Smals
ICT for society

Gestion intégrée des anomalies

28 octobre 2010

I. Boydens, D. Van Dromme, A. Hulstaert
Section Recherches

M. Dessart, E. Durwael
Data Support Group

Table des matières

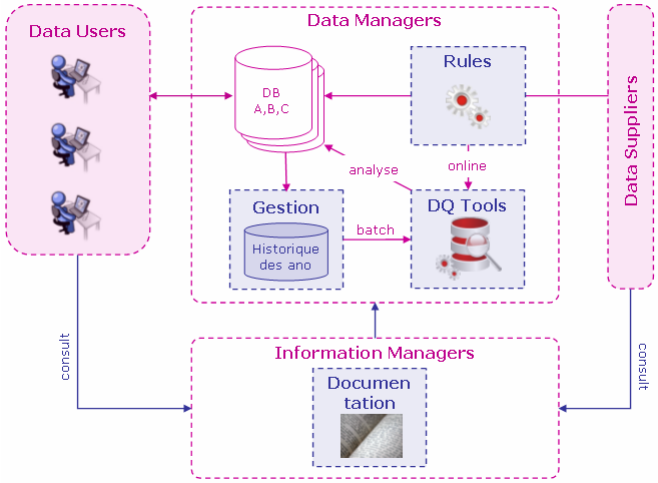
Table des matières

Smals ICT for society
28/10/2010

Contexte de l'étude

- Data quality : enjeux stratégiques pour l'egovernment
  - ex. DmfA : *Déclaration multifonctionnelle/ multifunctionele Aangifte*
  - 40 milliards € (cotisations et prestations sociales)
- Le « data quality competency center » de Smals
  - Consultances
  - Études (voir bibliographie)
- Retours d'expérience à travers la question des anomalies
  - Impacts : 15 % du revenu des entreprises
  - But : en maîtriser le traitement et en diminuer le nombre
  - Nouveautés de l'étude, quelques points :
    - Modélisation générique de l'historique des anomalies (principes et exemple de maquette) et stratégies de gestion
    - Documentation des anomalies (Falco)
    - Apports des « data quality tools »

Contexte de l'étude



Contexte de l'étude

Un système d'information est un fleuve : 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'anomalies (T. Redman)

