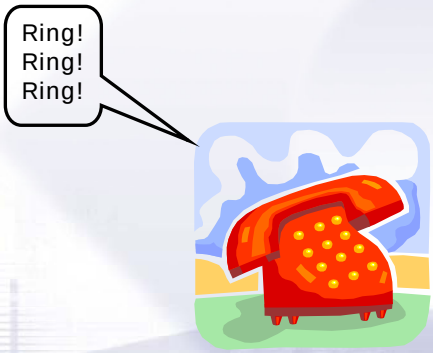
Data Quality II: Tools

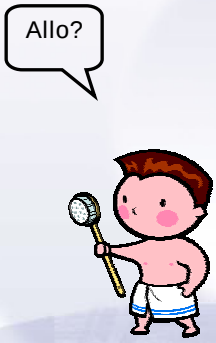
Yves Bontemps
SmalS-MvM, Section Recherches
21 septembre 2006

- Introduction
- Outils de Data Quality: concepts
  - Data Profiling
  - Standardisation
  - Matching
  - Monitoring
- Outils commerciaux
  - Case study
  - Architecture
- Conclusion

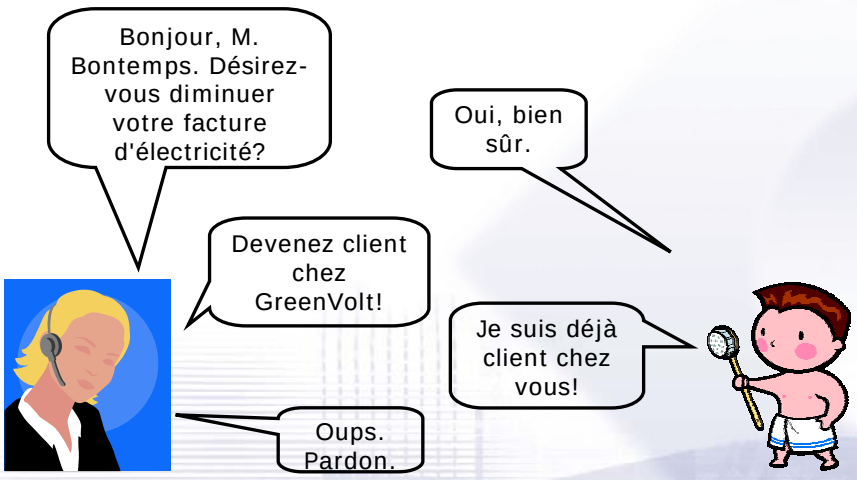
Introduction



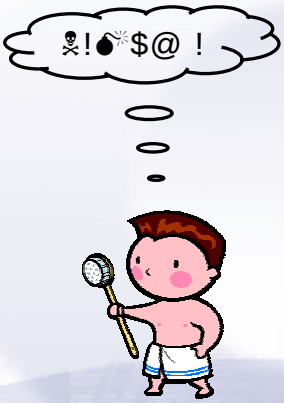
Introduction



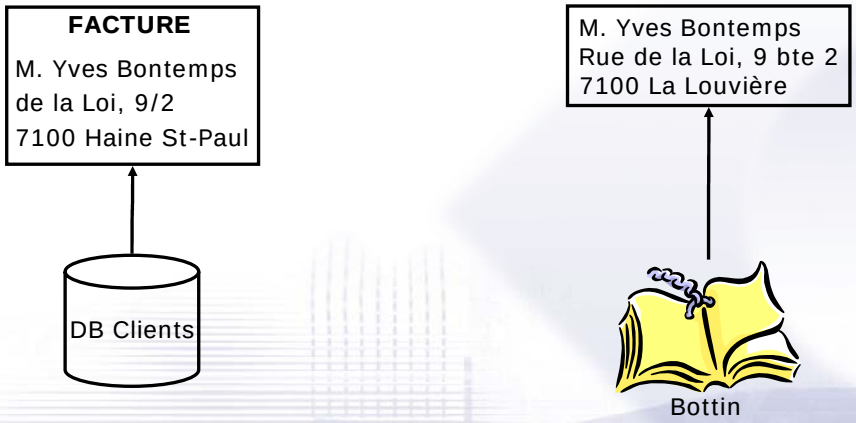
Introduction



Introduction



Introduction



Qualité des données
Définition

Fitness for use

- Remarques:
  - Fitness vs Perfection
  - Coûts-bénéfices
  - Use → présent & futur (!)

Introduction
Impacts/enjeux

- National Firearms Licensing Management Systems (UK)
  - *"During the pilot there were a number of data quality issues, which meant the system was returning errors, so the system was declined"*
  - *"If the Home Office really is incapable, over a period of eight years, of computerizing something as straightforward as a few hundred thousand firearms records, then it does suggest that they do not have a hope of making a success of the introduction of the national identity card scheme*"
- http://www.computing.co.uk/computing/news/2148820/delay-gun-register

Introduction
Impacts/enjeux

- Voter's registration system in California
  - Registration system of all voters, based on identification (driver's license). Checked against Calif. Dept of Motor Vehicles database.
  - "*The rigorous system will reject applications whose data doesn't exactly match the confirming documents. Even small discrepancies, such as a missing middle initial, could cause an application to be rejected.*"
  - "*The voter database has "been a disaster for anyone who is trying to register for the first time or reregister because they moved, got married and need to change their name or change parties,"*
  - http://www.computerworld.com/printthis/2006/0,4814,110861,00.html

Introduction
Impacts/enjeux

- Criminal Records Bureau (UK)
  - Check that someone has no criminal record prior to appointment (esp. unsupervised contacts with children).
  - *"The Criminal Records Bureau's first and foremost priority is to help protect children and vulnerable adults*"
  - *"The Criminal Records Bureau is only as effective as the information it can access."*
  - *Liberal Democrat home affairs spokesman Nick Clegg said the errors took "Home Office incompetence to new absurd levels". He added: "This latest fiasco will erase the last bit of public confidence in the Home Office."*
  - http://news.bbc.co.uk/1/hi/uk/5001624.stm

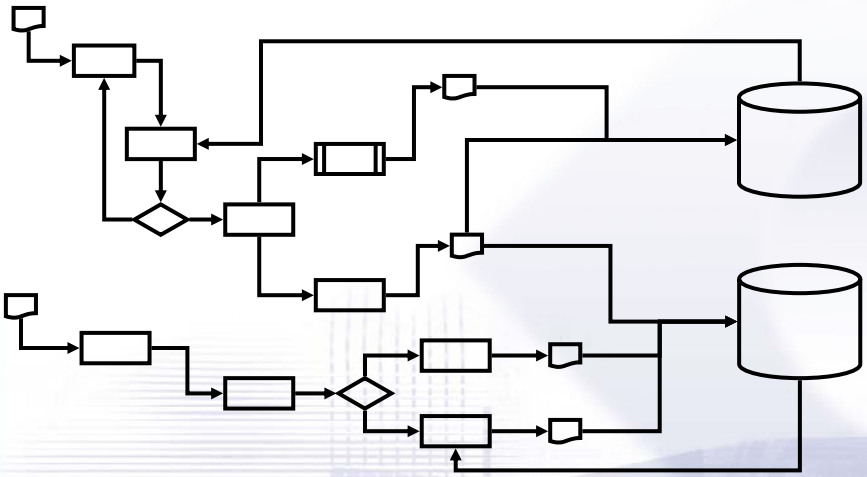
Introduction
Impacts de la qualité

- Coûts de correction (usine fantôme)
- Risques accrus → nouveaux dével.
- Décisions erronées
- Perte de confiance
- Abandon/Rejet du système

Introduction
Dimensions de la qualité

- Dimensions
  - Pertinente (*Relevant*)
  - Précise (*Accurate*)
  - Fraîche (*Timely*)
  - Complète (*Complete*)
  - Comprise (*Understood*)
  - Digne de confiance (*Trustworthy*)
- Contexte = utilisation des données

Introduction
Production des données



Introduction
Facettes d'une solution

Management

- Exec. DQ group
- Data Mgmt
- Data stewards
- ...

