Data quality (part I) : Best Practices

Isabelle Boydens
SmalS-MvM, Section Recherches
21 Mars 2006

Plan de l'exposé

- Position du problème et enjeux
- Analyse : dimensions de la qualité des données
- Méthodes d'amélioration de la qualité
- Conclusions

Position du problème et enjeux

- **Définitions**
- Symptômes de la "non qualité"
- Coûts de la "non qualité"
- Causes de la "non qualité"

Définitions : plan

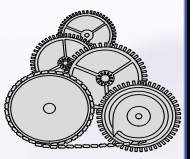
- **Les origines du concept de qualité**
- La qualité des bases de données
- Les bases de données administratives : caractéristiques

Les origines du concept de qualité

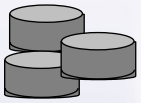
- *ποιος*, *qualis*, “quel ?”, "welk ?"
- "qualité" versus "quantité"
- degré plus ou moins élevé d'une échelle de valeurs pratiques

Les origines du concept de qualité

- Normes en matière de production industrielle (taylorisme, années 20)



- Apports :
  - Concept de "one best" :
    - La perfection est une "non valeur"
    - Arbitrage "coût-bénéfice"
    - La "sur-qualité" est de la "non-qualité"



  - Valorisation de la "qualité" au niveau du management

Les origines du concept de qualité

"Il est préférable de livrer en retard un produit qui fonctionne plutôt que de livrer à temps un produit qui ne fonctionne pas…"

Différence entre le "non fonctionnement" :

- D'un produit matériel (voiture en panne)
- D'une information ("non pertinence" des données en fonction des usages…)

Les origines du concept de qualité

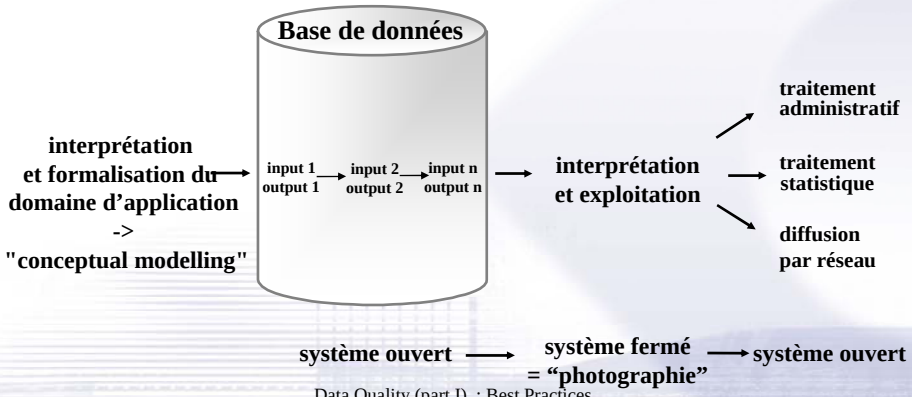
- Standards internationaux (ISO 9000, "total quality management", …) et certifications MAIS :
  - Beaucoup de généralités
  - Lourdeur et coût de mise en œuvre
  - Ponctualité de la certification : parfois, fin en soi (or, démarche continue indispensable)
  - Biais liés aux enjeux commerciaux des certifications
  - Distinction entre production industrielle et production d'information
- Essai d'application du suivi de la production aux bases de données (cfr "data tracking")

Définitions : plan

- Les origines du concept de qualité
- **La qualité des bases de données**
- Les bases de données administratives : caractéristiques

Bases de données et système d'information

Base de données : abstraction, "simplification" du réel observable s'inscrivant dans un système d'information



La qualité des bases de données

- Qualité d'une base de données : adéquation d'une base de données à ses objectifs ("fitness for use")
- Arbitrage coût/bénéfice : pas de "qualité totale"
- Enjeux stratégiques lorsque l'information est un instrument d'action sur le réel
- Approche pluridisciplinaire (techniciens, concepteurs, experts du domaine, …)
- Varie avec les caractéristiques du domaine d'application

Définitions : plan

- Les origines du concept de qualité
- La qualité des bases de données
- **Les bases de données administratives : caractéristiques**

Les bases de données administratives : caractéristiques

- **L'administration : définition et fonctions**
- Caractéristiques générales des bases de données administratives
- Deux types de systèmes d'information :
  - bases de données reposant sur un mode déclaratif régulier
  - Répertoires, référentiels ou sources authentiques

L'administration : définition et fonctions

- L'administration est constitutive de l'appareil d'Etat :
  - Prélèvement de contributions auprès des citoyens pour le fonctionnement de l'Etat
  - Exécution de services au profit des administrés
  - Production des règlements destinés à adapter la loi aux exigences de la pratique quotidienne

Les bases de données administratives : caractéristiques

- L'administration : définition et fonctions
- **Caractéristiques générales des bases de données administratives**
- Deux types de systèmes d'information :
  - bases de données reposant sur un mode déclaratif régulier
  - Répertoires, référentiels ou sources authentiques

Caractéristiques générales des bases de données administratives

- Souvent considérées, à tort, comme "simples" !
- Modifications législatives fréquentes et complexes → gestion des versions et historique
- Force probante des données
- "Idéalement", pas de tolérance à l'erreur (traitement équitable des dossiers des citoyens)
- Volume de données et d'anomalies important
- Incidences sociales et financières considérables

Les bases de données administratives : caractéristiques

- L'administration : définition et fonctions
- Caractéristiques générales des bases de données administratives
- **Deux types de systèmes d'information** :
  - bases de données reposant sur un mode déclaratif régulier
  - Répertoires, référentiels ou sources authentiques

Bases de données reposant sur un mode déclaratif régulier (DMFA)

- Objectif déclaratif et prélèvement régulier de l'information :
  - l'information est régulièrement mise à jour
  - contacts réguliers avec la population "cible"
- Modifications de schémas fréquentes et complexes
- Quelques chiffres (ordres de grandeur) :
  - enregistrements saisis chaque trimestre : 4.000.000
  - anomalies formelles : plusieurs centaines de milliers par trimestre
  - montants en jeu : 37 milliards d'euros

Répertoires (KBO), référentiels ou sources authentiques

- Contacts irréguliers avec la population "cible" :
  - communication ponctuelle d'événements : fusion d'entreprises, changement d'activité principale, d'adresse ...
  - information potentiellement plus obsolète (coût !)
- Pompe "aspirante-refoulante" (alimentation initiale : compromis entre besoins et sources disponibles)
- Peu de champs (l'exhaustivité des enregistrements prime sur la précision du schéma)
- Schéma plus stable

Position du problème et enjeux

- Définitions
- **Symptômes de la "non qualité"**
- Coûts de la "non qualité"
- Causes de la "non qualité"

Symptômes de la "non qualité"

- Plaintes des clients et préjudices (pertes financières, perte en crédibilité, procès, …)
- Ampleur des procédures de contrôle et de correction de l'information (concept "d'usine fantôme")
- Ampleur de la redondance dans les fichiers et les traitements
- Difficultés d'interprétation

Problèmes de redondance et d'interprétation

"A man with a watch knows what time it is.

A man with two is never sure."

M. Twain

Position du problème et enjeux

- Définitions
- Symptômes de la "non qualité"
- **Coûts de la "non qualité"**
- Causes de la "non qualité"

Les coûts de la "non-qualité"

- Vérification et correction de l'information
- Traitement des plaintes et procès
- Réparation des préjudices éventuels
- Difficultés lors de l'intégration de nouvelles technologies
- Crédibilité
- Erreurs de stratégie

Les coûts de la "non qualité"

- Selon une enquête aux USA (Redman, 1999) :
  - Taux d'erreur moyen dans les bases de données : 5 à 30 %
  - Dans les enregistrements médicaux (hôpitaux) : jusqu'à 80% d'erreurs formelles !
- Coûts moyens (Redman, 1999) :
  - 15% du revenu des entreprises
  - 50% des coûts de la conception d'un "datawarehouse"
- 59,5 milliards de $ de perte annuelle nationale aux USA (étude de 2002, citée dans Cinquin, 2006)

Position du problème et enjeux

- Définitions
- Symptômes de la "non qualité"
- Coûts de la "non qualité"
- **Causes de la "non qualité"**

Les "causes" de la "non qualité"

<u>Un système d'information est un fleuve</u> :

la mise en oeuvre exclusive de tests d'intégrité permet de nettoyer ponctuellement le fond du fleuve mais n'endigue pas l'arrivée de nouveaux flux d'information de qualité douteuse.