Contexte de l'étude

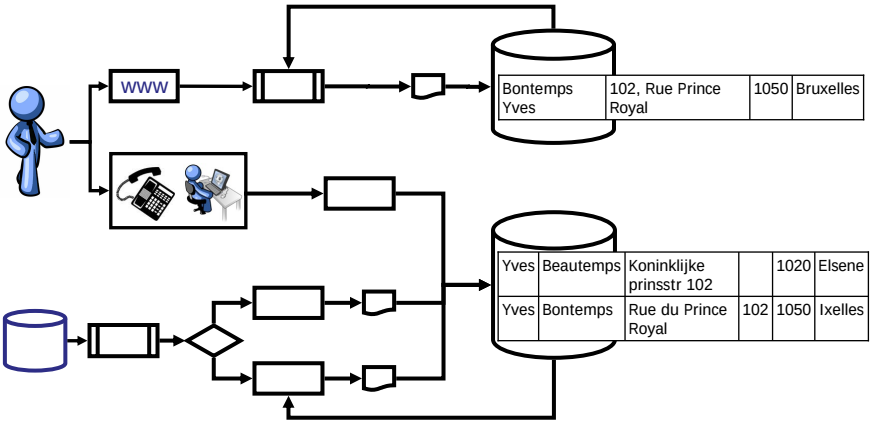


Table des matières

Smals
ICT for society

28/10/2010

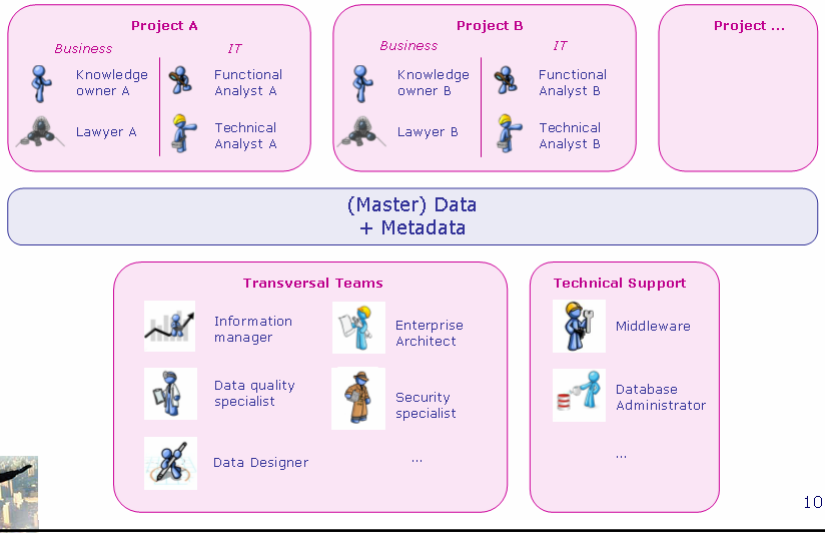
Définitions (rappels)
Qualité d'une base de données

- « Fitness for use » : adéquation aux objectifs
- Arbitrage « coûts bénéfices »
- Approche pluridisciplinaire (technique et métier)
- Enjeux : données = action sur le réel
  - Monde des entreprises
  - Domaines militaires, scientifiques, …
  - Egovernment
    - Coûts en vérification et correction des données
      - Services du contrôle affectés à cette tâche
      - En Europe, effectif de plusieurs dizaines de personnes par institution
    - Impact juridique et social
    - Image et crédibilité
    - Intégration de nouvelles technologies
    - Stratégie et management

Smals
ICT for society

28/10/2010

Définitions
Approche pluridisciplinaire



Définitions
Caractéristiques des DB administratives

- Modifications législatives fréquentes et complexes → gestion des versions
- Force probante des données → historique des records
- "Idéalement", pas de tolérance à l'erreur (traitement équitable des dossiers des citoyens)
- Volume de données et d'anomalies important
  - Ex : DmfA, par trimestre, environ :
    - 4 millions d'enregistrements
    - 10% d'anomalies formelles (cf autres secteurs : banques, …)
- Incidences sociales et financières considérables

Définitions
Typologie des systèmes d'information administratifs

- Bases de données administratives structurées
  - Bases de données reposant sur un prélèvement régulier d'information
  - Répertoire ou « référentiel »
- Systèmes d'information documentaires (incluant un SGBD pour les méta-données)
- « Sources authentiques » :
  - Stratégique dans les projets : données de référence
  - Approche pragmatique :
    - Qualité relative des données
    - Institution/service en charge de la gestion d'une source authentique

Définitions
Concept d'anomalie

- Trois questions
  - Qu'est-ce qu'une donnée ?
  - Qu'est-ce qu'une donnée « correcte » ?
  - Comment les données se construisent-elles progressivement ?

Définitions
Qu'est-ce qu'une donnée ?

- Triplet :
  - Concept (ex : salaire mensuel)
  - Domaine de définition (ex : « valeur numérique incluse entre 1000 € et 100.000 € »)
  - Valeur à un instant t : 3000 €
- Différence entre données
  - Déterministes : définition immuable
  - Empiriques : définition évolutive avec l'interprétation humaine du réel
- « Closed world assumption »

Définitions
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

Id.	Nom	Prenom	Categorie	Taux-cotisation
km-pod	Durant	Jean	énergie renouvelable	0.27 %

Categorie_taux

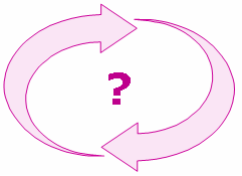
Catégorie	Taux-cotisation
énergie solaire	0.28%
énergie éolienne	0.27%
énergie biomasse	0.29%

Définitions
Qu'est-ce qu'une donnée « correcte » ?

- Typologie des violations de contrainte d'intégrité
  - Erreur formelle
  - Présomption formelle d'erreur : anomalie
  - Erreur indétectable formellement

Définitions
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

Définitions
Comment les données se construisent-elles progressivement ?