Information

- Meta-data
- Data Models
- Requirements
- Metrics, ...

Data as an Asset

Activités

- Root-cause analysis
- Data Integration
- Software Eng.
- Business Proc. Eng.
- Cleansing, etc.

Bénéfices

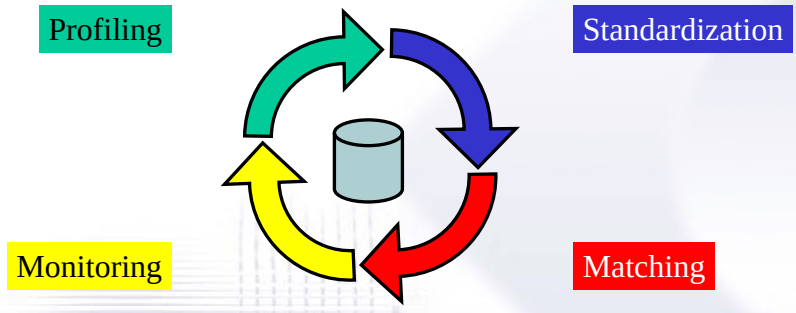
- Utilisateurs
- Développeurs
- Clients
- Affaires
- ...

Introduction
Outils Data Quality

- Marché existant et en forte croissance
- Question
  - Fonctionnalités proposées ?
  - Coûts vs Bénéfices ?
  - Par rapport dével. in-house?
- Place dans une approche globale?

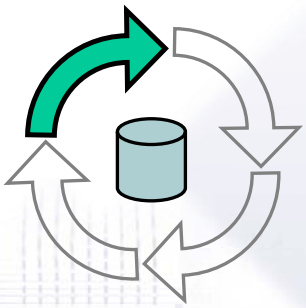
- Introduction
- **Outils de Data Quality: concepts**
  - Data Profiling
  - Standardisation
  - Matching
  - Monitoring
- Outils commerciaux
  - Case study
  - Architecture
- Conclusion

Cycle d'utilisation



Cycle d'utilisation

Profiling



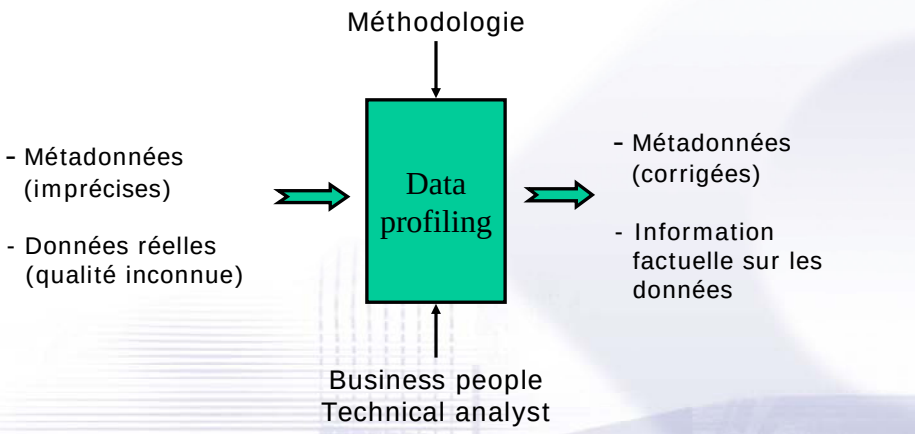
Data Profiling

- Expertiser l'état d'une base de données
  - Structure
  - Contenu
  - Validité < *Accuracy*
- Bird's eye view/ Rapport résumé

Data Profiling

- Information **sur** les données:
  - 10% des champs NUM_TEL sont vides
  - Le domaine de HEURES_REF est 3600 à 4000.
- Pas information **à partir** des données:
  - Les ouvriers de la construction sont significativement surreprésentés dans les accidents du travail.
  - 10% des contrats de travail dans l'enseignement durent moins de 3 mois.

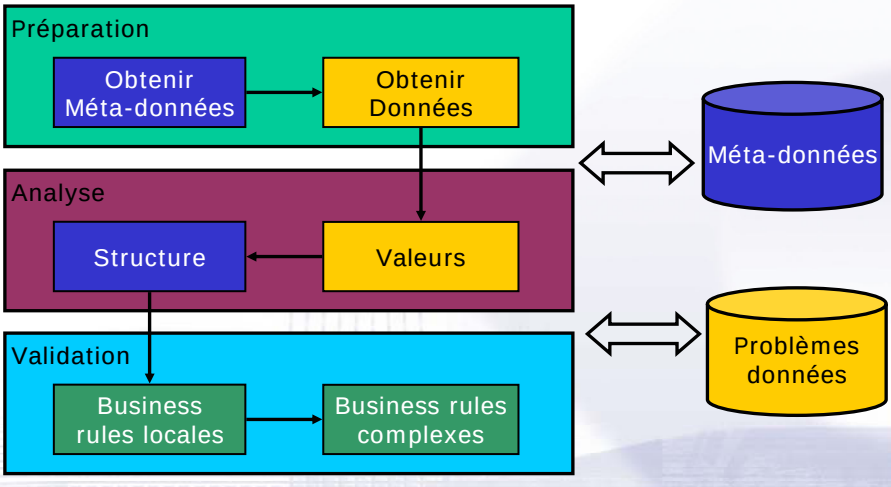
Data Profiling
Modèle



Data Profiling
Problèmes

- Métadonnées:
  - Inexistantes
  - Imprécises
  - Plus à jour
- Données:
  - Incohérentes avec métadonnées
  - Invalides/incorrectes
  - Contenu différent des attentes

Data Profiling
Processus



Data Profiling
Notes

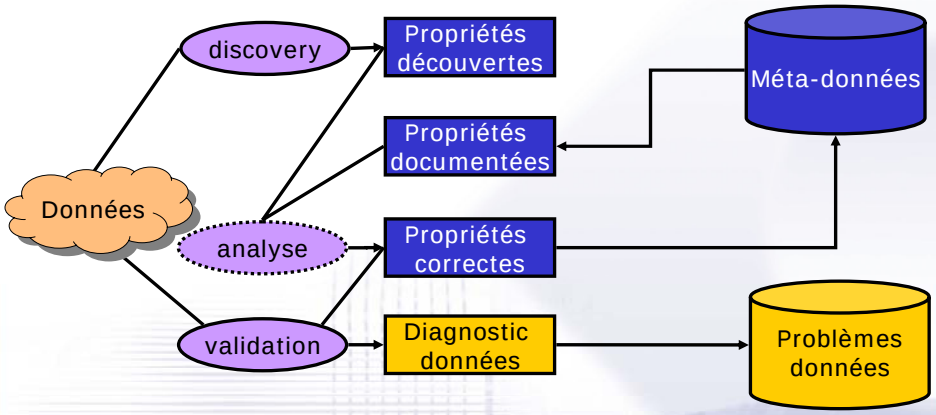
- Préparation
  - Rassembler les métadonnées, les participants, etc.
  - Extraction des données (!)
  - Effort principal
- Analyse & Validation
  - 2 phases pour chaque étape:
    - Discovery
    - Assertion testing

Data Profiling
Préparation (Méta-données)

- Systèmes de gestion de l'information
  - Dictionnaires des données (ex: glossaires)
  - Meta-data repositories
- Data definitions
  - Copybooks COBOL
  - Catalogs
- Logiques/Règles business
  - Programmes
  - Analyses fonctionnelles
  - Instructions aux utilisateurs
- People
  - DBA, Data Architect, Business Analyst

Data Profiling

2 phases: exemple (valeurs)



Data Profiling

Discovery

- Méta-données (valeurs champ)
  - Signification business
  - Type
  - Longueur/précision
  - Valeurs valides (liste/range)
  - Null autorisé
  - Unique
  - Séquence
  - Formatage
  - Conventions d'encodage (null, etc)

Data Profiling
Discovery

- Analyse automatique
- Niveau de complexité variable
- Détermine **ce qui est**, pas ce qui est possible ou ce qui est permis.

