laisser l'utilisateur du SGBRB spécifier quel degré de similitude il désire utiliser lorsqu'il compare ou recherche des éléments de la BRB, et lui laisser la possibilité de modifier ce degré de similitude dynamiquement en fonction de ses objectifs.

Dans la suite, nous commençons par décrire comment l'utilisateur pourra définir différents degrés de similitude entre des éléments de base d'une BRB, à

savoir, les mots. Ensuite, on s'appuyera sur cette définition pour définir des niveaux de similitude pour des éléments plus complexes (les phrases et les champs), pour enfin définir des niveaux de similitude entre références.

2.5.1 Similitude des Mots

Les mots d'une BRB sont des suites non interrompues de caractères non blancs. Donc, dans un souci de simplification, on ne traitera pas le cas de mots composés.

La notion de similitude entre des mots ne tiendra pas seulement compte de la syntaxe de représentation des mots (les lettres qui composent les mots) mais aussi du sens des mots. A cette fin, on se basera sur une liste de synonymes.

Définition des Synonymes On supposera l'existence d'un fichier contenant la liste des mots synonymes (incluant également des mots équivalents dans des langues différentes). La relation de synonymie est définie sur l'ensemble des mots. La relation de synonymie étant réflexive, si le couple (a,b) est présent dans le fichier, le couple (b,a) y sera également. De plus, la relation étant transitive, si (a,b) et (b,c) se trouvent dans la relation, (a,c) y sera également, et par conséquent (c,a) y sera aussi. On s'assurera que la relation de synonyme respecte à tout moment ces contraintes.

La relation de synonymie sera représentée dans un fichier de texte dans lequel chaque ligne représentera la liste de tous les mots qui sont synonymes entre eux, séparés par un ou plusieurs espaces. Ce fichier assurera la pérennité de la relation de synonymie d'une exécution à l'autre. Le fichier des synonymes sera de taille quelconque et celle-ci pourra évoluer de manière non prévisible.

On veillera à ce que l'utilisation et la gestion des synonymes soit efficace. On devra donc nécessairement charger le fichier en mémoire au début de l'exécution du programme. On s'assurera aussi que la structure de donnée choisie en mémoire soit efficace.

Comparaison des Mots Pour comparer les mots, on considèrera leur synonymes, leur longueur (nombre de caractères significatifs) et le nombre de fois que chaque caractère significatif apparaît dans le mot. Par caractère significatif, on entend toutes les lettres de l'alphabet (accentuées ou non) et les chiffres. Lorsque le programme devra comparer des mots, l'utilisateur pourra spécifier le degré de similitude entre mots désiré en précisant :

- si la liste des synonymes peut être utilisée pour déterminer si les mots sont (sémantiquement) similaires. Si l'utilisateur précise que la liste de synonymes peut être utilisée et que les deux mots comparés apparaissent comme une paire de la liste, ils seront considérés comme similaires. Sinon, seul le contenu des mots (les lettres qu'ils contiennent seront utilisés pour déterminer s'ils sont similaires).
- une limite supérieure à la différence entre les longueurs des mots en caractères significatifs. Par exemple, si l'utilisateur précise une limite de 0.2, les

mots ne pourront être considérés comme similaires (en longueur) que si la différence entre leur deux longueurs en caractères significatifs ne dépasse pas 20% de leur longueur en caractères significatifs. Avec cette valeur, un mot de huit caractères significatifs ne pourra donc jamais être considéré comme similaire à un mot de cinq caractères significatifs.