Evolution des normes (législation, théorie)

Evolution des représentations administratives et informatiques

Evolution du réel observable, objet de la norme et de la représentation informatique

Table des matières

Modélisation de l'historique des anomalies

- Extension originale du "conceptual modelling" en vue :
  - de passer d'un monde clos à un monde ouvert sous contrôle
  - de rendre opérationnels les principes de gestion des anomalies posés et expérimentés (exemple de maquette)
  - du suivi dans le temps du traitement des anomalies :
    - détection / correction / validation …
    - impact fonctionnel et organisationnel adapté au contexte (nouvelle db, reengineering ou db existante)
  - de concevoir des indicateurs de qualité et des stratégies de gestion pour diminuer structurellement le nombre d'anomalies

Modélisation de l'historique des cas d'anomalies
Prérequis structurels concernant la DB applicative

- Chaque table de la DB applicative concernée par la gestion des anomalies doit gérer son historique par la clôture de l'enregistrement courant et l'insertion d'un nouvel enregistrement
- La gestion de l'historique au travers de ses identifiants uniques permet toujours de retrouver la clef fonctionnelle associée au contenu sémantique

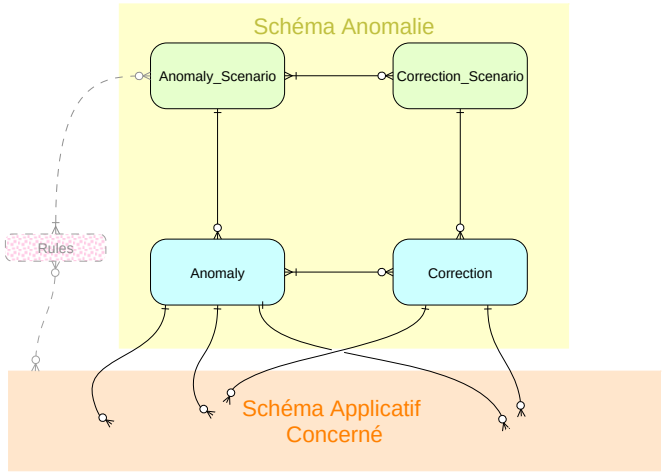
Généralement on ajoute à la clef fonctionnelle :
- Date et heure de création de l'enregistrement (CREATED_TMS)
- Date et heure de fin de l'enregistrement (END_DATE)
  ou au minimum un indicateur (ACTIVE_IND)

Modélisation de l'historique des cas d'anomalies
Généricité et incidence minimum sur la DB applicative

- L'implémentation doit pouvoir trouver sa place, de la même manière, dans des DB existantes comme au sein des nouvelles architectures
- La démarche doit pouvoir supporter la gestion automatique comme manuelle
- Le couplage entre l'application et la gestion des anomalies doit être le plus léger possible et ne nécessiter aucune modification des applications déjà en place

Ajout d'un champ dans chaque table de la structure originelle :
- La clef unique du numéro d'enregistrement en anomalie (RECORD_ANOMALY_ID)

Modélisation de l'historique des cas d'anomalies
Schéma Anomalie à côté du schéma Applicatif

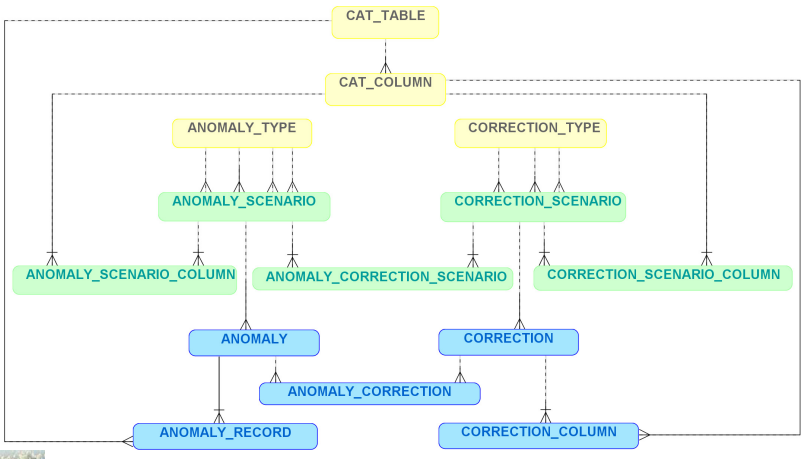


Modélisation de l'historique des cas d'anomalies
Structuration en trois couches

- Le catalogue des tables et des champs du schéma originel et le typage pour le regroupement statistique
- Les tables de références fonctionnelles reprenant toutes les descriptions des cas d'anomalies et de corrections répertoriés
- L'inventaire des anomalies et des corrections identifiées dans la base de données applicative et dont les enregistrements sont marqués par le RECORD_ANOMALY_ID

