UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

Master SIAD

Projet d'étude

Réalisation d'un outil d'évaluation de la qualité

Arnaud Giacometti, Patrick Marcel, Verónika Peralta

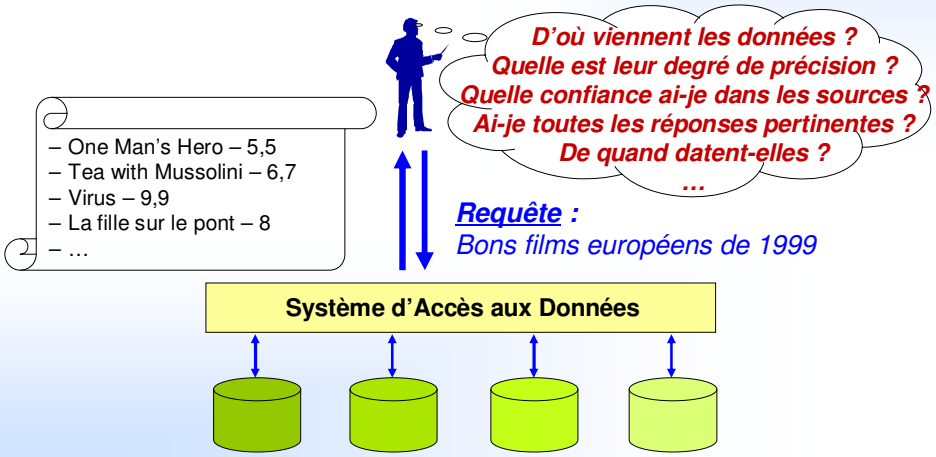
Novembre 2008

Informations générales

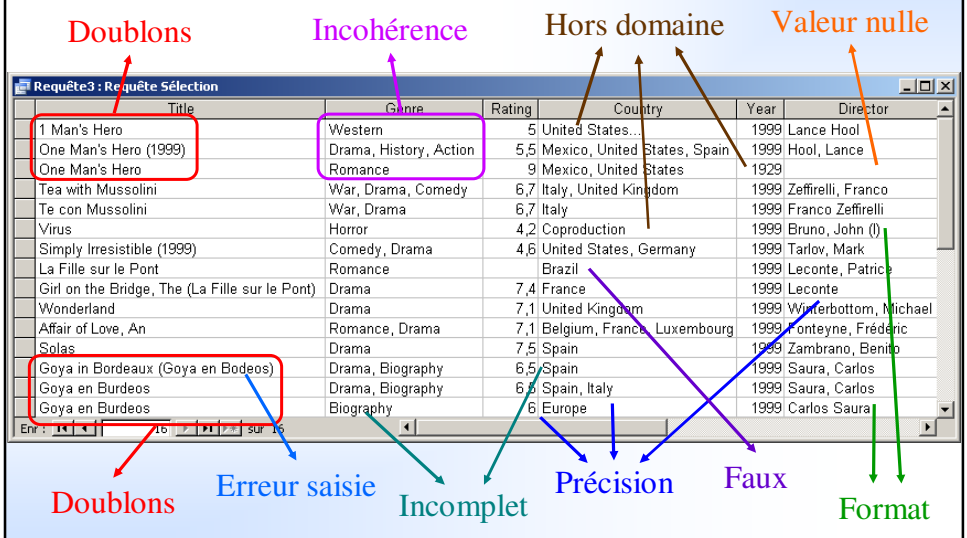
Enseignants	Arnaud Giacometti, Patrick Marcel et Verónika Peralta
Emails	prenom.nom@univ-tours.fr
Objectifs généraux	Faire une synthèse de vos connaissances acquises au travers d'un problème concret Faire découvrir de nouveaux concepts
Groupes	6 groupes de 4 ou 5 étudiants 1 chef de projet par groupe
Encadrement	Réunions pour faire le point sur l'avancement Présentations orales Rapports

Contexte

Interrogation des sources de données multiples, distribuées, autonomes, hétérogènes