- une limite supérieure à la somme des différences entre les fréquences respectives de chaque caractère significatif dans chacun des deux mots. Pour les lettres, la fréquence sera calculé indépendemment du caractère accentué ou non de la lettre et indépendemment du caractère majuscule ou minuscule de la lettre. Par exemple, si l'utilisateur donne la valeur 0.2, deux mots ne pourront être considérés comme similaires que si les fréquences respectives de chacun de leurs caractères significatifs ne diffèrent pas, au total, de plus de 20%.
- si l'ordre des caractères doit être strictement respecté. On remarquera qu'avec les deux dernières limites définies ci-dessus, l'utilisateur ne pourrait pas imposer l'équivalence stricte de deux mots car deux anagrammes (des mots dont l'un est constitué une permutation des caractères de l'autre) respecteront toujours les deux contraintes (leurs tailles et la fréquence des leurs lettres sont respectivement égales). Dans le cas particulier où l'utilisateur choisit des valeurs nulles pour ces deux paramètres, on lui permettra également d'imposer ou non la contrainte supplémentaire de respect strict de l'ordre des caractères.

2.5.2 Similitude de Phrases

Des phrases sont des suites de mots séparés pas des espaces. L'utilisateur pourra spécifier le degré de similitude désiré en précisant :

- une limite supérieure à la différence entre les longueurs des phrases en mots (le nombre de mots dans chacune). Par exemple, si l'utilisateur précise une limite de 0.2, les phrases ne pourront être considérés comme similaires (en longueur en mots) que si la différence entre leur deux longueurs en mots ne dépasse pas 20% de leur longueur. Avec cette valeur, une phrases de huit mots ne pourra donc jamais être considéré comme similaire à une phrase de cinq mots.
- une limite supérieure à la différence entre les longueurs des phrases en caractères significatifs. Par exemple, si l'utilisateur précise une limite de 0.2, les phrases ne pourront être considérés comme similaires que si la différence entre leur deux longueurs en caractères significatifs ne dépasse pas 20% de leur longueur en caractères significatifs. Avec cette valeur, une phrase de trente caractères significatifs ne pourra donc jamais être considéré comme similaire à une phrase de quarante caractères significatifs.
- une limite supérieure à la proportion de mots d'une des deux phrases qui ne sont similaires à aucun mot de l'autre phrase. Par exemple, si l'utilisateur précise une limite de 0.2, les phrases ne pourront être considérés

comme similaires que si moins de vingt pourcents des mots de cette phrase ne sont similaires à aucun mot de l'autre phrase.

- une limite supérieure à la somme des différences entre les fréquences respectives de chaque caractère significatif dans chacune des deux phrases. Par exemple, si l'utilisateur donne la valeur 0.2, deux phrases ne pourront être considérés comme similaires que si les fréquences respectives de chacun de leurs caractères significatifs ne diffèrent pas, au total, de plus de 20%.
- les paramètres utilisés pour déterminer la similitude entre les mots des deux phrases (utilisation ou non de la liste des synonymes, limite supérieure à la différence de taille en caractères significatifs, limite supérieure à la somme des différences de fréquences des caractères significatifs et respect strict de l'ordre).

Remarque : L'utilisateur devra veiller à ce que ces contraintes soient judicieusement combinées avec les contraintes sur la similitude des mots. Par exemple, si on permet que la similitude des mots soit basée sur les synonymes, des contraintes trop fortes pour la seconde et la dernière mesure (limite à la différence de longueur en caractères significatifs et limite des fréquences des caractères significatifs) ne seront pas adaptées.

2.5.3 Similitude de Champs

Un champ peut être considéré comme une agrégat constitué de deux phrases (une pour le nom du champ et une pour le contenu du champ). En fonction du but recherché lors de la comparaison des champs, l'utilisateur pourra préciser :

- si seuls les noms des deux champs doivent être similaires, si seuls les contenus des deux champs doivent être similaires ou si à la fois les noms et contenus des deux champs doivent être respectivement similaires.
- les paramètres utilisés pour comparer les noms des champs.
- les paramètres utilisés pour comparer les contenus des champs.