(T. Redman)

Les "causes" de la "non qualité"

- Vision à "court terme"
- Importance insuffisante accordée :
  - Aux usages ("*use it or lose it*"), au contexte de l'information ("périmètre")
  - A la documentation des données et des processus
- Séparation excessive entre la phase de conception d'une base de données et le suivi de sa qualité
- Concentration sur les nouvelles technologies et négligence des questions que posent les applications de gestion courante

Data quality : best practices

- Position du problème et enjeux
- **Analyse : dimensions de la qualité des données**
- Méthodes d'amélioration de la qualité
- Conclusions

Analyse : les dimensions de la qualité des données

- **Introduction**
- Qu'est-ce qu'une donnée ?
- Qu'est-ce qu'une donnée correcte ?
- Comment les données se construisent-elles progressivement ?
- Indicateurs de qualité

Les dimensions de la qualité des données : introduction

- Pas de qualité sans système d'évaluation homogène :
  - Permettre des comparaisons dans le temps et de l'espace
  - Suivre l'impact des décisions, les progrès éventuels, …
  - Éviter les dérives ("*data quality act*")
- Quels indicateurs d'évaluation choisir ?
  S'interroger sur l'objet : données administratives

Analyse : les dimensions de la qualité des données

- Introduction
- **Qu'est-ce qu'une donnée ?**
- Qu'est-ce qu'une donnée correcte ?
- Comment les données se construisent-elles progressivement ?
- Indicateurs de qualité

Qu'est-ce qu'une donnée ?

- 3 composantes (triplet) :
  - un intitulé/concept : ex. salaire mensuel
  - un domaine de définition : ex. valeur numérique incluse entre 1.000 € et 100.000 €
  - une valeur : ex. 3.000 €

Identifiant	nom	prénom	salaire	catégorie	Taux cotisation	Année, mois	date – update
lkm-pod	Durant	Jean	3.000 €	chimie	0.23 %	jan 1998	25/5/1998

Qu'est-ce qu'une donnée ?

- Quelques propriétés :
  - Interactions entre composantes
  - Données déterministes vs données empiriques :
    - Les bases de données répertorient essentiellement des données empiriques ("concepts mobiles")
  - "Closed World Assumption"

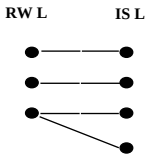
Identifiant	nom	prénom	salaire	catégorie	Taux cotisation	Année, mois	date – update
lkm-pod	Durant	Jean	3.000 €	chimie	0.23 %	jan 1998	25/5/1998

Analyse : les dimensions de la qualité des données

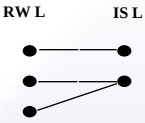
- Introduction
- Qu'est-ce qu'une donnée ?
- **Qu'est-ce qu'une donnée correcte ?**
- Comment les données se construisent-elles progressivement ?
- Indicateurs de qualité

Qu'est-ce qu'une donnée correcte ?

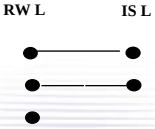
représentation correcte



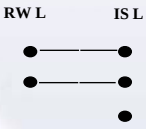
représentation ambiguë



représentation incomplète



état non significatif



Légende (programme TDQM – MIT) :
RWL : ***"lawful state space of a real-world system"***
ISL : ***"lawful state space of an information system representing the real world system"***

Qu'est-ce qu'une donnée correcte ?

Comment déceler une incohérence entre
une donnée A (catégorie) et une donnée B (taux-cotisation) ?
Et comment identifier avec certitude l'information "correcte"?

Employeur

Identifiant	nom	prénom	catégorie	taux-cotisation
km-pod	Durant	Jean	banques de données	0,27 %

Catégorie_taux

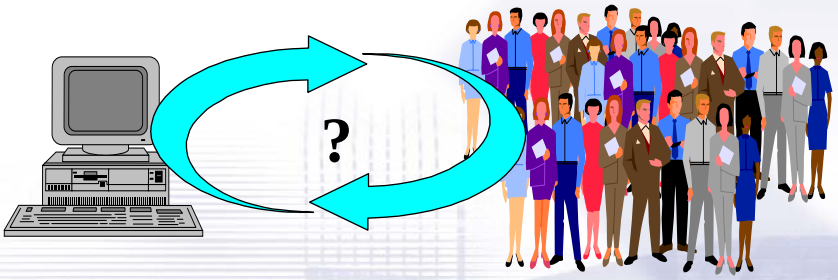
Catégorie	taux-cotisation
conseil informatique	0,28%
traitement de données	0,27%
banques de données	0,29%

Qu'est-ce qu'une donnée correcte ?

- Typologie des violations de contraintes d'intégrité :
  - Erreur formelle
  - Présomption formelle d'erreur (anomalie)
    - A priori
    - A posteriori
  - Erreur indétectable formellement
- La catégorie "autres"

Les données ne sont pas données

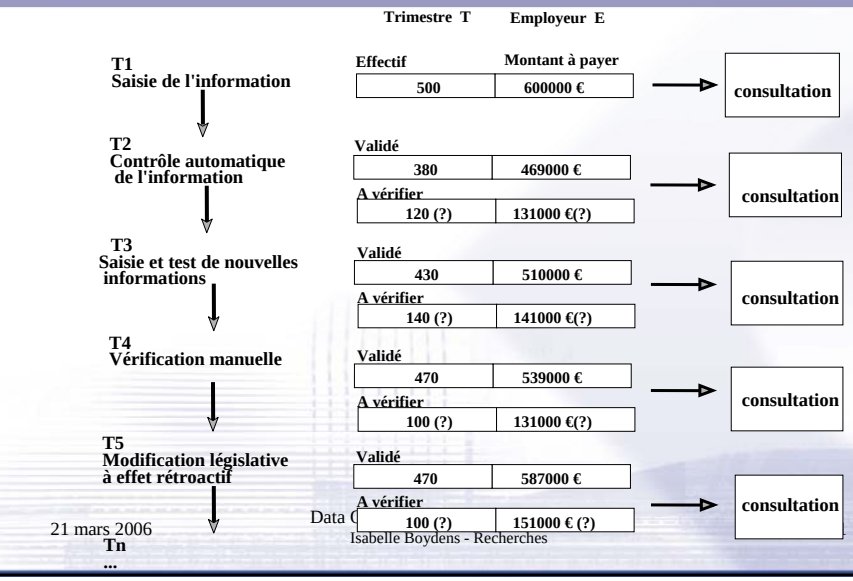
On ne dispose d'aucun référentiel "absolu" en vue de tester la correction d'une vaste base de données empiriques



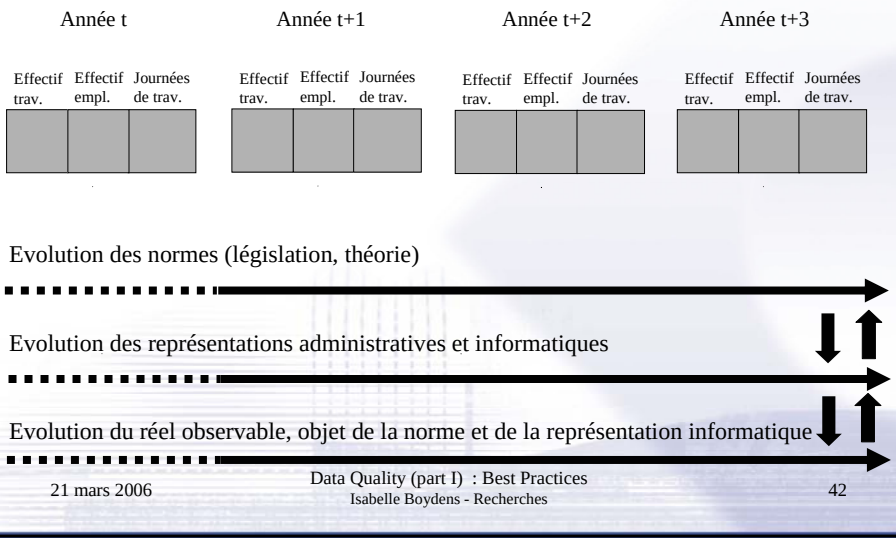
Analyse : les dimensions de la qualité des données

- Introduction
- Qu'est-ce qu'une donnée ?
- Qu'est-ce qu'une donnée correcte ?
- **Comment les données se construisent-elles progressivement ?**
- Indicateurs de qualité

Comment l'information se construit-elle progressivement ?