Modélisation de l'historique des cas d'anomalies

Schéma Anomalie : Modèle conceptuel



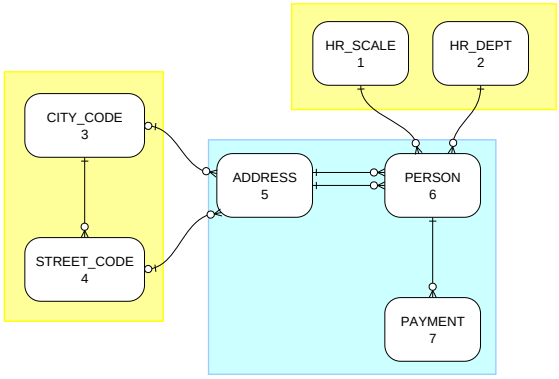
Modélisation de l'historique des cas d'anomalies

Déroulement pour implémenter la démarche

- Remplir le dictionnaire technique et business :
  - Le catalogue avec les tables et les colonnes concernées du schéma applicatif
  - Les types d'anomalie et de correction, dont la spécification et la granularité dépendent de la base de données originelle
  - La description des scénarios d'anomalies et de corrections à compléter et à affiner au fur et à mesure des besoins
- Remplir l'inventaire avec les anomalies et les corrections réelles détectées dans la DB applicative

Modélisation de l'historique des cas d'anomalies

Exemple : une base de données applicative



Exemple : Phase de détection des cas d'anomalie

Schéma Applicatif

PERSON (6)

PK	CREATED_ TMS	END_ TMS	yyy		...	xxx	...	RECORD_ ANOMALY_ID
536 814	*Tms 1*	NULL	γγγ		...	α α	...	2564

Détection anomalie

ANOMALY_ID counter: 1433

RECORD_ANOMALY_ID counter: 2564

ANOMALY

ANOMALY_ID	CREATED_ TMS	ANOMALY_ SCENARIO_ID	RESOLVED_IND
1433	*Tms 2*	A101-0610	N

ANOMALY_RECORD

ANOMALY_ID	RECORD_ ANOMALY_ID	TABLE_NR	ORIGINAL_TMS
1433	2564	6	*Tms 1*

Schéma Anomalie

Exemple : Phase de correction des cas d'anomalie

Schéma Applicatif

PERSON (6)

PK	CREATED_ TMS	END_ TMS	...	xxx (8)	...	RECORD_ ANOMALY_ID
xxx	*Tms 1*	*Tms 3*	...	ααα	...	2564
xxx	*Tms 3*	*NULL*	...	ωωω	...	2564

ANOMALY

ANOMALY_ID	CREATED_ TMS	ANOMALY_ SCENARIO_ID	RESOLVED_IND
1433	*Tms 2*	A101-0610	Y

Anomalie correction

CORRECTION

CORRECTION_ ID	CREATED_ TMS	CORRECTION_ SCENARIO_ID	ANOMALY_ID
870	*Tms 3*	C101-0610	1433

CORRECTION_COLUMN

CORRECTION_ ID	RECORD_ ANOMALY_ID	COLUMN_ ID	ORIGINAL_TMS
870	2564	8	*Tms 1*

ANOMALY_ID counter: 1433

RECORD_ANOMALY_ID counter: 2564

CORRECTION_ID counter: 870

Schéma Anomalie

Modélisation de l'historique des cas d'anomalies
Alternatives et mises en garde

- Substituer un numéro de version à la START_DATE
- Remplacer la END_DATE par un indicateur : actif/inactif
  - Voire se passer de l'information de clôture et la déduire de l'existence d'un enregistrement plus récent
- La dissociation de la gestion des anomalies par rapport à l'applicatif traité doit faire l'objet d'une surveillance constante

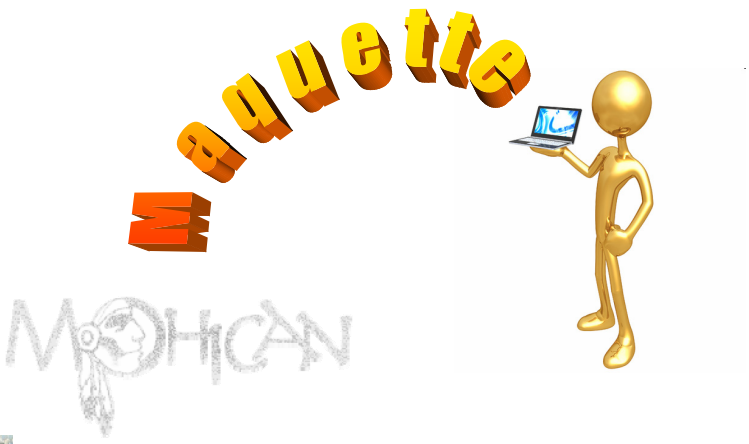
Cohérence des données

Implémentation de la contrainte relationnelle entre ANOMALIE_RECORD et les enregistrements de chaque table