- Méta-données (valeurs attribut)
  - Signification business
  - Type
  - Longueur/précision
  - Valeurs valides (liste/range)
  - Null autorisé
  - Unique
  - Séquence
  - Formatage
  - Conventions d'encodage (null, etc)

Data Profiling
Discovery

- Discovery (*Attribute properties*)
  - Fréquence
  - Patterns
  - Types de données effectifs (date, entiers)
  - Statistiques descriptives (min, max, médiane, moyenne, écart-type, distribution).
  - Nombre de *null*
  - Outliers
- Possibilité de "*drill down*"

Data Profiling
Exemples

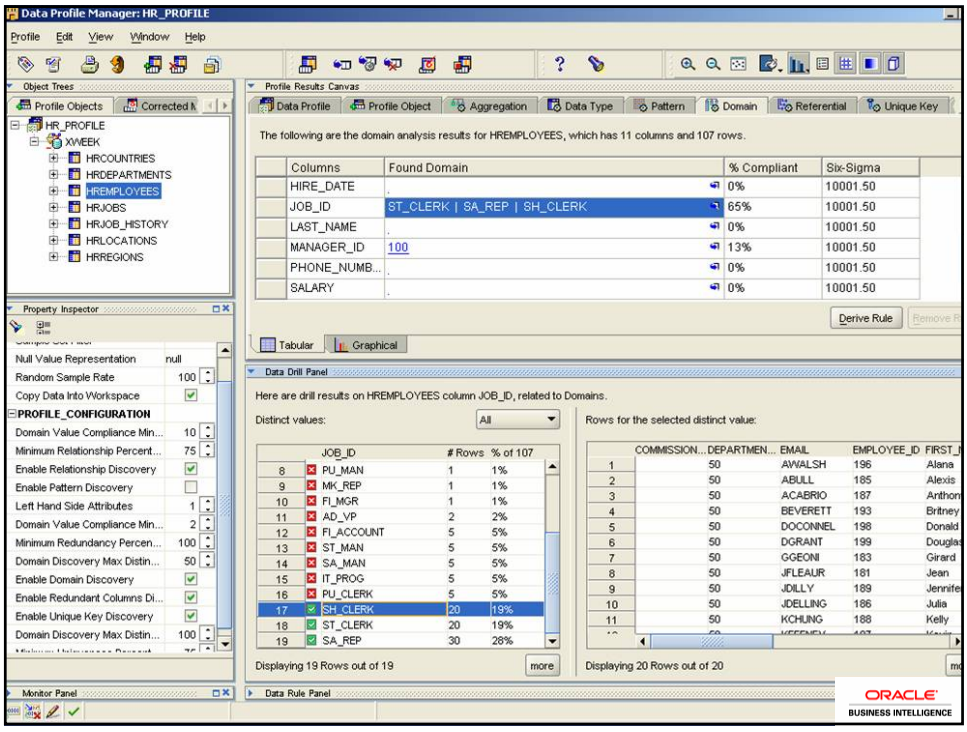
From http://www.dataflux.com

"Loan"

METRIC NAME	METRIC VALUE
Data Type	double
Primary Key Candidate	no
Unique Count	1140
Uniqueness	70.11
Pattern Count	(not applicable)
Minimum Value	-223000
Maximum Value	9999999
Minimum Length	(not applicable)
Maximum Length	(not applicable)
Null Count	2
Blank Count	(not applicable)
Actual Type	double
Count	1628
Data Length	53 bit
Mean	114348.170972
Median	4888499.5
Mode	0
Non-Null Count	1626
Nullable	YES
Ordinal Position	7
Decimal Places	0
Standard Deviation	429438.361236
Standard Error	10649.778281

"Phone number"
Pattern analysis

PATTERN	COUNT	PERCENTAGE
999-999-9999	3166	96.73
(999)999-9999	42	1.28
(999) 999-9999	34	1.04
999 99 9999 999	20	0.61
999 999 9999	5	0.15
999-999-AAAA	2	0.06
9-999-999-9999	2	0.06
a	1	0.03
99 99 9999 999	1	0.03



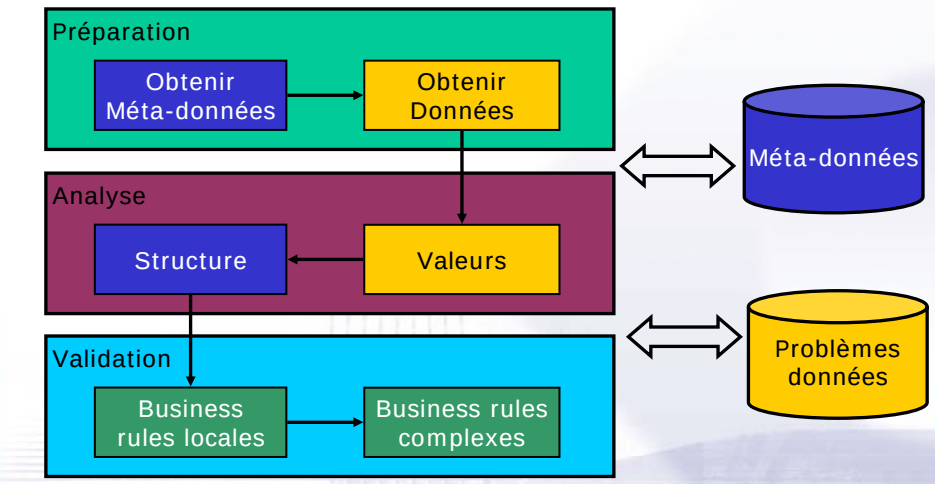
Data Profiling
Indices de problèmes

- Sur base de l'analyse automatique,
  - Attributs non utilisés ou peu utilisés
  - Représentations incohérentes
  - Représentations de NULL (vide, N/A, etc).
  - Valeurs inutilisées
  - Outliers (~ valeur inhabituellement grande ou petite)

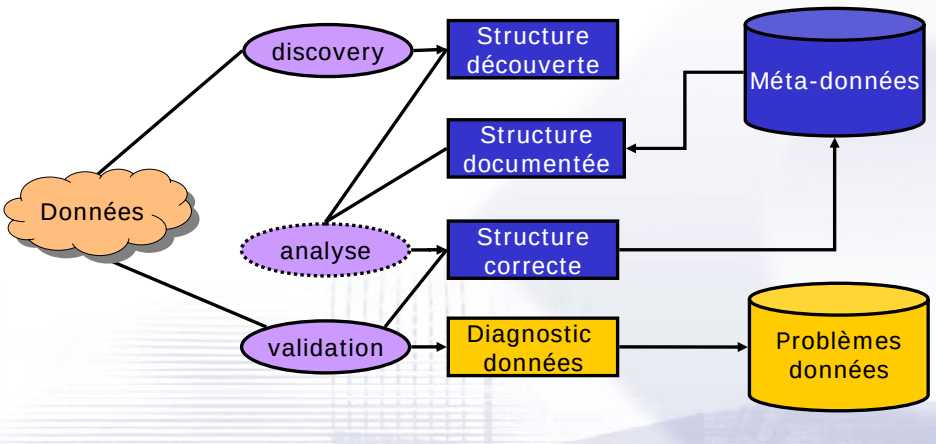
Data Profiling
Intervention humaine

- Inspection visuelle
  - Valeurs extrêmes
  - Caractères spéciaux
  - Random walks
- Signification "business"
- Valeurs autorisées

Data Profiling
Processus



Data Profiling
Structure



Data Profiling
Structurel

- Propriétés:
  - Identifiants
    - ONSS_NR
  - Relations (*Jointures*)
    - FORM_JUR est un code documenté dans la table annexe FORMES_JURIDIQUES.
  - Dépendances fonctionnelles
    - INS_CODE → VILLE
  - Synonymes

Data Profiling
Besoin d'un outil ?