Exemples de problèmes de qualité



Besoin : Afficher la qualité

- … des cellules, tuples, attributs ou le tout

Requête3 : Requête Sélection

Title	Genre	Rating	Country	Year	Director
1 Man's Hero	Western	5	United States...	1999	Lance Hool
One Man's Hero (1999)	Drama, History, Action	5,5	Mexico, United States, Spain	1999	Hool, Lance
One Man's Hero	Romance	9	Mexico, United States	1929	
Tea with Mussolini	War, Drama, Comedy	6,7	Italy, United Kingdom	1999	Zeffirelli, Franco
Te con Mussolini	War, Drama	6,7	Italy	1999	Franco Zeffirelli
Virus	Horror	4,2	Coproduction	1999	Bruno, John (I)
Simply Irresistible (1999)	Comedy, Drama	4,6	United States, Germany	1999	Tarlov, Mark
La Fille sur le Pont	Romance		Brazil	1999	Leconte, Patrice
Girl on the Bridge, The (La Fille sur le Pont)	Drama	7,4	France	1999	Leconte
Wonderland	Drama	7,1	United Kingdom	1999	Winterbottom, Michael
Affair of Love, An	Romance, Drama	7,1	Belgium, France, Luxembourg	1999	Fonteyne, Frédéric
Solas	Drama	7,5	Spain	1999	Zambrano, Benito
Goya in Bordeaux (Goya en Bodeos)	Drama, Biography	6,5	Spain	1999	Saura, Carlos
Goya en Burdeos	Drama, Biography	6,5	Spain, Italy	1999	Saura, Carlos
Goya en Burdeos	Biography	6	Europe	1999	Carlos Saura

Enr : 16 sur 16

Besoin : Afficher la qualité des données

- … avec diverses styles de visualisation

Requête3 : Requête Sélection

Title	Genre	Rating	Country	Year	Director
1 Man's Hero <0.8>	Western	5	United States...	1999	Lance Hool
One Man's Hero (1999) <0.5>	Drama, History, Action	5,5	Mexico, United States, Spain	1999	Hool, Lance
One Man's Hero <1>	Romance	9	Mexico, United States	1929	
Tea with Mussolini	War, Drama, Comedy	6,7	Italy, United Kingdom	1999	Zeffirelli, Franco
Te con Mussolini	War, Drama	6,7	Italy	1999	Franco Zeffirelli
Virus	Horror	4,2	Coproduction	1999	Bruno, John (I)
Simply Irresistible (1999)	Comedy, Drama	4,6	United States, Germany	1999	Tarlov, Mark
La Fille sur le Pont	Romance		Brazil	1999	Leconte, Patrice
Girl on the Bridge, The (La Fille sur le Pont)	Drama	7,4	France	1999	Leconte
Wonderland	Drama	7,1	United Kingdom	1999	Winterbottom, Michael
Affair of Love, An	Romance, Drama	7,1	Belgium, France, Luxembourg	1999	Fonteyne, Frédéric
Solas	Drama	7,5	Spain	1999	Zambrano, Benito
Goya in Bordeaux (Goya en Bodeos)	Drama, Biography	6,5	Spain	1999	Saura, Carlos
Goya en Burdeos	Drama, Biography	6,5	Spain, Italy	1999	Saura, Carlos
Goya en Burdeos	Biography	6	Europe	1999	Carlos Saura

Enr : 16 sur 16

précision: 1
complétude: 0.5
exactitude: 1

Plus précisément…

- **L'application devra permettre de :**
  - Mesurer la qualité des données stockées dans la BD
  - Stocker ces mesures de qualité
  - A chaque requête utilisateur:
    - Récupérer les mesures de qualité
    - Afficher les données avec leurs mesures de qualité