Comment l'information se construit-elle progressivement ?



Analyse : les dimensions de la qualité des données

- Introduction
- Qu'est-ce qu'une donnée ?
- Qu'est-ce qu'une donnée correcte ?
- Comment les données se construisent-elles progressivement ?
- **Indicateurs de qualité**

Indicateurs de qualité

- La "correction" n'est pas un indicateur valable
- Les indicateurs de qualité sont nécessairement "latéraux"; certains sont quantifiables, d'autres pas
- Indicateur principal : pertinence des concepts et des processus (non quantifiable)
  - interaction entre besoins et sources disponibles
  - arbitrages de type coûts bénéfices

  → "Master Data Management"

Indicateurs de qualité

- Autres indicateurs potentiellement importants :
  - Précision (schéma)
  - Usability (schéma)
  - Comparabilité, accessibilité, clarté (schéma)
  - Fraîcheur (extension)
  - Validité formelle des valeurs (extension)
  - Processus de traitement des anomalies (flux)
  - Ponctualité par rapport aux besoins (flux)
- Arbitrages entre indicateurs concurrents
  - Rapidité vs validité formelle vs coût
  - Exhaustivité vs précision

Indicateurs de qualité : stratégie de mise en oeuvre

- Démarche descendante :
  - Cibler les besoins sur la base des objectifs (éviter une multiplicité de chiffres)
  - Aller des concepts au calcul opérationnel
  - Définir plusieurs niveaux d'agrégation
  - Travail de synthèse, de clarification et d'interprétation (méta-informations)
  - Industrialiser la production (méthode, organisation et suivi continu)
  - Définir des stratégies d'amélioration

Source : P. Rivière, INSEE, 2005

Indicateurs de qualité : exemple (BCE)

- Sujets d'intérêt principaux :
  - L'identifiant
  - Les variables
- Principe d'évaluation :
  - Exemple : "faux actifs" : taux d'unités non présentes à l'adresse indiquée
- Méthode d'évaluation opératoire
  - Champ temporel et spatial
  - Variable d'intérêt (exemple :identifiant)
  - Domaine-cible : sous-populations concernées
  - Mode de calcul ou d'observation

Source : P. Rivière, INSEE, 2005

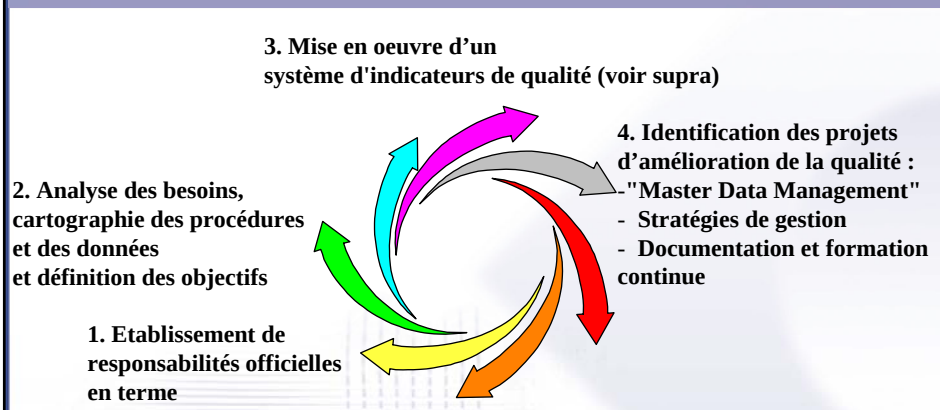
Indicateurs de qualité : exemples de méthodes d'observation

- Enquête sur la base d'échantillons :
  - "one shot"
  - cher si récurrent (traitement des "non réponses")
  - crédibilité vis-à-vis de clients contactés plusieurs fois si on respecte le principe de l'échantillonnage (problème de la base de sondage)
- Analyse de la cohérence interne (tools)
  - Au niveau des données (exemple : chiffre d'affaire/effectif)
  - Au niveau temporel
- Comparaison avec une source concurrente (tools)

Data quality : best practices

- Position du problème et enjeux
- Analyse : dimensions de la qualité des données
- **Méthodes d'amélioration de la qualité**
- Conclusions

Méthodes d'amélioration : cycle itératif



Méthodes d'amélioration : points clés

- Appui et suivi du management (cycle)
- Mise en place de rôles ("data quality stewardship") et d'un comité de suivi (groupes de travail pluridisciplinaires incluant les utilisateurs)
- Mise en œuvre de procédures dont les gains seront mesurables et continus : éviter les mesures ponctuelles prises dans l'urgence, les opérations "coup de poing"…

Méthodes d'amélioration : plan

- **Examen et amélioration de l'architecture de base : "Master Data Management"**
- Production d'informations en vue du déploiement ultérieur de stratégies de gestion de la base de données
- Documentation du système d'information et formations continues

Examen et amélioration de l'architecture de base

- **Introduction**
- Les concepts
  - Identifiant unique
  - Codifications principales
- Les processus : quelques pistes
  - Identification des individus
  - Alimentation de la base

  (liens étroits avec stratégies de gestion)

Architecture de base : introduction

- Approche globale : inventaire des intervenants, enjeux et besoins (concept de "royaume-émissaires")
- Relief :
  - Identification des concepts les plus importants, (employeur, entreprise, travailleur …) : périmètre
  - Identification des événements pouvant les affecter : processus
  - Examen des éléments organisationnels stratégiques
  - Identification des supports correspondants : bases de données, documentation (cartographie)

Architecture de base : "Master Data Management"

- Analyse, représentation et gestion :
  - Des données, de leurs relations et règles
  - Des composants, processus et services
  - Des applications correspondantes (applications transactionnelles, "reporting", …)
  - Des liens entre sources internes et sources externes (données, services, applications)
  - De l'évolution dans le temps de chacun de ces éléments

Architecture de base : plan

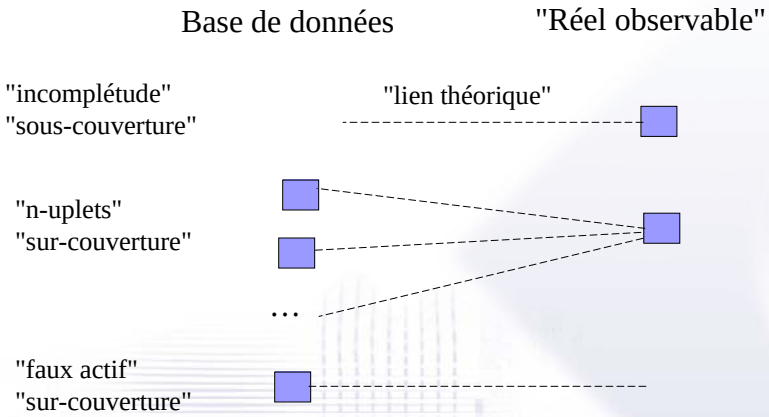
- Introduction
- **Les concepts**
  - Identifiant unique
  - Codifications principales
- Les processus : quelques pistes
  - Identification des individus
  - Alimentation de la base

  (Liens étroits avec les stratégies de gestion)

Les concepts : l'identifiant unique

- Référentiel de l'identifiant : le destinataire et non l'organisation interne
- Forme de l'identifiant : proscrire tout identifiant porteur d'information
- Test des champs associés (conversion des caractères spéciaux, …)
- Flux producteurs de l'identifiant