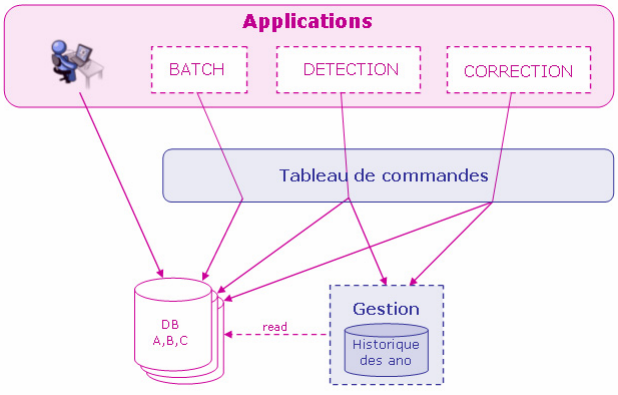
Performance & **Indépendance**

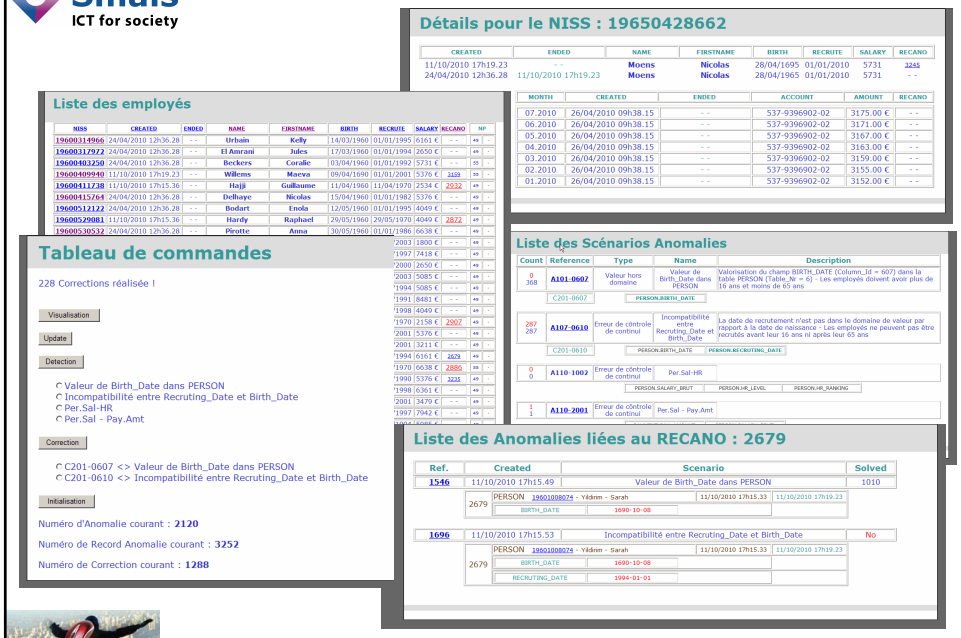
Seuls les programmes de gestion des anomalies ont les droits de modification sur les RECORD_ANOMALIE_ID applicatifs

Modélisation de l'historique des cas d'anomalies
Démo



Modélisation de l'historique des cas d'anomalies
Positionnement de la maquette





Modélisation de l'historique des cas d'anomalies

Règles de mises en œuvre pour une démarche automatique

- Toute phase de correction pour un scénario d'anomalie donné doit être obligatoirement précédée d'une phase de détection de ce même scénario d'anomalie (ceci afin que la correction s'effectue sur l'enregistrement actif détecté en anomalie)
- La phase de détection d'un scénario d'anomalie doit également positionner dans un statut d'obsolescence (RESOLVED_IND = "O") les anomalies non traitées et liées à des enregistrements historisés (RESOLVED_IND = "N" and END_TMS not null)

Modélisation de l'historique des cas d'anomalies
Statistiques : quelques requêtes SQL de gestion