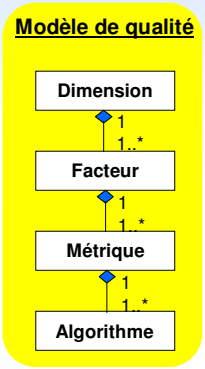
Qualité des données

Accuracy	Correction, Precision, Syntax, Level of detail, Error free.
Objectivity	Objectivity, Non ambiguity, Factuality, Impartiality…
Credibility	Credibility, Confidence…
Reputation	Reputation…
Access	System availability, Source availability, Ease of use, Localization…
Security	Security, Privileges…
Pertinence	Pertinence, Relevance, Utility…
Added value	Importance, Added value, Contents…
Freshness	Currency, Age, Volatility, Timeliness, Obsolescence…
Completeness	Density, Coverage, Suffisance…
Data quantity	Volume, Data quantity…
Interpretation	Interpretation, Modifiability, Traceability, Appearance, Presentation…
Comprehension	Comprehension, Readability, Clearness, Signification, Comparability…
Concise repr.	Minimality, Uniqueness, Concise representation…
Consistent repr.	Consistency, Format, Syntax, Alias, Semantics, Control of versions…

Qualité des données

- **La qualité est multi dimensionnelle**

- **Concepts importants:**
  - Une **dimension de qualité** décrive une facette de la qualité des données
  - Une **facteur de qualité** représente un aspect particulier d'une dimension
  - Une **métrique de qualité** décrit la façon de mesurer un facteur
  - Une **algorithme d'évaluation** calcule une métrique



Exactitude Complétude Fraîcheur Cohérence Unicité

Exactitude

- **Définition :**
  - La correction et la précision avec laquelle les objets du monde réel sont représentés dans le système d'information

- **Concerne :**
  - **Correction sémantique** : Indique si les données correspondent aux objets du monde réel.
  - **Correction syntaxique** : Indique si les données sont exemptes d'erreurs syntaxiques ou de format.
  - **Précision** : Indique si le niveau de détail des données est suffisant.

Exactitude

Correction syntaxique

Requête3 : Requête Sélection

Title	Genre	Rating	Country	Year	Director
1 Man's Hero	Western	5	United States...	1999	Lance Hool
One Man's Hero (1999)	Drama, History, Action	5,5	Mexico, United States, Spain	1999	Hool, Lance
One Man's Hero	Romance	9	Mexico, United States	1929	
Tea with Mussolini	War, Drama, Comedy	6,7	Italy, United Kingdom	1999	Zeffirelli, Franco
Te con Mussolini	War, Drama	6,7	Italy	1999	Franco Zeffirelli
Virus	Horror	4,2	Coproduction	1999	Bruno, John (I)
Simply Irresistible (1999)	Comedy, Drama	4,6	United States, Germany	1999	Tarlov, Mark
La Fille sur le Pont	Romance		Brazil	1999	Leconte, Patrice
Girl on the Bridge, The (La Fille sur le Pont)	Drama	7,4	France	1999	Leconte
Wonderland	Drama	7,1	United Kingdom	1999	Winterbottom, Michael
Affair of Love, An	Romance, Drama	7,1	Belgium, France, Luxembourg	1999	Fonteyne, Frédéric
Solas	Drama	7,5	Spain	1999	Zambrano, Benito
Goya in Bordeaux (Goya en Bodeos)	Drama, Biography	6,5	Spain	1999	Saura, Carlos
Goya en Burdeos	Drama, Biography	6,5	Spain, Italy	1999	Saura, Carlos
Goya en Burdeos	Biography	6	Europe	1999	Carlos Saura

Enr : 16 sur 16

Correction sémantique

Précision

Exactitude

- **Métriques :**
  - **Correction sémantique** :
    - Correction sémantique booléenne
      - 1= correcte – 0 = fausse
  - **Correction syntaxique** :
    - Correction syntaxique booléenne
      - 1= correcte – 0 = erronée
    - Déviation syntaxique
      - Déviation par rapport à une valeur correcte ; [0,1]
  - **Précision** :
    - Dégrée de précision
      - Appréciation du niveau de détail ; [0,1]

Comparaison avec la réalité ou un référentiel

Exemple:
- <client,telephone> appartient aux pages blanches

Règles de format ou dictionnaires

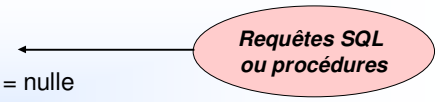
Exemples:
- Rue appartient à un catalogue
- Portable commence par 06