2.5.4 Similitude de Références

Une référence étant une suite de champs, l'utilisateur pourra déterminer le degré de similitude recherché en précisant :

- une limite supérieure à la différence entre les longueurs des références en champs (le nombre de champs dans chacune). Par exemple, si l'utilisateur précise une limite de 0.2, les références ne pourront être considérés comme similaires (en longueur en champs) que si la différence entre leur deux longueurs en champs ne dépasse pas 20% de leur longueur. Avec cette valeur, une référence contenant huit champs ne pourra donc jamais être considéré comme similaire à une référence contenant cinq champs.
- une limite supérieure à la différence entre les longueurs des références en mots. La longueur d'une référence en mots peut être calculés à partir de la

longueur du contenu de tous ses champs, en omettant les noms de champs. Par exemple, si l'utilisateur précise une limite de 0.2, les références ne pourront être considérés comme similaires (en longueur en mots) que si la différence entre leur deux longueurs en mots ne dépasse pas 20% de leur longueur en mots. Avec cette valeur, une référence de trente mots ne pourra donc jamais être considéré comme similaire à une référence de cinquante mots.

- une limite supérieure à la différence entre les longueurs des références en caractères significatifs. La longueur d'une référence en caractères significatifs peut-être calculé sur base de la longueur en caractères significatifs du contenu de tous ses champs, en omettant les noms de champs. Par exemple, si l'utilisateur précise une limite de 0.2, les références ne pourront être considérés comme similaires que si la différence entre leur deux longueurs en caractères significatifs ne dépasse pas 20% de leur longueur en caractères significatifs. Avec cette valeur, une référence d'une longueur de cent caractères significatifs ne pourra donc jamais être considéré comme similaire à une référence d'une longueur de cent-cinquante caractères significatifs.
- une limite supérieure à la proportion de champs d'une des deux références qui ne sont similaires à aucun champ de l'autre référence. Par exemple, si l'utilisateur précise une limite de 0.2, les références ne pourront être considérés comme similaires que si moins de vingt pourcents des champs de cette référence ne sont similaires à aucun champ de l'autre référence.
- une limite supérieure à la somme des différences entre les fréquences respectives de chaque caractère significatif dans chacune des deux références. Par exemple, si l'utilisateur donne la valeur 0.2, deux références ne pourront être considérés comme similaires que si les fréquences respectives de chacun de leurs caractères significatifs ne diffèrent pas, au total, de plus de 20%.
- les paramètres utilisés pour déterminer la similitude entre les champs de la référence (quelles parties du champ – nom et/ou contenu – doivent être similaires et quels paramètres sont utilisés respectivement pour comparer le nom et le contenu du champ).

2.6 Stratégie de Comparaison d'éléments de BRB

On remarquera que presque tous les critères de définition des degrés de similitudes définis ci-dessus utilisent des mesures quantitatives qui peuvent être calculées une fois pour toutes pour chaque référence d'une BRB. On calculera donc une seule fois ces mesures et on les réutilisera d'une exécution à l'autre du programme.

Dans l'utilisation des mesures pour déterminer la similitude entre des éléments de la BRB, on veillera à utiliser les mesures les moins coûteuse en calculs d'abord pour rendre la comparaison la plus efficace. Par exemple, pour comparer deux références, on utilisera d'abord les mesures de longueur (en champs,

en mots et en caractères significatifs) puisque ces mesures ne nécessitent que peu de calcul. Pour affiner le résultat de la comparaison, on appliquera ensuite d'autres mesures telles que la similitude entre les champs. De cette façon, on évitera dans la plupart des cas de nombreux calculs inutiles.

2.7 Fonctionnalités du SGBRB

Le SGBRB devra fournir les fonctionnalités suivantes. Chaque fonctionnalité est décrite de manière informelle en termes des informations nécessaires (paramètres en entrée) et informations produite (résultats).

On utilisera obligatoirement la notion de similitude définie ci-dessus pour toutes les fonctionnalités qui nécessitent une recherche ou une comparaison dans la BRB. Dans la mesure du possible, on n'accédera pas au contenu du fichier de la BRB lui-même mais on utilisera uniquement les données quantitatives décrites ci-dessus. On ajustera éventuellement les paramètres afin d'obtenir le degré de similitude adapté à chaque fonctionnalité.