Questions touchant l'identifiant unique



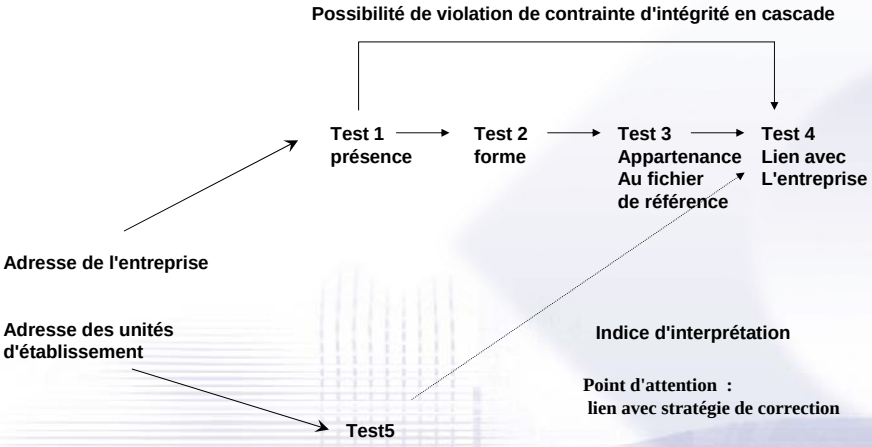
Traitement de l'identifiant unique

- Traitement des doublons ou n uplets
  - Détection préventive lors de la saisie ("warning")
  - Détection ex post (voir "tools")
  - Eléments organisationnels :
    - règle homogène de sélection d'un numéro et des valeurs correspondantes
    - feedback légal (auprès de l'instance concernée et au niveau des documents légaux)

Examen des codifications principales

- Aspects sémantiques et fonctionnels :
  - adéquation aux actions visées
  - partitions sans omissions, ni doubles emplois
  - clarté du code, des procédures de saisie, des tables de passage
- Prise en compte de la dynamique des codifications empiriques : adoption de compromis dans la conception des tables de passage
- Documentation des codes

Contrôles et séquences de tests

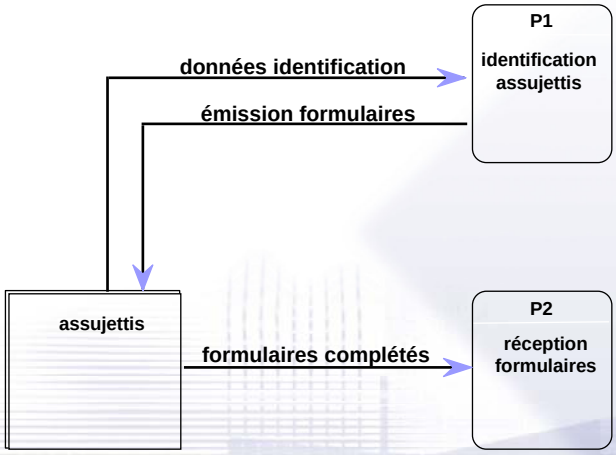


Architecture de base : plan

- Introduction
- Les concepts
  - Identifiant unique
  - Codifications principales
- **Les processus : quelques pistes**
  - Identification des individus
  - Alimentation de la base

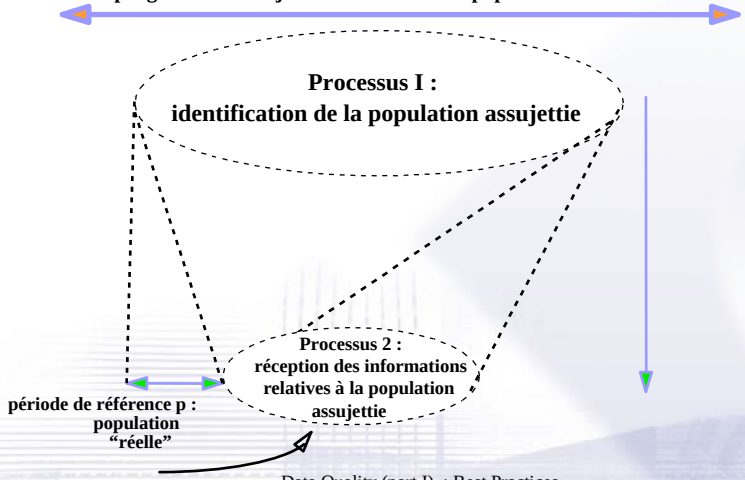
  (Liens étroits avec les stratégies de gestion)

Identification des individus



Identification des individus

Manifestation progressive et toujours inachevée de la population "réelle" relative à la période p



Suivi des décalages temporels entre bases de données

Répertoire de référence (source authentique)

immatriculation						
Matricule_I	catégorie_I	adresse_I	année_I	trimestre_I	mois_I	jour_I
xzs_pml	klm	Mons	1997	1	janvier	5

Base de données (mode "déclaratif régulier")

employeur				
matricule_E	catégorie_E	trimestre_E	année_E	adresse_E
xzs_pml	klm	1	1997	Mons

```
SELECT I.matricule_I
FROM Immatriculation I
WHERE NOT EXIST (SELECT *
                 FROM Employeur
                 WHERE I.matricule_I=matricule_E)
```


Alimentation de la base : pistes complémentaires

- Workflow de procédure pour gérer les états transitoires (cas en cours de traitement ou de validation)
- Traitement des données structurées et des documents justificatifs :
  - formulaires électroniques
  - système des codes à barre associé au format PDF

Méthodes d'amélioration

- Examen et amélioration de l'architecture de base "Master Data Management"
- **Production d'informations en vue du déploiement de stratégies de gestion de la base de données**
- Documentation du système d'information et formations continues

Production d'informations en vue du déploiement de stratégies de gestion

- **Prérequis**
- Suivi des anomalies et stratégie de gestion
- Data tracking et BPR
- Les outils
  - Aide à la décision : profiling, matching, monitoring, filtering
  - Action directe sur la base de données

Les prérequis

- Indicateurs de qualité (différents niveaux d'agrégation : cfr supra)
- Nécessité d'un système de détection d'anomalies "ex ante" et "ex post"
- Des procédures (qui traite / quoi / quand / comment) doivent être mises en place
- Un historique des anomalies (par type) et de leurs corrections/validations est indispensable (voir exemple en annexes)

Production d'informations en vue du déploiement de stratégies de gestion

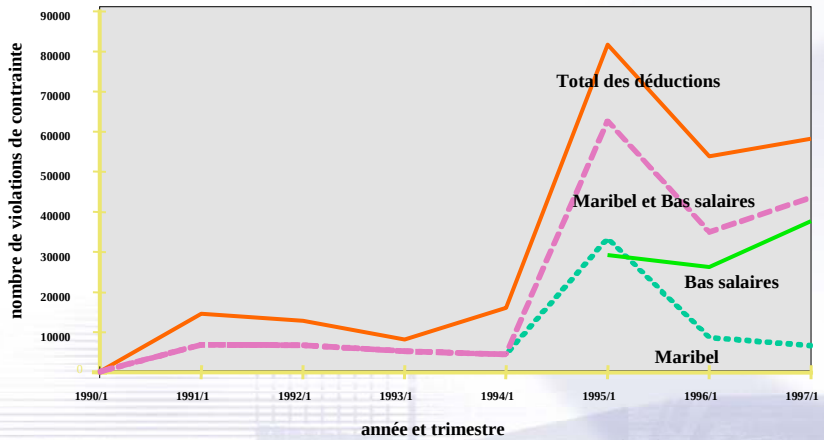
- Prérequis
- **Suivi des anomalies et stratégie de gestion**
- Data tracking et BPR
- Les outils
  - Aide à la décision : profiling, matching, monitoring, filtering
  - Action directe sur la base de données

Suivi des anomalies et stratégies de gestion

- Evaluer le processus de décision auquel sont confrontés les gestionnaires de la base :
  - temps et nature des traitements
  - nombre de validations d'anomalies formelles par donnée (anomalies formelles jugées valides au terme de l'interprétation humaine)
- Adapter ponctuellement le schéma de la base en vue de diminuer le nombre d'anomalies fictives à traiter