- Méthode ad-hoc (SQL, ...) ?
  - Effort nécessaire (écriture requêtes)
  - Manque de support méthodologique
    - ➔ Risque de manquer des informations importantes
  - Certaines analyses pas possibles:
    - Dépendances fonctionnelles
    - Synonymes
    - Drill-down
- Outils dédiés dépassent ces limitations

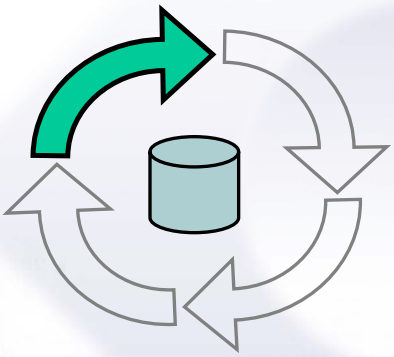
Data Profiling
Reference Book

- **Data Quality: the Accuracy Dimension**, *Jack Olson*

  Elsevier, 2002, The Morgan-Kaufmann Series in Database Management.

Data Profiling
Fin du processus

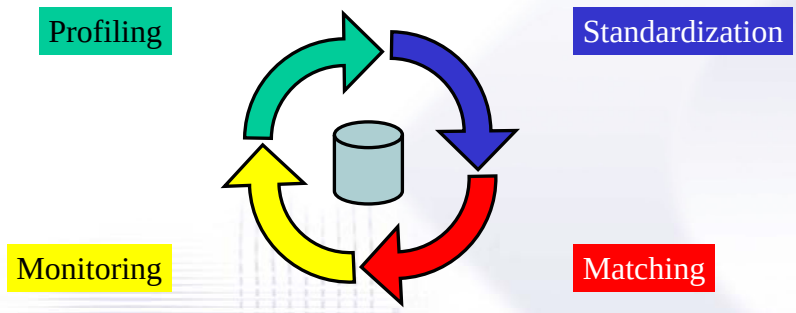
- Repositories:
  - Méta-données
    - Colonnes
    - Structure
    - Règles
  - Contenu
    - Distribution
    - ...
  - Problèmes de données
    - Colonnes
    - Structure
    - Règles



Data Profiling
Outils commerciaux

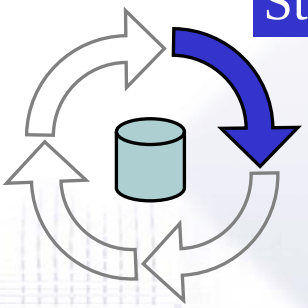
- Points d'attention
  - Support méthodologique et collaboration
  - Ouverture du repository
  - Fonctionnalités de "discovery"
  - Expressivité du moteur de règles
    (lang. contraintes built-in, reg-ex, SQL, javascript, etc)
  - Performance (!) et gestion des jobs
  - Intégration avec phases ultérieures de DQ
  - Possibilités de trends analysis (monitoring)

Cycle d'utilisation



Cycle d'utilisation

Standardization



Standardisation

- Buts
  - Cohérence des conventions de représentation
  - Correction de certains problèmes identifiés par le profiling
  - Enrichissement des données
    - Structuration de champs non-structurés
    - Utilisation de look-up tables

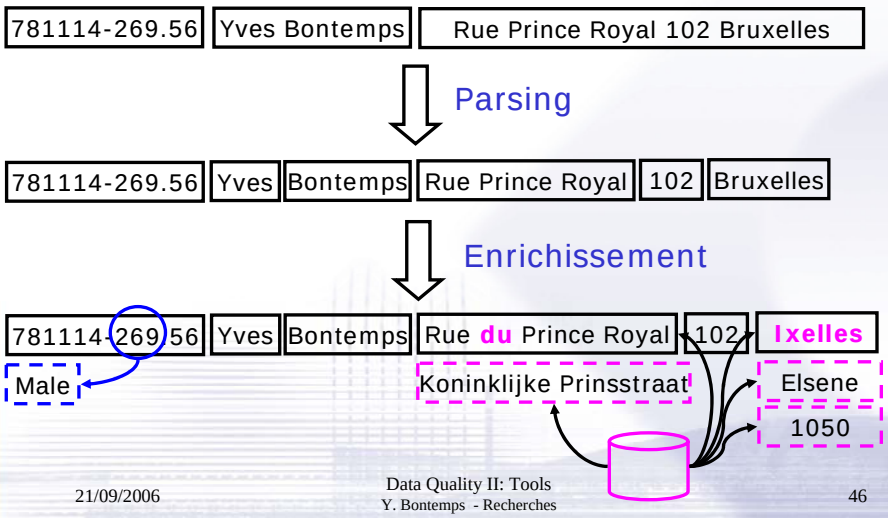
Standardisation

Conventions de représentation

- Domaines restreints
  - Définition d'une transformation
    Dom Invalide → Dom Valide
  - Directement dans outil de Profiling
- Domaines plus complexes (noms, adresses, etc)
  - Parsing + enrichissement
  - Domain-specific information
  - Données de références externes

Standardisation

Domaines Complexes



Standardisation
Domaines Complexes

- Programmation
  - Parsing
  - Enrichissement
- "Encodage" de la connaissance du domaine

```
public class PersonStandardiser {

    protected List<String> decomposeNN(String nationalNumber){
        List<String> decomposition = new ArrayList<String>(5);
        assert(nationalNumber.length() == 14);
        decomposition.add(nationalNumber.substring(0,2));
        decomposition.add(nationalNumber.substring(2,4));
        decomposition.add(nationalNumber.substring(4,6));
        decomposition.add(nationalNumber.substring(7,10));
        decomposition.add(nationalNumber.substring(11,13));
        return decomposition;
    }

    protected void enrich(Person pers, List<String> decomposedNN){
        if (Integer.parseInt(decomposedNN.get(3)) % 2 == 1){
            pers.setGender(Sex.MALE);
        }
        else{
            pers.setGender(Sex.FEMALE);
        }
        pers.setBirthDate(          decomposedNN.get(2)
                                +"/"+ decomposedNN.get(1)
                                +"/"+ decomposedNN.get(0)
                );
    }
```


Standardisation
Parsing

- Patterns, grammaires, expressions régulières

Pattern:	99.99.99-999.99
Décomp:	*AN M J NR CD*
Règle enrich:	*NR* mod 2 = 0 ➔ sexe = 'M'
	NR mod 2 = 1 ➔ sexe = 'F'
	DateNaiss = *J* '/' *M* '/' *AN*

Standardisation
Parsing – Outils commerciaux

- Base de connaissance
  - Corpus de règles
  - Gigantesque (50 000+ règles)
  - Années d'expérience
- Points d'attention:
  - Flexibilité/Exceptions
  - Fréquence des mises à jour
  - Adaptation des règles au contexte particulier (pays, région, langue, culture)

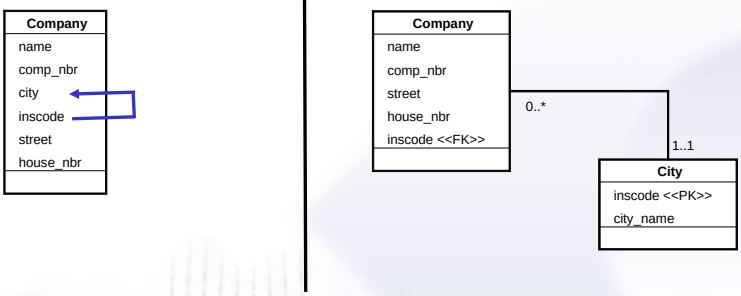
Standardisation
Parsing - Futur

- Techniques d'Intelligence Artificielle
- Modèles probabilistes (Hidden Markov Models)
- Apprentissage → adaptation au contexte.
- Sur le marché dans 2 à 3 ans.