Hiérarchie de valeurs pour le domaine

Exemple:
- rue complète → 1
- manque numéro porte → 0.8

Complétude

- **Définition :**
  - Tous les objets pertinents du monde réel sont représentés dans le système d'information
- **Concerne :**
  - **Couverture** : Si tous les tuples sont représentés
  - **Densité** : Si tous les attributs sont représentés (pas nuls)
- **Métriques :**
  - Densité booléenne
    - 1= pas nulle – 0 = nulle



Fraîcheur

- **Définition :**
  - L'âge et l'actualité des données d'un système d'information
- **Concerne :**
  - **Age** : Indique le temps passé depuis la création des données
  - **Actualité** : Indique le temps passé depuis l'extraction des données
- **Métriques :**
  - Dégrée d'actualité :
    - Ratio entre le temps passé depuis l'extraction de la donnée et le temps de validité de la donnée

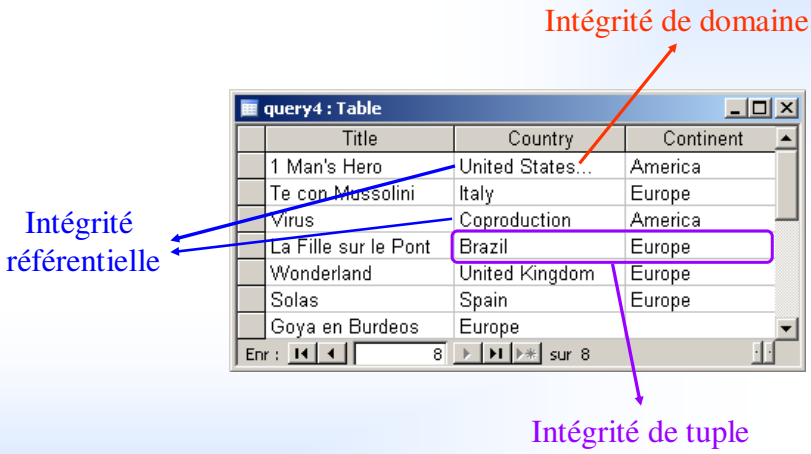
      Max { 0, 1 – temps_extraction / temps_validité }

Obs: On stockera la date d'extraction de chaque donnée et la validité de chaque attribut

Cohérence

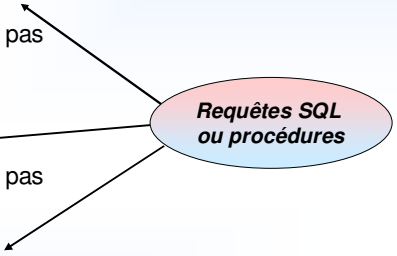
- **Définition :**
  - La satisfaction des règles d'intégrité d'un système d'information
- **Concerne :**
  - **Intégrité de domaine** : Indique si les données satisfont des règles de domaine
  - **Intégrité de tuple** : Indique si les tuples satisfont des règles inter-attributs
  - **Intégrité référentielle** : Indique si les tuples satisfont des règles d'intégrité référentielle (clés étrangères)

Cohérence



Cohérence

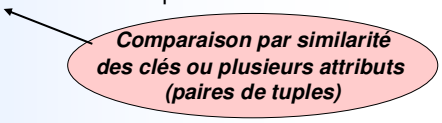
- **Métriques :**
  - **Intégrité de domaine** :
    - Intégrité de domaine booléenne
      - 1= satisfait – 0 = ne satisfait pas
  - **Intégrité de tuple** :
    - Intégrité de tuple booléenne
      - 1= satisfait – 0 = ne satisfait pas
  - **Intégrité référentielle** :
    - Intégrité référentielle booléenne
      - 1= satisfait – 0 = ne satisfait pas



Attention: Les 2 dernières métriques donnent des mesures pour chaque tuple

Unicité

- **Définition :**
  - Les données d'un système d'information ne sont pas dupliquées
- **Concerne :**
  - **Unicité** : Indique si les données sont uniques ou dupliquées
- **Métriques :**
  - Unicité booléenne
    - 1= unique – 0 = répétée
  - Degré d'unicité
    - Appréciation de la similarité à d'autres tuples