Un exemple : nombre d'anomalies par donnée

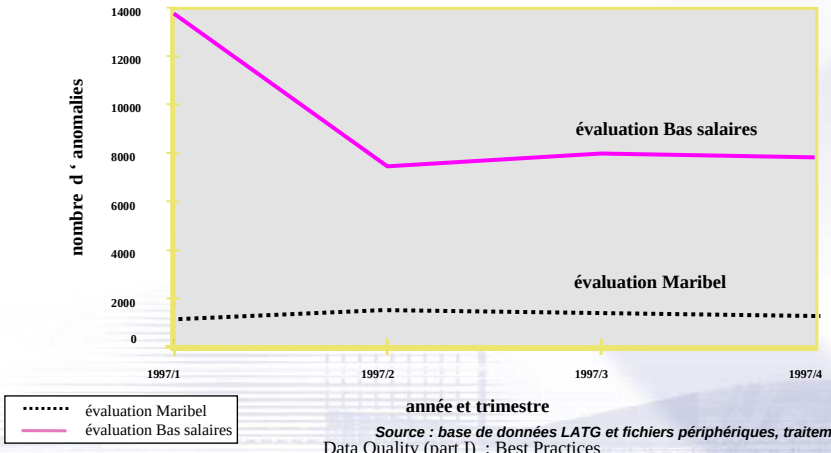
Nombre total d'anomalies par déductions de cotisation (premiers trimestres 1990-1997)



Source : base de données LATG et fichiers périphériques, traitements propres

Résultat d'une adaptation du schéma

Evolution des anomalies "à vérifier" (déduction *Bas salaires* et *Maribel*, 1997)

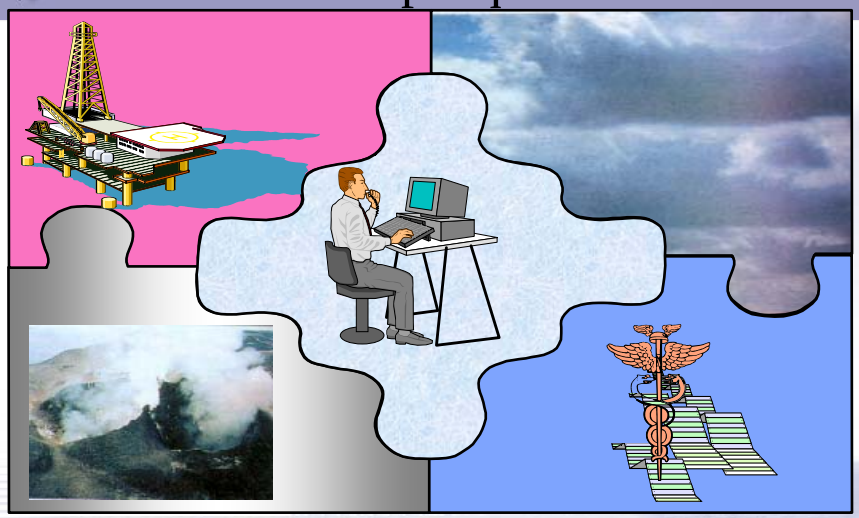


Source : base de données LATG et fichiers périphériques, traitements propres

Bénéfices de l'opération

- traitement plus homogène et rapide de la base de données
- meilleure connaissance de la signification de l'information
- diminution de la charge de travail manuel
- traitement plus fiable des flux financiers et des avantages sociaux

Généralisation à l'information empirique



Autres indicateurs utiles et stratégies de gestion associées

- Nombre d'anomalies traitées (validées ou corrigées) et temps de stabilisation
  - déterminer le moment le plus opportun pour exploiter la base
- Identifier et traiter les plages qui ne seraient jamais corrigées
- Identifier et catégoriser les pics d'anomalies
  - identification des causes (modifications législatives, lisibilité des instructions, …)

Production d'informations en vue du déploiement de stratégies de gestion

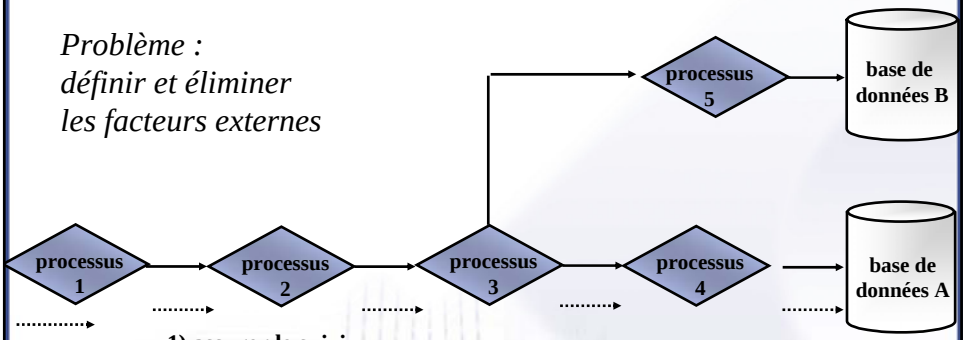
- Prérequis
- Suivi des anomalies et stratégie de gestion
- **Data tracking et BPR**
- Les outils
  - Aide à la décision : profiling, matching, monitoring, filtering
  - Action directe sur la base de données

Le "data tracking" : étude du processus d'assemblage des données

- Application des méthodes statistiques issues de l'industrie aux bases de données (AT&T Labs)
- Application spécifique en cours à la DmfA : "top 50 des employeurs commettant le plus d'anomalies prioritaires"

Le "data tracking" : étude du processus d'assemblage des données

Problème :
définir et éliminer
les facteurs externes



échantillon de l'input

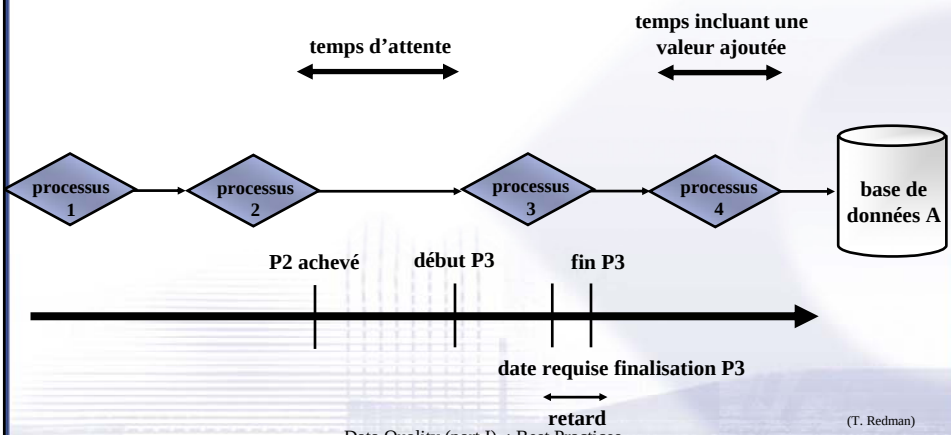
1) assurer le suivi des enregistrements de l'échantillon
2) comparer les données
3) réaliser des graphes de contrôle

(T. Redman)

Exemple de suivi d'un enregistrement

	Processus 1	Processus 2	Processus 3	Processus 4	Base de données A
Attribut a	XYZ1	XYZ1	XYZ1	**XYZ1-001**	**XYZ1-001**
Attribut b	Oui	Oui	**Non**	Non	Non
Attribut c			K	K	K
Attribut d		1500	**5100**	5100	5100
Attribut e		Z	Z	**Z**	**1**
Attribut f					**OK**
Date entrée	01/03/89	02/03/89	20/03/89	04/04/89	04/04/89
Date sortie	01/04/89	10/03/89	01/04/89	25/04/89	
Date prévue		9/03/89	30/03/89	25/04/89	01/04/89

Cycle de transformation de l'information



(T. Redman)

Data tracking : opportunité

- méthode adaptée à :
  - la détection des erreurs formelles (erreurs de programmation)
  - la diminution des files d'attente dans les traitements
  - l'analyse de collections de données dont l'évolution est stable et linéaire