Standardisation
Etapes

- Enrichissement
  - Calculs
  - Look-up tables
    - Enterprise Reference Data (Master Data)
    - Tables "annexes" (ex: codes postaux, etc).
    - Outil commercial (geographique, démographique, etc).
  - Information annexe
    - *"Adresse invalide","N° de maison inexistant" p.ex.*

Standardisation
Structure

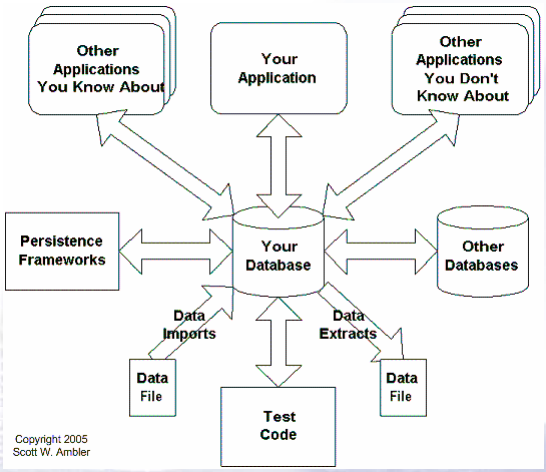


Dépendance fonctionnelle

Normalisation

Standardisation
Structure

- Modification du schéma DB opérationnelle
- Difficile à mettre en œuvre car couplage fort
  ➔ co-évolution
- Pas de support logiciel spécifique



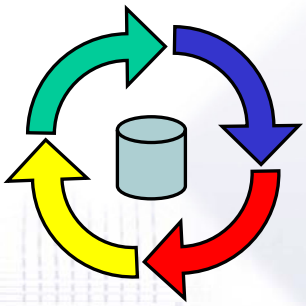
Standardisation
Résumé

- Assurer la cohérence des représentations
- 2 objectifs:
  - Préparation du matching
  - Amélioration de la base de donnée source (en continu)
- Outils commerciaux laissent la structure intacte

Cycle d'utilisation

Profiling

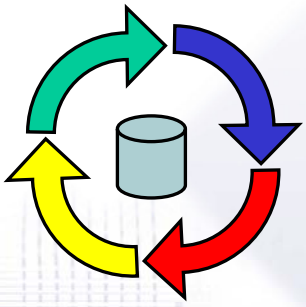
Standardization



Monitoring

Matching

Cycle d'utilisation



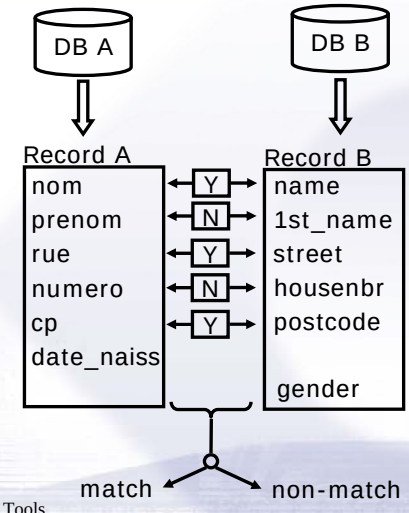
Matching

Matching

- Déduplication
  - Détection des doublons
- Matching
  - Eviter doublons (ex: Data Integration)
  - Relier des BD → déduire de l'information, détecter des incohérences

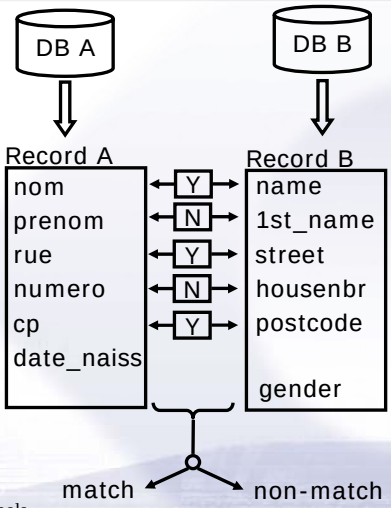
Matching

- Champ par champ
- Par record
- Remarques
  - Performance
  - Outils commerciaux



Matching

- **Champ par champ**
- Par record
- Remarques
  - Performance
  - Outils commerciaux



Matching
Champ par champ

- Chaînes de caractères (noms, noms de rue, …)
- Deux champs sont "les mêmes"
  - AFSCA "=" Agence Fédérale pour la Sécurité de la Chaîne Alimentaire
  - Yves "=" Yves
  - Yves "=" yves
  - Yves "=" Yvse
  - Bontemps "=" Bontan
  - Bontemps, Yves "=" Yves Bontemps

Matching
Familles de comparaisons

	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Matching
Familles de comparaisons

	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching
Comparaisons booléennes

- Relations booléennes
  - is-equal-to
  - is-a-prefix-of
  - is-a-suffix-of
  - is-an-abbreviation-of
  - stems-from
  - ...
- Exemples:
  Bontemps ~ Bontemps
  Bont ~ Bontemps
  Emps ~ Bontemps
  S.A. ~ Soc. Anon.
  Bakkerij ~ Bakker

Matching
Rule-based

- Règles codées à la main
  - Hard-wired (Java, C, COBOL, ...)
  - Rule engine (declarative rules: prédéfinie et user-defined)
- Peuvent être fournies dans un outil commercial (black-box ?)
- ☺ Meilleure utilisation de la connaissance du domaine
- ☹ Fastidieux
- ☹ Coûteux (à créer/maintenir)

Matching
Familles de comparaisons

Booléennes	**Prédicats et règles** • Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	**Phonétiques** • Soundex • Metaphone *Bontant ~ Bontemps*
Similarité	**Word-based** • Edit distances • Lettres communes *Bontemps ~ Bnotmps*	**Token-based** • Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching
Phonétique

- Erreurs de retranscription de l'oral vers l'écrit.
- Ex: Bontemps:
  - Montand
  - Bontant
  - Bonton
  - Bontand
  - Bautemps
  - ...

Data Matching
Phonétique

- Principe:
  - Mot *m* → représentation de sa prononciation *p*(*m*).
  - $p(m) = p(m') \Leftrightarrow m \sim m'$
- cfr indexation par hashing

Data Matching
Phonétique

- Russel Soundex Algorithm (1918)
  - Garder première lettre
  - Supprimer a,e,h,i,o,u,w,y
  - Codage:
    - 1: B,F,P,V
    - 2: C,G,J,K,Q,S,X
    - 3: D,T
    - 4: L
    - 5: M, N
    - 6: R
  - Garder 4 premiers symboles (padding avec 0)
- Exemple
  - Bontemps
    BNTMPS
    B53512
    B535
  - Bondant
    BNDNT
    B5353
    B535

Data Matching
Phonétique

- Variantes:
  - Metaphone et Double Metaphone
  - NYSIIS et NameSearch®
  - Daitch-Mokotoff
    - Langues Slaves et Germaniques
    - 54 entrées
  - Fonem
    - Français
    - 64 rules pour mettre les règles en forme normale
  - Phonex, etc…

Matching
Familles de comparaisons

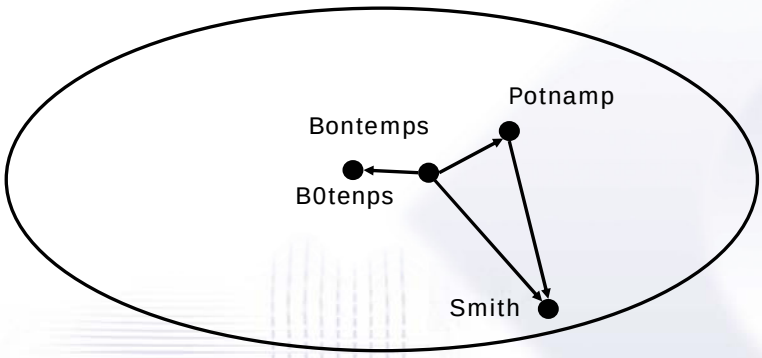
	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

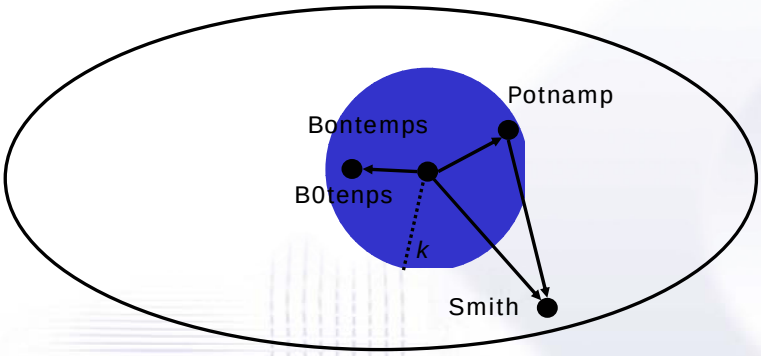
Data Matching
Similarité

- Remplacer l'approche booléenne (0-1) par une approche plus fine
- Idée:
  - Similarité variable entre
    - Bontemps,
    - Smith,
    - Potnamp,
    - B0tenps.