Attention: Les métriques donnent
des mesures pour chaque tuple

Modes de visualisation

- **Deux granularités de mesure :**
  - Détaillée : affichage des valeurs de chaque métrique
    - Ex. <déviation syntaxique = 0.8, dégrée d'actualité = 0.9, densité booléenne = 1>
  - Agrégée : affichage d'une valeur globale, obtenue en combinant les valeurs des métriques
    - Ex. 0.87
    - Formule:

      0.5 *déviation syntaxique + 0.3 *dégrée d'actualité + 0.2 *densité booléenne
- **Selon le type d'objet mesuré :**
  - Valeur d'attribut, tuple, attribut ou l'ensemble des résultats

L'utilisateur ...

- **peut :**
  - Choisir les métriques à visualiser
  - Choisir les poids pour obtenir des valeurs agrégées
  - Choisir les niveaux de qualité minimales pour chaque métrique
  - Ordonner les tuples par rapport à leur qualité
  - Filtrer les tuples qui n'ont pas une qualité suffisante
- **ne peut pas :**
  - Choisir les algorithmes d'évaluation à utiliser
  - Ajouter des nouvelles dimensions, facteurs et métriques

cela correspond aux programmeurs

  - Il faut prévoir les mécanismes

Résumé des fonctionnalités

- **L'application doit permettre :**
  - de se connecter à une base de données hébergée par un SGBD
  - de sélectionner/désélectionner les facteurs de qualité à visualiser et les métriques correspondantes
  - de sélectionner le mode de visualisation et des poids de combinaison
  - de poser des requêtes sur la base de données
  - de visualiser les résultats des requêtes en incluant de façon intuitive les niveaux de qualité des données
  - d'ordonner les résultats des requêtes par rapport à leur qualité
  - de filtrer des tuples du résultat qui n'ont pas une qualité suffisante
  - de configurer les algorithmes d'évaluation à utiliser
  - de mesurer plusieurs facteurs de qualité en utilisant plusieurs métriques
  - d'évaluer les requêtes sur la base de données et d'obtenir les résultats
  - de calculer des valeurs globales de qualité à partir des valeurs détaillées

D'autres précisions

- **Base de données de tests**
  - BD décrivant des films, acteurs, directeurs, cinémas...
    - Le schéma sera fourni
    - Quelques référentiels seront fournis
  - Il faut:
    - Collecter des données de plusieurs sites web
    - S'assurer d'avoir des doublons et des données erronées
- **Langages**
  - L'interface utilisateur et les schémas de bases de données seront en anglais
- **Outils informatiques**
  - A choisir par les étudiants

Phases du projet

1. **Etude bibliographique**
2. **Spécification (préparation)**
   - Cahier des charges (besoins, cadrage, planning prévisionnel, répartition des taches, analyse de risque), modalités des tests, indicateurs du reporting
3. **Analyse détaillée**
   - Cahier des charges, schéma UML
4. **Développement et test**
   - Choix des outils, développement
5. **Présentation**
   - Bilan du projet, rapport final, soutenance

Calendrier

Novembre 2008

L	M	M	J	V
3	4	5	6	7
10	11	12	13	14
17	18	19	20	21
24	25	26	27	28

Décembre 2008

L	M	M	J	V
1	2	3	4	5
8	9	10	11	12
15	16	17	18	19
22	23	24	25	26
29	30	31		

Janvier 2009

L	M	M	J	V
			1	2
5	6	7	8	9
12	13	14	15	16
19	20	21	22	23
26	27	28	29	30

Février 2009

L	M	M	J	V
2	3	4	5	6
9	10	11	12	13
16	17	18	19	20
23	24	25	26	27

Mars 2009

L	M	M	J	V
2	3	4	5	6
9	10	11	12	13
16	17	18	19	20
23	24	25	26	27
30	31			

Avril 2009

L	M	M	J	V
		1	2	3
6	7	8	9	10
13	14	15	16	17
20	21	22	23	24
27	28	29	30	

Mai 2009

L	M	M	J	V
				1
4	5	6	7	8
11	12	13	14	15
18	19	20	21	22
25	26	27	28	29