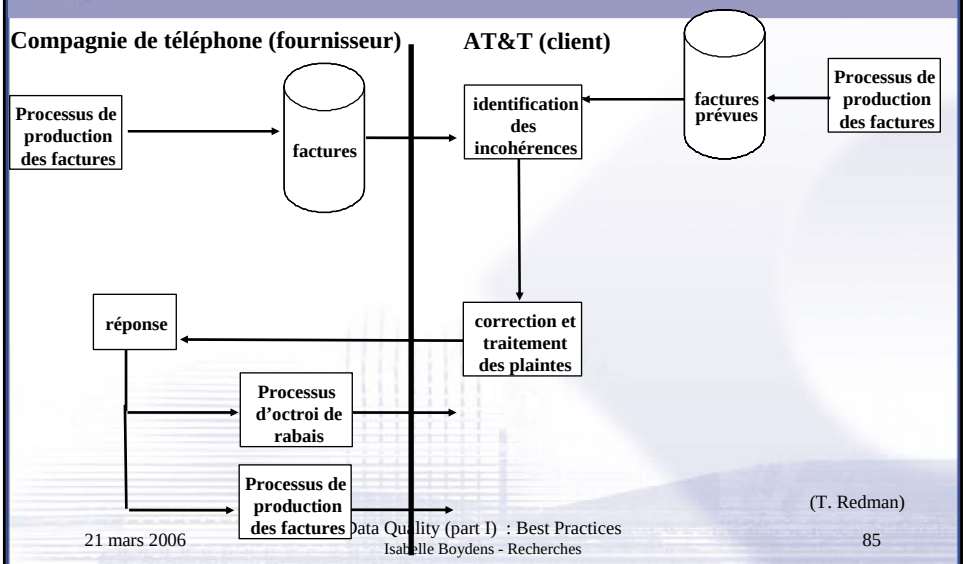
Application spécifique en cours à la DmfA

- "Top 50 des employeurs commettant le plus d'anomalies prioritaires"
- Particularités (Y. Bontemps)
  - Échantillon "non aléatoire" car connaissance a priori
  - "Tracking" arrière
- Diagnostic (variété des causes d'erreur) et actions correctrices
- Amélioration des processus et recommandations plus générales

Reengineering des processus (BPR)

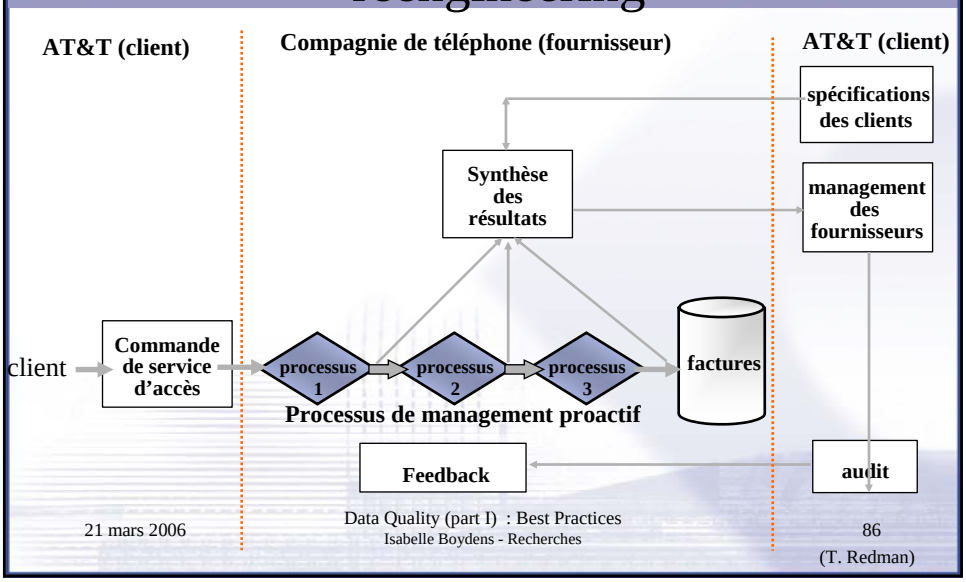
- objectifs :
  - diminution de la redondance et du risque d'émergence d'erreurs formelles
  - allègement du travail de test et de correction de l'information
- un exemple remarquable : le processus de facturation d'AT&T Laboratories

Processus de facturation initial



(T. Redman)

Processus de facturation après reengineering



Les bénéfices du reengineering

- partenariat entre clients et fournisseurs de l'information et partage de la responsabilité
- baisse significative des coûts liés à la correction de l'information (gains en personnel et en matériel) et à la gestion des plaintes et litiges
- amélioration de la qualité de l'information (liée à la suppression de la redondance initiale)

Production d'informations en vue du déploiement de stratégies de gestion

- Prérequis
- Suivi des anomalies et stratégie de gestion
- Data tracking et BPR
- **Les outils**
  - **Aide à la décision : profiling, matching, monitoring, filtering**
  - **Action directe sur la base de données: standardization, cleansing**

Les outils

- Aide à la décision :
  - Data profiling
  - Data matching
  - Data monitoring
  - Data filtering
- Action directe sur la base de données :
  - Data standardization
  - Data cleansing

→ Session "Data quality (part II) : tools" : Y. Bontemps

Méthodes d'amélioration

- Examen et amélioration de l'architecture de base "Master data management"
- Production d'informations en vue du déploiement ultérieur de stratégies de gestion de la base de données
- **Documentation du système d'information et formations continues**

Documentation et formations continues

- **Utilité**
- Définition et arbitrages
- Un exemple d'application pratique

Documentation et formations continues : utilité

- L'aspect documentaire s'inscrit dans l'une des trois fonctions de l'administration ("méta-informations")
- Trois niveaux interagissants :
  - Information juridique
  - Information administrative
  - Information technique
- L'information peut être interprétée distinctement en fonction des usages (exemple : la population active)

Documentation du système d'information

- Utilité
- **Définition et arbitrages**
- Un exemple d'application pratique

Documentation du système d'information : définition

- Méta-information (particule grecque "méta") :
  - "méta-physique" : "information sur l'information"
  - "métastase", … : notion de changement
- Plusieurs niveaux d'emploi de "méta" en informatique :
  - "méta-classe"
  - "méta-information" : schéma d'une base de données et documentation afférente
  - "méta-langage" : formalisme de modélisation
  - …

Documentation du système d'information : arbitrages

- Paradoxes :
  - Infinité des niveaux d'ordre "méta"
  - Décalages temporels entre données et méta-données
  - Importance des ressources humaines requises
    - NASA : "the metadata myth"
    - "Data tagging"
    - Bases de données temporelles, incertaines, …

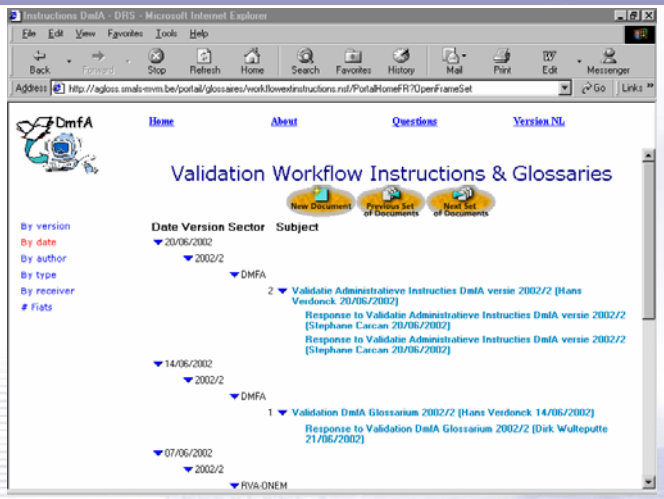
Documentation du système d'information

- Utilité
- Définition et arbitrages
- **Un exemple d'application pratique**

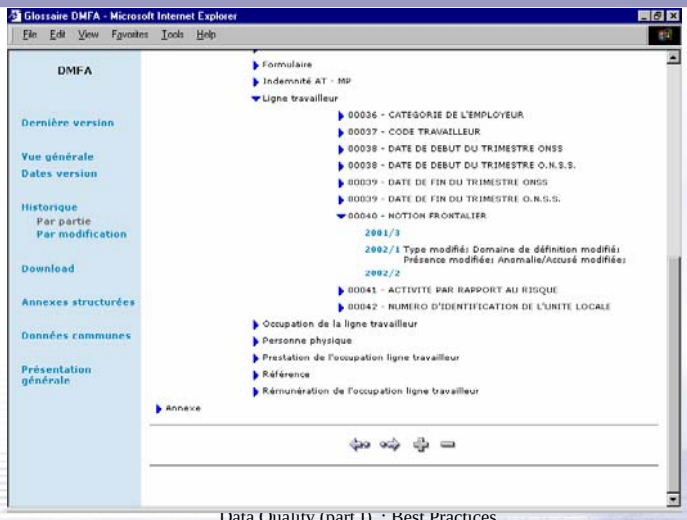
Un exemple dans le contexte de "l'e-governement"