Data Matching
Similarité



Data Matching
Similarité



Matching
Familles de comparaisons

	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Matching
Edit distance

- Distance entre deux mots = nombre de "fautes" à effectuer pour passer de l'un à l'autre
  - Insertion (I)
  - Effacement (D)
  - Substitution (S)
- Coût par faute
  - OCR, Clavier, etc.

- Exemple
  - Bontemps
  - Bnontemps (I)
  - Bnontemps (D)
  - Bnontemps (D)

➔ 3 opérations

Data Matching
Lettres communes

- Jaro:
  - Fenêtre de (m/2) – 1
- Exemple
  - 3 caractères communs
  - 1 transposition
  - d = 1/3 * 3/7 + 1/3 * 3/7 + 1/3 * (3 – 1)/3 = 0.508

	S	L	U	I	T	E	N
S	1						
T					1		
R							
U			1				
D							
E						1	
L							

Matching
Familles de comparaisons

	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching
Token-based

- Plusieurs éléments dans un record
- Exemples
  - Yves Bontemps
  - Organisme national des déchets radioactifs et des matières fissiles.

Data Matching
Token-based metrics

- Ignorer l'ordre des mots
- Métrique de Jaccard

$$\frac{|S \cap T|}{|S \cup T|}$$

- Bontemps, Yves =?= Yves Bontemps
- {{Bontemps;Yves}} {{Yves; Bontemps}}
- 2/2 = 1

Data Matching
Token-based

- Term Frequency / Inverse Document Frequency (TF-IDF)
- Rareté d'un mot ➔ importance.
- Exemples:
  - Organisme National des Déchets Radioactifs et des Matières Fissiles
  - Office des Matières Fissiles et Déchets Radioactifs

Data Matching
Token-Based

- Approches récursives:
  - Utiliser une technique "word-based" (Levenshtein, etc) pour déterminer si 2 tokens sont les mêmes (*Office ~ Ofice*)
  - Utiliser une technique "token-based" pour combiner ces résultats
  - Réf: Soft TF-IDF, etc.

Matching
Familles de comparaisons

	Prédicats et règles	Phonétiques
Booléennes	• Egale, • Suffixe de • Stems from *Boulanger ~ Boulangerie* *S.A. ~ Société Anonyme*	• Soundex • Metaphone *Bontant ~ Bontemps*

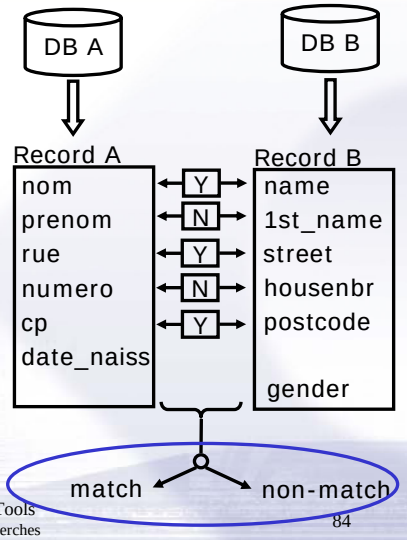
	Word-based	Token-based
Similarité	• Edit distances • Lettres communes *Bontemps ~ Bnotmps*	• Cosinus • Recursives *Office National des matières fissiles ~ Office des matières fissiles*

Data Matching
En pratique

- Comparaison phonétique (KBO):
  - Décomposition en mots
  - Mots mis sous forme phonétique.
    - Variante de Soundex
    - Ignore termes habituels (S.A., ...)
    - Stemming
  - Comparaison
    (prise en compte de l'ordre des mots)

Matching

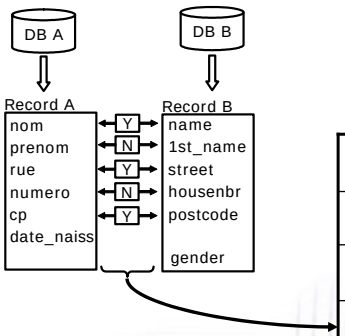
- Champ par champ
- **Par record**
- Performance



Data Matching
Par record

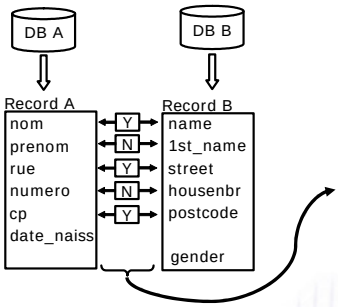
- Approches:
  - Déterministe
  - Probabiliste (*global scoring*)

Data Matching
Approche déterministe



	Nom	Prénom	Rue	Num	Cp	**Result**
1	Y	Y	-	-	Y	**Match**
2	N	-	-	-	-	**No**
3	Y	N	Y	Y	Y	**Match**
4	Y	N	Y	N	Y	**Maybe**
4	-	-	-	-	-	**No**

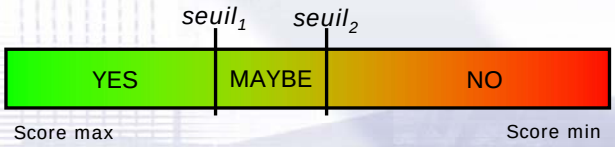
Data Matching
Global Scoring



- *Poids*
  - Nom = 0.95
  - Prenom = 0.45
  - Rue = 0.25
  - Num = 0.01
  - CP = 0.05

➔ Score = 1.25 (p.ex.)

- Règle:

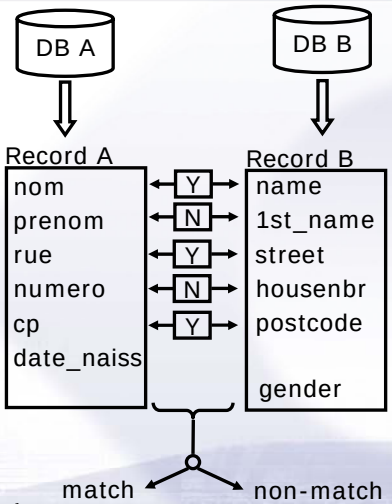


Data Matching
Probabiliste (Fellegi & Sunter, 1969)

- Pour chaque comparaison, deux probabilités:
  - m = Prob. de Y si A et B sont un vrai match
  - u = Prob. de Y si A et B sont un vrai unmatch.
- Le score est calculé sur base de ces probabilités (m/u)
- Classification en fonction du score
  - càd, score élevé si Y *et* N très discriminants

Matching

- Champ par champ
- Par record
- **Performance**



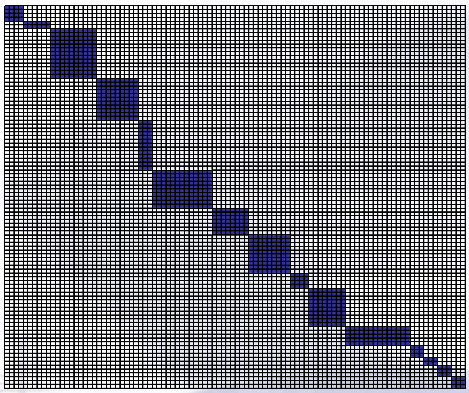
Data Matching
Performances

- Matching = comparer tous les records de A avec tous les records de B
- Exemple:
  - A : 100 000 rec.
  - B : 100 000 rec.
  - ➔ 10 000 000 000 de comparaisons.