2.7.1 Création d'une BRB vide

Paramètres en Entrée Un nom de BRB.

Résultats Une nouvelle BRB vide dont le nom est celui donné en entrée.

2.7.2 Ajout d'une référence à une BRB

Paramètres en Entrée Une BRB, une référence sans champ clé, une valeur de clé.

Résultats Si la BRB en entrée ne contient pas de référence ayant comme valeur de clé la même valeur que celle données en entrée, la référence est ajoutée à la BRB. Sinon, la BRB est inchangée.

2.7.3 Suppression d'une référence d'une BRB

Paramètres en Entrée Une BRB, une valeur de clé.

Résultats Si la BRB en entrée contient une référence ayant comme valeur de clé la même valeur que celle données en entrée, la référence correspondante est considérée comme supprimée. Afin que cette fonctionnalité reste efficace, on veillera à ne pas supprimer la référence directement mais à faire toutes les suppressions ensembles au moment opportun.

2.7.4 Vérification de l'existence d'une référence à un document dans une BRB

Paramètres en Entrée Une référence, une BRB, une liste de synonymes, les paramètres définissant le degré de similitude désiré des mots, phrases, champs et références.

Résultats La liste de toutes les références similaires à la référence donnée.

2.7.5 Rechercher les doubles dans deux BRB

Paramètres en Entrée Deux BRB, une liste de synonymes, les paramètres définissant le degré de similitude désiré des mots, phrases, champs et références.

Résultats La liste de toutes les paires de références similaires. Chaque paire contient une référence de la première BRB et une de la deuxième.

2.7.6 Recherche de tous les documents d'un auteur dans une BRB

Paramètres en Entrée Une BRB, une liste de synonymes, un nom d'auteur, les paramètres définissant le degré de similitude désiré du nom d'auteur (qui est un mot).

Résultats La liste de toutes les références dont le champ auteur ou un champ synonyme est similaire (au vu du degré de similitude donné en entrée) avec le nom d'auteur spécifié en entrée.

2.7.7 Fusion de deux BRB

Paramètres en Entrée Deux BRB, une liste de synonymes, les paramètres définissant le degré de similitude désiré des mots, phrases, champs et références.

Résultats Une nouvelle BRB contenant toutes les références des deux BRB dans laquelle aucune référence n'est similaire à une autre.

2.7.8 Ajout d'une paire de mots à la relation de synonymie

Paramètres en Entrée La relation de synonymie, une paire de mots synonymes.

Résultats La relation de synonymie est mise à jour pour que les deux mots de la paire donnée en entrée apparaisse comme synonymes.

2.7.9 Suppression d'une relation de synonymie

Paramètres en Entrée La relation de synonymie, deux listes de mots dont tous les mots sont actuellement synonymes.

Résultats La relation de synonymie est mise à jour pour que les mots qui apparaisse dans chacune des listes restent synonymes entre eux mais que les mots qui sont dans des listes différentes ne soient plus synonymes.

2.7.10 Modification des paramètres de niveau de similitude

Paramètres en Entrée les paramètres définissant le degré de similitude désiré des mots, phrases, champs et références.

Résultats Les paramètres définissant le degré de similitude désiré des mots, phrases, champs et références, modifiés par l'utilisateur.

3 Modalités Pratiques

3.1 Séries et Groupes

Pour le TP, afin de mettre en évidence la difficulté supplémentaire amenée par le travail en groupes, on optera pour la structure suivante. Les étudiants seront répartis en quatre *séries* de 12 à 15 étudiants. Chaque série devra réaliser la totalité du TP et est décomposée en quatre *groupes* de 3 à 4 étudiants. Chaque groupe réalisera une partie du TP en fonction d'une répartition choisie par l'ensemble de la série. A l'intérieur d'un groupe, les tâches sont réparties comme le groupe l'entend mais les rapports doivent mentionner les noms des étudiants qui ont participé à chaque partie des rapports.