- Glossaires de la sécurité sociale en production depuis 2001
- Section "gestion de l'information"
- Fonctionnalités :
  - Workflow de validation
  - Gestion des versions
  - Structuration de champs multilingues (thesaurus juridique)
  - Héritage et réutilisation (OO concept)
  - WOPM (*Write Once Publish Many*)
  - "*Multibase search tool*"

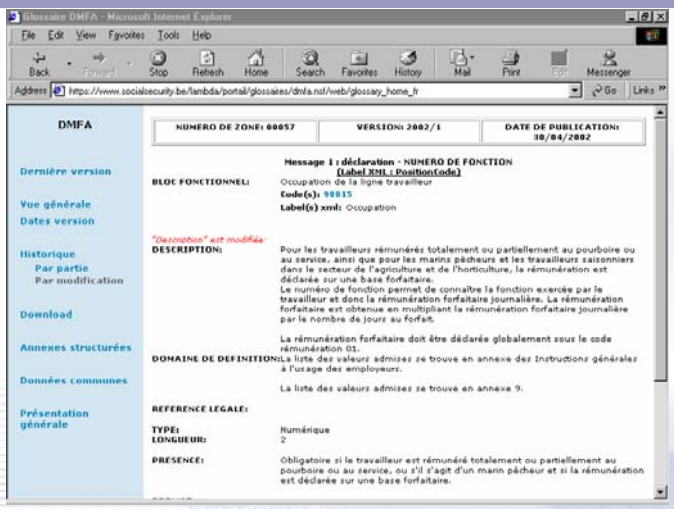
Workflow de validation



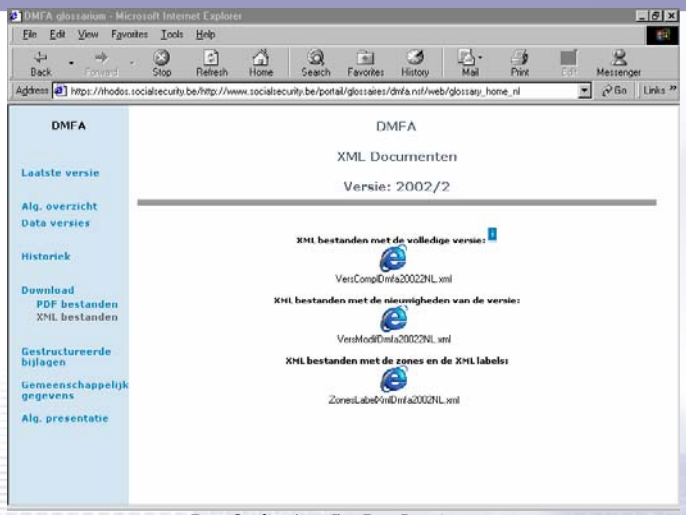
Gestion des versions (1)



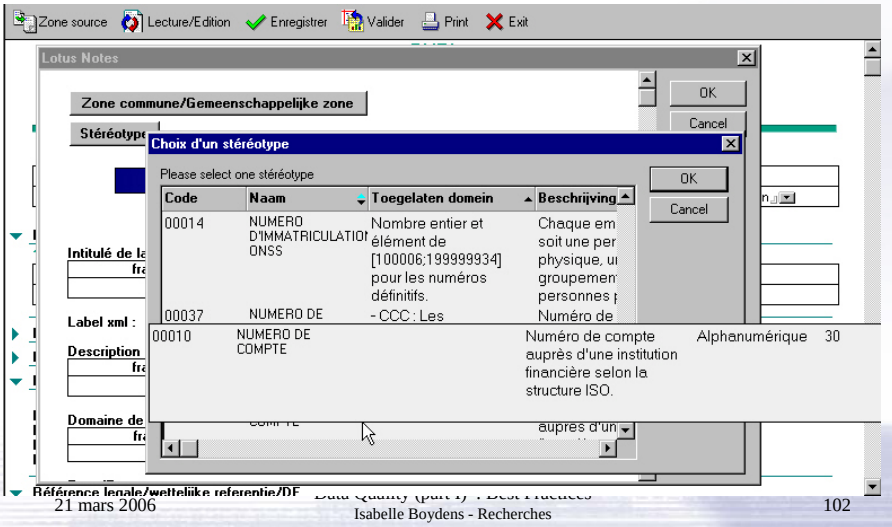
Gestion des versions (2)



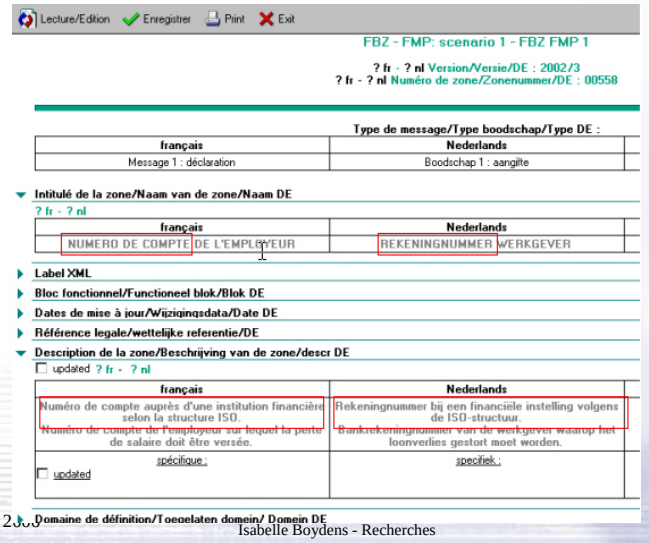
Gestion des versions (3)



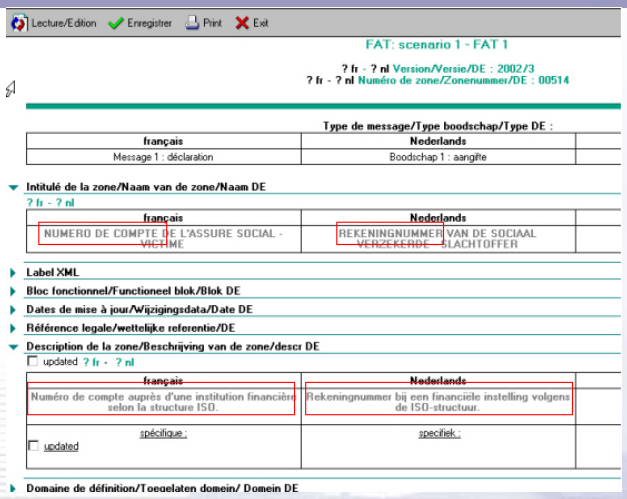
Héritage et réutilisation (1)



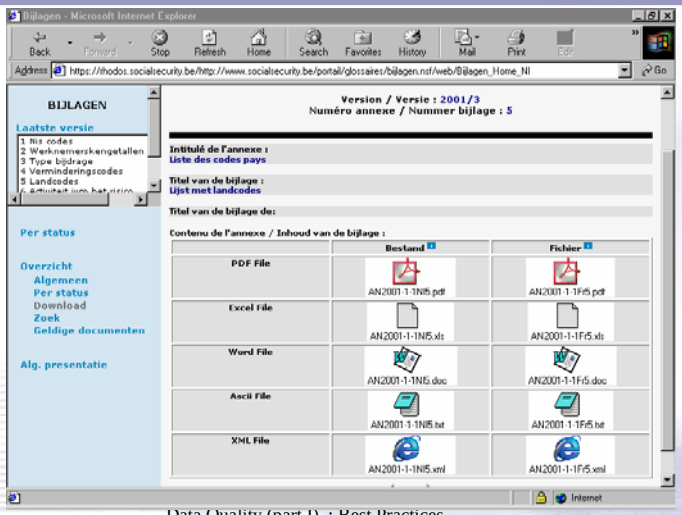
Héritage et réutilisation (2)