➔Optimisations nécessaires...

Data Matching
Performances

- Blocking:
  - Donnée critique → former des blocs
  - Les records sont comparés bloc par bloc.
  - Ex: Code Postal, …
  - Trade-off précision et performance
  - Nécessite indexation/tri.
  - Multi-passes possible
- Sliding window, etc



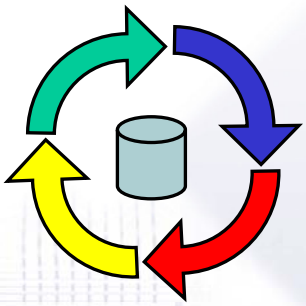
Data Matching
Outils commerciaux

- Comparaison champ-par-champ
  - Best-of-breed, best-of-market (black-box ?)
- Agrégation
  - Global scoring (dans la plupart des cas)
  - Déterministe
- Performances
  - Blocking inclus (obligatoire)
  - DBMS extérieur ou gestion propre.
- Mode online et mode batch.

Cycle d'utilisation

Profiling

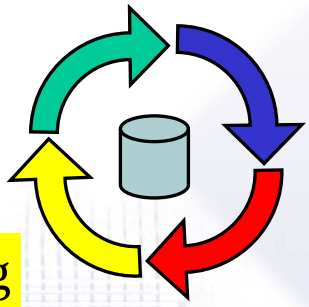
Standardization



Monitoring

Matching

Cycle d'utilisation



Monitoring

Monitoring

- Surveillance de l'évolution de la qualité
  - Régulière
  - Planifiée
  - Automatisée
- Tableaux de bord et alarmes
- Trends analysis

Monitoring

- Profiling tools
  - Violation des règles business (méta-données prescriptives)
  - Evolution des fréquences de valeur
    - Ex: +150% de N/A, etc.
- Possibilité de stocker des profils d'analyse et de les répéter au cours du temps

Monitoring

- Aspect nécessaire,
- Data Quality = Processus continu

- Pertinence des outils
  - Intégration
    - Architecture IT (impact sur performances).
    - Decision Support Systems
  - Fonctionnalités proposées (faibles)

- Introduction
- Outils de Data Quality: concepts
  - Data Profiling
  - Standardisation
  - Matching
  - Monitoring
- **Outils commerciaux**
  - Case study
  - Architecture

- Conclusion

Data Quality
Marché

- Fournisseurs principaux
- Case studies/Proof of concept
- Offre logicielle

Fournisseurs principaux

- Avril 2006: Magic Quadrant Gartner → marché reconnu.
- Deux phases de consolidation
  - Profiling → DQ
  - DQ → BI/ETL
- Futur: DQ = commodity?

Outils commerciaux
Offre

- Data Profiling
  - Discovery
  - Assertive testing
  - Collaboration (via email)
  - Langage de règles variable
- (Monitoring)

- Standardisation et matching
  - Base de connaissance (règles, adresses, …)
  - Performance
  - Templates de projet
  - Connecteurs
  - Updates

Case studies

- Répertoire des employeurs ONSS
- Détection de doublons
  - Insertion des doubles immatriculations
- 230 000 entrées et 10 colonnes analysées, dont
  - Dénomination,
  - Adresse
  - Forme juridique, …

Case studies

- "Workshops" d'une journée
- Planning
  - Profiling (2h)
  - Matching (4h)
  - Debriefing (1h30)

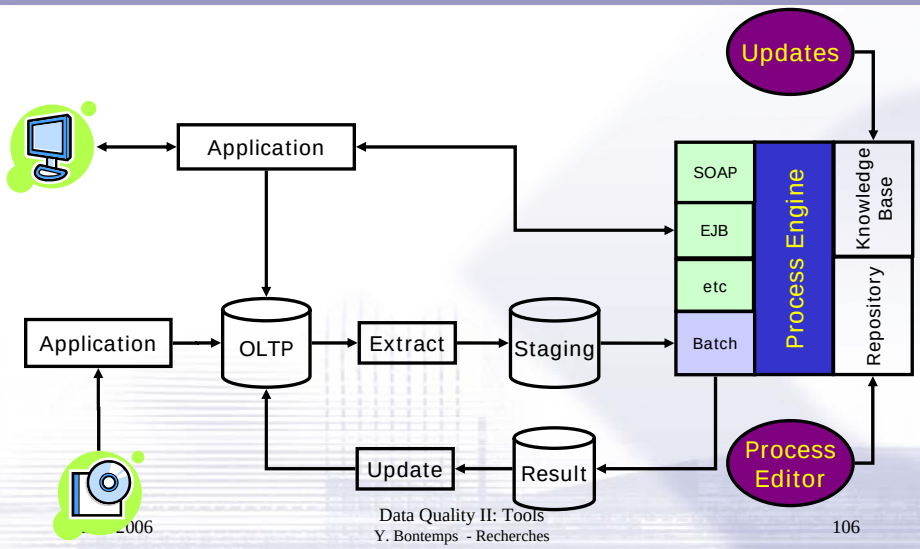
Case studies
Enseignements

- Performances
  - Rapidité de mise en œuvre (situation simple)
  - Performances (6 min. à 30 min. machine low-end)
  - Facilité d'utilisation (env. graphique, pas de programmation)
- Standardisation
  - Adresses bilingues: support variable.
  - Dénomination: pers. physiques vs pers. morales
- Matching
  - Champ par champ: black-box (pas un obstacle)
  - Record par record: Déterministe → traçabilité
  - Qualité des résultats

Outils commerciaux

- Ordres de prix (licences)
  - Profiling:
    - 25 000 € (1 named user)
    - 10 000 € par named user suppl.
    - Peu de users dans les entreprises
  - Standardisation, Matching:
    - 100 000 €

Architecture



- Introduction
- Outils de Data Quality: concepts
  - Data Profiling
  - Standardisation
  - Matching
  - Monitoring
- Outils commerciaux
  - Case study
  - Architecture
- **Conclusion**

Conclusion

- Data Quality Tools ?
  - Fonctionnalités ?
    - Profiling
    - Standardisation/Enrichissement
    - Matching/Deduplication
    - Monitoring
  - Apport ?
    - Economie de l'expérience !
    - Performances

Conclusion
Scénarios

- Profiling
  - Au début d'un projet, extraction, documentation et vérification des données
  - Quelques jours (analyse)
  - Validation de toutes les hypothèses !
  - Diminution risques projet

Conclusion
Scénarios d'utilisation

- Standardisation/Déduplication batch
  - Création d'entités via Web ou batch
  - Déduplication en batch
  - Répétitif (mensuel, hebdomadaire, …)
  - Traitement:
    - Surviving record
    - Gestion manuelle
    - Linkage des records (trace/info supplémentaire)

Conclusion
Scénarios d'utilisation

- Validation on-line
  - Lors de l'encodage,
    - Validation
    - Standardisation
    - Déduplication
  - Feed-back vers l'utilisateur
  - Impact sur l'application "front-office"

Conclusion
Facettes d'une solution

Management

- Exec. DQ group
- Data Mgmt
- Data stewards
- ...

Information

- Meta-data
- Data Models
- Requirements
- Metrics, ...

Data as an Asset

Activités

- Root-Cause Analysis
- Data Integration
- Software Eng.
- Business Proc. Eng.
- Cleansing, etc.

Bénéfices

- Utilisateurs
- Développeurs
- Clients
- ...

Conclusion

Facettes d'une solution

Management

- Exec. DQ group
- Data Mgmt
- Data stewards
- ...

Tools

Information

- **Meta-data**
- Data Models
- Requirements
- *Metrics*, ...