3.2 Rapports

Le TP se déroulera en plusieurs étapes détaillées dans la section 3.3. Chaque étape sera réalisée en série ou en groupe et donnera lieu à un rapport (par série ou par groupe). Les **dates de remise** et le **contenu précis** des rapports sont spécifiées pour chacune des étapes dans la section 3.3.

Pour chaque rapport, une **page de garde** respectera le format prédéfini fourni dans le fichier Word qui accompagne ce document. Cette page de garde précisera notamment le numéro du groupe ou de la série et l'identité de ses membres, le titre du rapport et la date de remise.

Il est fondamental que chaque groupe et série **garde une copie** de tous les rapports remis. Certains rapports seront en effet utilisés pour les étapes suivantes du TP. Il vous appartient de faire ces copies vous-même.

3.3 Etapes et Déroulement du TP

Le travail sera réalisé en suivant les étapes décrites ci-dessous.

3.3.1 Conceptualisation/structuration du problème

Objectif Sur base de l'énoncé et des techniques vues au cours, chaque série élabore une conceptualisation et une structuration du problème. Le problème est décomposé en modules qui permettraient une répartition de travail équitable entre les groupes de la série pour l'étape suivante.

Rapport à remettre Un rapport sera remis par série contenant :

1. une conceptualisation du problème comprenant :
 - (a) un diagramme de contexte pour le programme ;
 - (b) un diagramme fonctionnel pour toutes les fonctions principales du programme (identification des entrées/sorties et décomposition en sous-fonctions) ;
 - (c) un diagramme fonctionnel pour toutes les sous-fonctions identifiées (entrées/sorties) ;

- (d) une description de tous les types de données identifiés ;
 - (e) un diagramme de flux pour chaque fonction décomposée en sous-fonctions.
2. une structuration du problème en terme de modules comprenant :
 - (a) une définition du contenu de chaque module (liste des fonctions et types de données dans le module) ;
 - (b) l'idée sous-jacente au groupement des éléments dans chaque module ;
 - (c) une justification de la découpe en modules en terme de couplage et cohésion.

Date de remise Mercredi 28/2

Solution Après la remise des rapports, une solution élaborée par les enseignants vous sera remise. Elle servira de base pour l'étape suivante. En préparation à l'examen, il vous appartiendra de comparer de manière critique votre solution à la solution proposée par les enseignants.

3.3.2 Spécification des sous-problème

Objectif La série décide d'une répartition des modules définis dans la solution proposée par les enseignants à ses différents groupes. Chaque groupe spécifie ensuite son/ses modules avec les techniques rappelées au cours. On essaye d'éviter les interactions entre groupes au sein de la série lors de cette étape.

Rapport à remettre Un rapport sera remis par groupe contenant :

1. les spécifications du contenu du/des modules (fonctions et données avec l'approche la plus appropriée (pré/post, type abstrait,...). Chaque ligne de la spécification sera numérotée afin de rendre possible l'étape suivante du TP (révision de spécifications).
2. une identification de chaque problème rencontré (question soulevée) lors de la rédaction des spécifications. Chaque problème sera décrit par :
 - (a) une explication du problème ou de la question soulevée ;
 - (b) la cause du problème (incomplétude ou ambiguïté de l'énoncé en langage naturel, incomplétude des graphiques de la conceptualisation, problèmes liés à la structuration adoptée,...) ;
 - (c) comment le problème a été résolu (questions aux enseignants sur la signification de l'énoncé, complétion ou détail de la conceptualisation,...).

Date de remise Mercredi 21/3.

3.3.3 Révision des spécifications

Objectif Sur base du rapport remis à l'étape précédente, des révisions de spécifications sont réalisées dans chaque série. Les spécifications réalisées par chaque groupe sont données à une autre groupe de la série. Lors d'un réunion, chaque groupe tente d'identifier un maximum d'erreurs sur base :

- de la spécification qu'il a préalablement reçue et lue.
- d'un ensemble de règles de rédaction de bonnes spécifications données par l'enseignant.