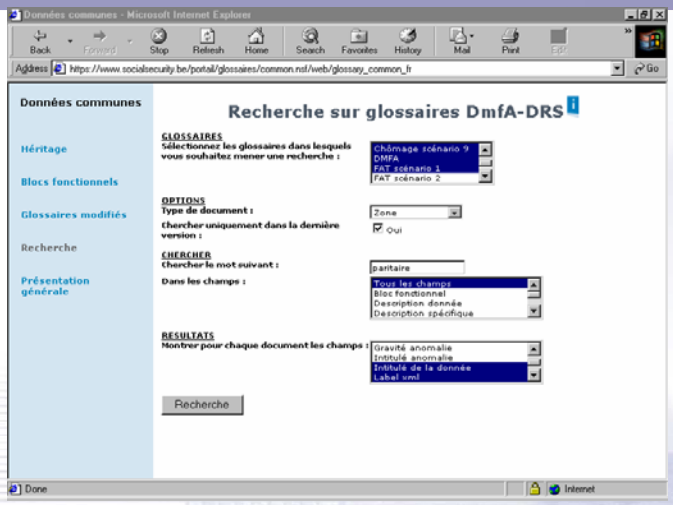
Héritage et réutilisation (3)



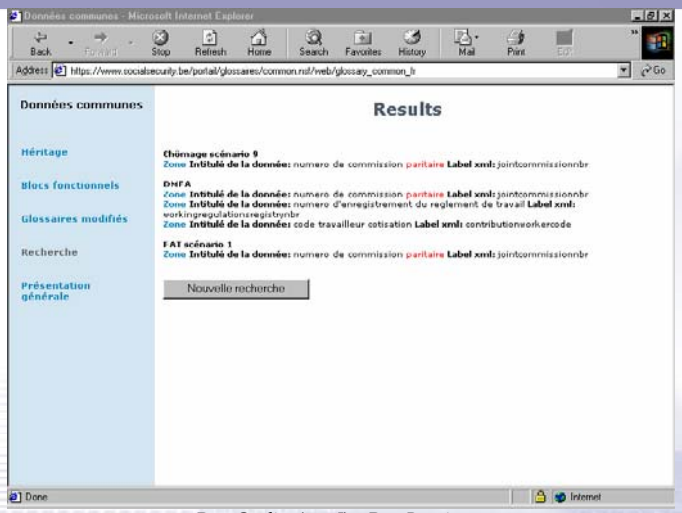
"Write Once publish Many"



"Multibase Search Tool" (1)



"Multibase Search Tool" (2)



Bénéfices d'un système de méta-information

- Documenter une base de données tout au long de son cycle de vie et en fonction des usages
- Gérer plus précisément les droits sociaux
- Permettre une “réutilisation” plus aisée des applications
- Interpréter des données issues de sources hétérogènes
- Constituer une base à la mise en place de formations continues

Plan de l'exposé

- Position du problème et enjeux
- Analyse : dimensions de la qualité des données
- Méthodes d'amélioration de la qualité
- **Conclusions**

Quelques points-clés

- Qualité des données :
  - adéquation aux objectifs et usages
  - L'information parfaite n'existe pas
  - arbitrage "coût-bénéfice"
  - La "sur-qualité" est de la "non qualité"
  - Relief : privilégier les données et les processus stratégiques
  - Indicateur crucial : "pertinence" des données

"Data Quality @ SmalS-MvM"

- Cellule "data quality" (section "recherches")
  - Études et publications de travaux
  - Consultances au sein de l'administration fédérale belge
  - Collaborations flexibles avec plusieurs équipes de la société (section "gestion de l'information", section "statistiques", …)
  - Mise en place de groupes de travail
  - Définition et suivi de projets novateurs

Questions

Orientation bibliographique (1)

- Boydens I., Evaluer et améliorer la qualité des bases de données. *Techno, publication technique de la SmalS-MvM*, janvier 1998, n° 7.
- Boydens I., Informatique, normes et temps. Bruxelles : Bruylant, 1999.
- Boydens I., Les bases de données sont-elles solubles dans le temps? In La Recherche hors série ("Ordre et désordre"). Hors série n° 9, novembre-décembre 2002, p. 32-34.
- Boydens I., E-gouvernement en Belgique. Un retour riche d'expériences. In *L'informatique professionnelle* (Dossier spécial "*Services publics*"). Paris : Editions Gartner France, Numéro 217, octobre 2003, p. 29-35
- Bloch L. Système d'information : obstacles et succès. Paris : Vuibert, 2005.
- Charlesworth I., Kellett A. et Thompson M., Data Quality and Integrity. Essential Steps for Exploiting Business Information. Hull : Butler Group, decembre 2004.

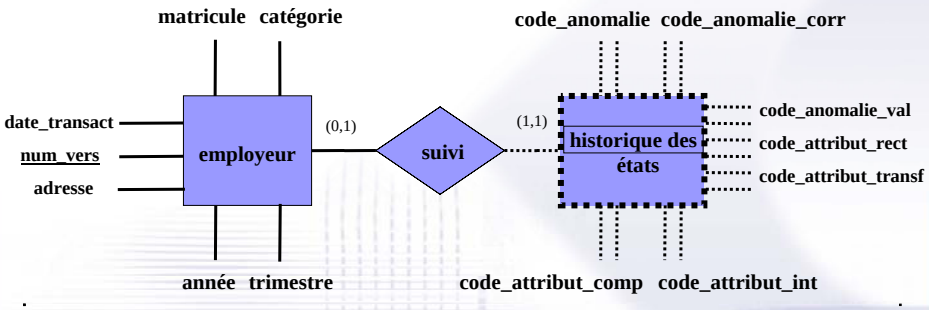
Orientation bibliographique (2)

- Moles A., Les sciences de l'imprécis. Paris : Seuil, 1995.
- Elmasri R. et Navathe S. B., Fundamentals of Database Systems, Addison Wesley, 2003.
- Redman T. C., Data Quality for the Information Age. Boston-London : Artech House Publishers, 1996.
- Redman T. C., Data Quality. The Field Guide. Boston : Digital Press, 2001.
- Rivière P., Approche coût-qualité pour l'amélioration des processus de production statistique, *Courrier des statistiques*, juin 2003, n°105-106, p. 55-65.
- Rivière P., Indicateurs de qualité en matière de production de données : quelques éléments de réflexion , *Courrier des statistiques*, septembre 2005, n°115, p. 35-40.
- Thomasset C. et Bourcier D., éds, Interpréter le droit : le sens, l'interprète, la machine. Bruxelles : Bruylant, 1997.

Annexes

Historique des anomalies et de leurs corrections

Représentation de l'historique d'instances hétérogènes



Historique des anomalies et de leurs corrections

employeur

num_vers	précédent	suivant	matricule	catégorie	trimestre	année	adresse	date-transact
lkm-1		lkm-2	xyz-56M	036	1	1997		05/02/1998
lkm-2	lkm-1		xyz-56M	**036**	1	1997	**Mons**	10/05/1998
dfm-1		dfm-2	xyz-96P	036	1	1997	Arlon	06/09/1998
dfm-2	dfm-1	dfm-3	xyz-96P	**035**	1	1997	Arlon	10/09/1998
dfm-3	dfm-2		xyz-96P	**037**	1	1997	Arlon	13/09/1998

anomalie

id_trait	num_vers	code_attribut	code_anomalie	commentaire
05	lkm-1	xzsm	01	adresse absente
02	lkm-1	spzo	53	catégorie incompatible avec la catégorie identifiée lors de l'immatriculation

anomalie_corr

id_trait	num_vers	code_attribut	code_anomalie	commentaire
05	lkm-2	xzsm	01	l'adresse absente est complétée

anomalie_val

id_trait	num_vers	code_attribut	code_anomalie	commentaire
02	lkm-2	spzo	53	validation de la catégorie incompatible avec la catégorie identifiée lors de l'immatriculation

Historique des anomalies et de leurs corrections

attribut_rect

id_trait	num_vers	code_attribut_rect	commentaire
045	dfm-2	spzo	rectification de la catégorie

attribut_int

id_trait	num_vers	code_attribut_int	commentaire
021	dfm-3	spzo	interprétation et modification de la catégorie