Data as an Asset

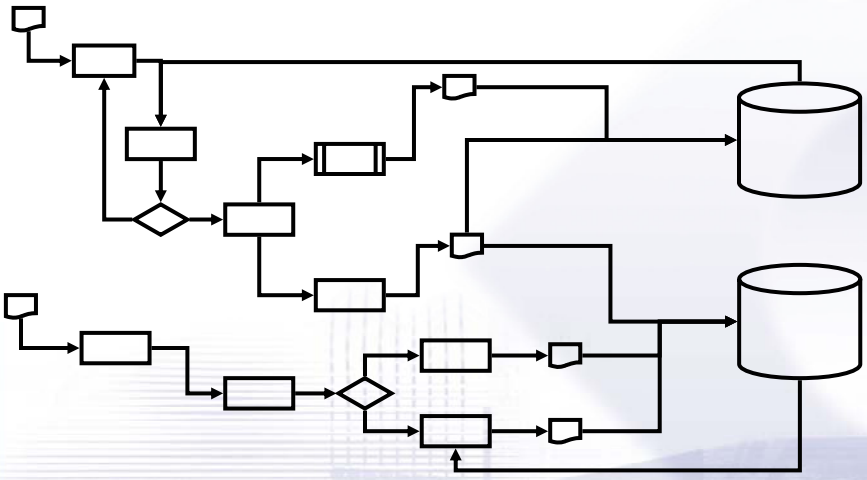
Activités

- Root-Cause Analysis
- **Data Integration**
- **Software Eng.**
- Business Proc. Eng.
- **Cleansing**, etc.

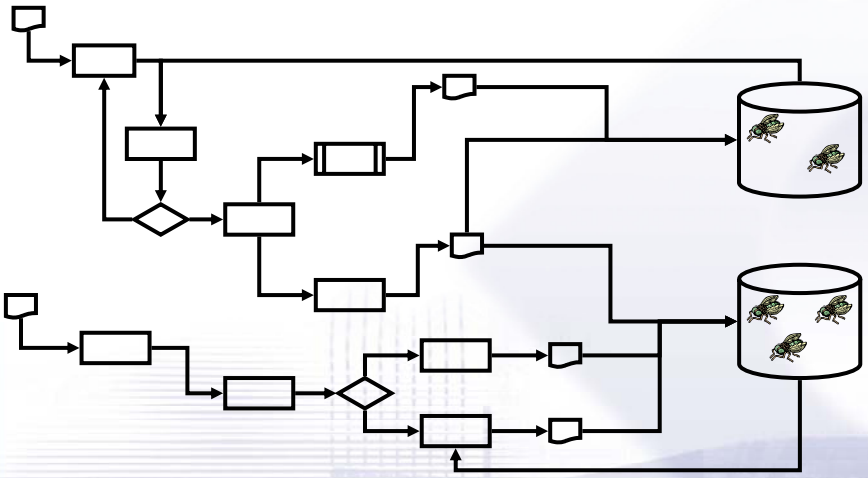
Bénéfices

- Utilisateurs
- Développeurs
- Clients
- Affaires
- ...

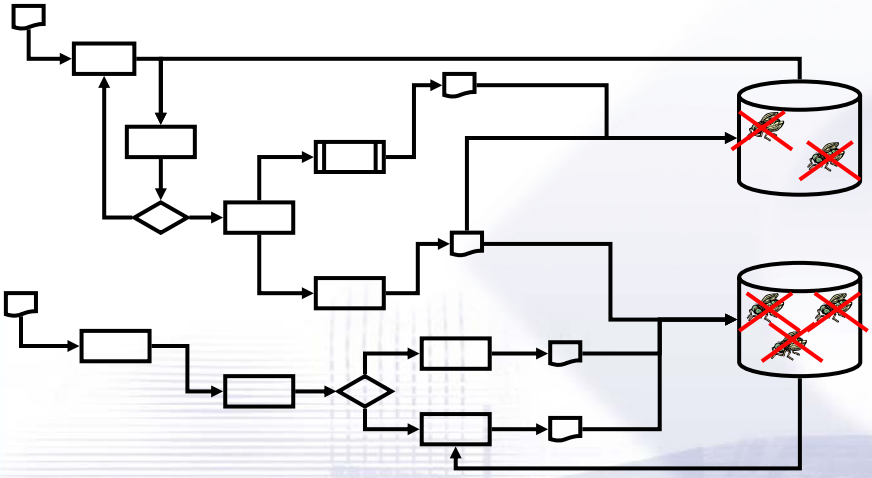
Data Quality



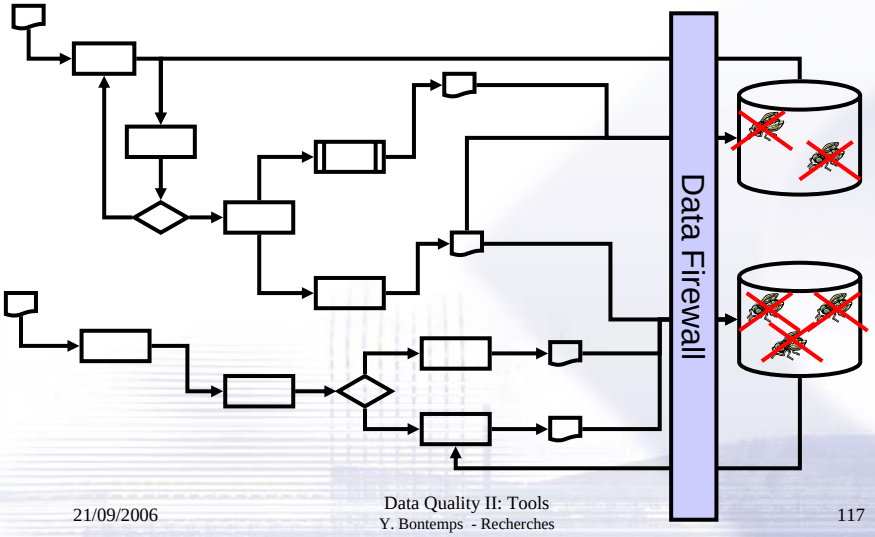
Data Quality
Profiling



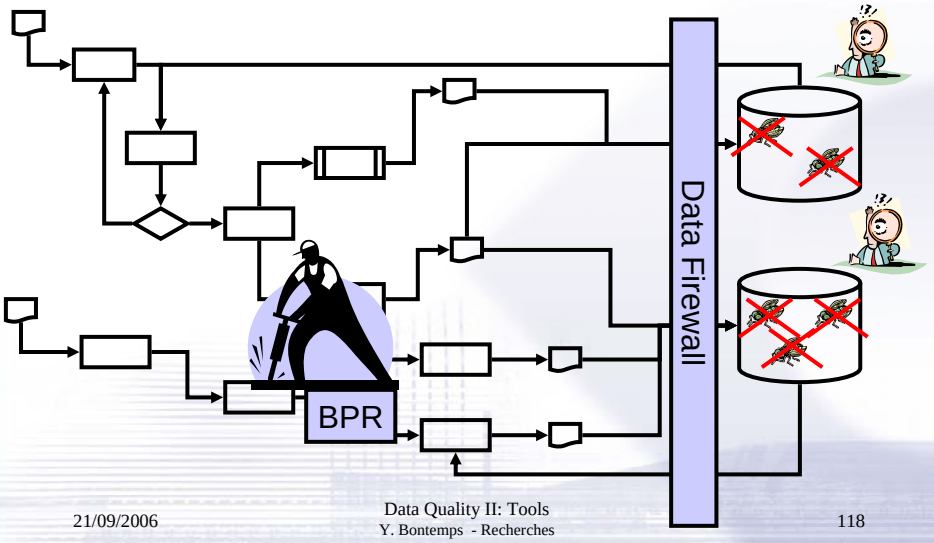
Data Quality
Standardisation & Matching



Data Quality
Infrastructure



Data Quality



"Data Quality @ SmalS-MvM"

Cellule "data quality" (section "recherches")

- En collaboration avec les autres équipes de la société:
  - Sensibilisation à la qualité des données,
  - Formations,
  - Mise en place d'indicateurs,
  - Mise en place de groupes de travail & de suivi,
  - Actions spécifiques (root-cause analysis, etc),
  - Analyses de l'existant (impact, …),
  - Aide à la mise en place d'outils,
- Etudes et publications de travaux
- Consultances au sein de l'administration fédérale belge