Rapport à remettre Un rapport de révisions est remis par groupe. Il décrit chaque erreur rencontrée en terme de :

1. sa localisation dans la spécification (numéro de page et de ligne) ;
2. le numéro de la règle qu'elle viole ;
3. une estimation de sa gravité ;
4. l'identification de sa cause possible, si approprié (imprécision de l'énoncé, faiblesse de la conceptualisation, . . .).

Date de remise Mercredi 28/3.

Solution A la fin de cette étape, la spécification élaborée par les enseignants sera distribuée. Elle servira de base pour l'étape suivante (implémentation et tests).

3.3.4 Implémentation et Définition des Jeux de Tests

Objectif Chaque module identifié dans la solution spécifiée par les enseignants est attribué à un groupe de la série. Chaque groupe implémente ensuite ce module en Pascal. Pour cela, il se base sur la définition des structures de données définies par les enseignants.

Chaque groupe définit des jeux de tests pour un certain nombre de procédures indiquées par les enseignants en utilisant les approches vues au cours (black box et white box). Les jeux de tests sont ensuite utilisés pour vérifier la correction du programme.

Rapport à Remettre Un rapport d'implémentation par groupe contenant :

1. le code Pascal correspondant au module traité par le groupe ;
2. une description des jeux de tests demandés définis en termes de :
 - (a) classes de valeurs d'entrées considérées et valeurs représentatives choisies ;
 - (b) valeurs de sorties correspondantes attendues ;
 - (c) chaque erreur détectée en les utilisant, en mentionnant la valeur d'entrée testée qui a permis de trouver l'erreur et la cause de l'erreur.

Date de remise Mercredi 2/5.

3.3.5 Intégration des implémentations

Objectif Les implémentations des différents groupes sont intégrées. Des jeux de test sont définis et utilisés pour s'assurer de la correction de l'implémentation intégrée (jeux de tests pour les fonctions principales).

Rapport à remettre Un rapport final d'implémentation par série qui contient :

1. le code final de la série (résultant de l'intégration des codes des groupes et de leur correction) ;
2. une description des jeux de tests définis en termes de :
 - (a) classes de valeurs d'entrées considérées et valeurs représentatives choisies ;
 - (b) valeurs de sorties correspondantes attendues ;
 - (c) chaque erreur détectée en les utilisant, en mentionnant la valeur d'entrée testée qui a permis de trouver l'erreur et la cause de l'erreur.
3. une évaluation finale du TP contenant une critique positive et négative des techniques utilisées dans le TP en termes de son adéquation à palier à la complexité du problème traité.

Date de remise Mercredi 11/5.

3.4 Planning des Cours

Le planning provisoire suivant indique les dates de cours théoriques et de TP ainsi que le contenu prévu de chacun de ceux-ci. Les dates de remise des rapports sont également reprises.

Mercredi 24/1	Théorie	Conceptualisation/structuration de problèmes
Mercredi 31/1	Théorie	Conceptualisation/structuration du problème + Spécification
Mercredi 7/2	Pratique	Conceptualisation/structuration du problème
Mercredi 14/2	Pratique	Conceptualisation/structuration du problème
Mercredi 21/2	Théorie	Spécification
Mercredi 28/2	Pratique	Spécification
	Rapport	Conceptualisation/structuration du problème
Mercredi 7/3	Pratique	Spécification
Mercredi 14/3	Théorie	Techniques de révision
Mercredi 21/3	Pratique	Révisions et élaboration de rapports de révision
	Rapport	Spécification
Mercredi 28/3	Théorie	Elaboration de jeux de test
	Rapport	Révisions de spécifications
		Congés de Pâques
Mercredi 18/4	Pratique	Implémentation par groupe
Mercredi 25/4	Théorie	Elaboration de jeux de tests (2)
Mercredi 2/5	Pratique	Elaboration de jeux de tests et Intégration des implémentations
	Rapport	Implémentation par groupe et tests
Mercredi 11/5	Théorie	Débriefing final : leçons apprises
	Rapport	Implémentation par série