- Suivi des anomalies sur l'ensemble des périodes de référence :

```
SELECT YEAR(created_tms)as year
    ,MONTH(created_tms)as month
    ,anomaly_scenario_id as anomaly_scenario ,count(*)
FROM anomaly
GROUP BY year, month, anomaly_scenario;
```

- Nombre d'anomalies détectées mais finalement valides :

```
SELECT COUNT(*)
FROM correction COR, anomaly_correction ANCO
WHERE COR.correction_id = ANCO.correction_id
AND COR.correction_scenario_id = 10       -- ne rien faire
AND COR.created_tms BETWEEN '2010-01-01' and '2010-08-31';
```

Table des matières

Indicateurs de qualité
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 et d'interprétation (méta-informations)
- Industrialiser la production
- Définir des stratégies d'amélioration

Source : P. Rivière, INSEE, 2005

Stratégies de gestion (case studies)
Suivi et amélioration de la structure de la DB (1)

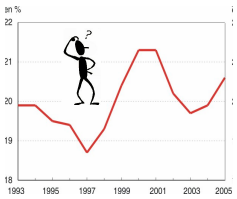
- Comment diminuer le nombre d'anomalies ?
- Evaluer le processus de décision auquel sont confrontés les gestionnaires de la base :
  - temps et nature des traitements
  - évolution du nombre de validations d'anomalies formelles (anomalies formelles "fictives" jugées valides au terme de l'interprétation humaine)
- Adapter ponctuellement le schéma de la base en vue de diminuer le nombre d'anomalies fictives à traiter
  - exemple d'application concrète (LATG – DmfA) :
    - déduction de cotisation pour les "bas salaires"
    - baisse structurelle du nombre d'anomalies de 50 % (14.000/7.000)

Stratégies de gestion (case studies)
Suivi et amélioration de la structure de la DB (2)

- 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

Stratégies de gestion (case studies)
Autres indicateurs utiles et stratégies associées

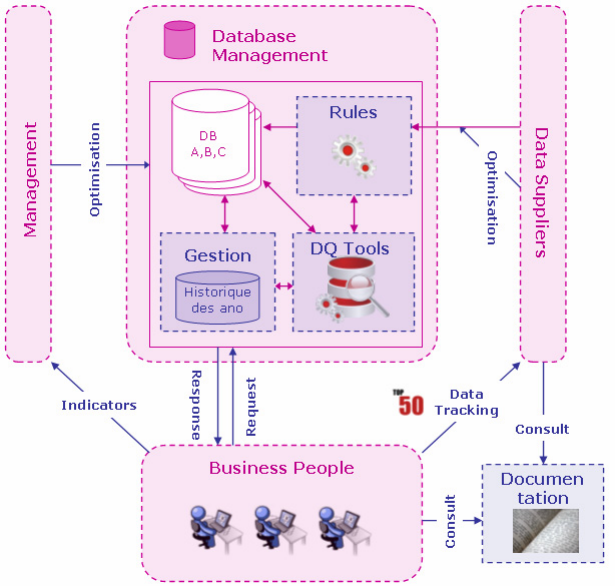
- Nombre d'anomalies traitées (validées ou corrigées) et temps de stabilisation
  - déterminer le moment le plus opportun pour exploiter la DB
- Identifier et traiter les plages qui ne seraient jamais corrigées
- Identifier et catégoriser les pics d'anomalies
  - identification des causes (modifications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



Stratégies de gestion (case studies)
Data tracking

- Méthode de T. Redman, AT&T (validité formelle des données et temps de traitement)
- Application à la sécurité sociale
  - "Top 50" des employeurs commettant le plus d'anomalies
  - Diagnostic (variété des causes) et actions correctrices structurelles
  - Opération peu coûteuse et positive en terme de relations avec les citoyens
  - Perspectives d'application pour la traçabilité des processus internes et externes et leur amélioration

Aspects organisationnels



Pause

Table des matières

Modélisation de la séquence des contrôles
Introduction

- Contexte
  - Sur la base de consultances menées sur le terrain
  - Contrôles formels (<> contrôle de conformité à la réalité)
  - Les contrôles sont exécutés séquentiellement
  - Dépendance des contrôles
  - L'ordre des contrôles est important

- Enjeux : métiers - financiers
  - Les contrôles influent sur la qualité des données
  - Conséquence sur la charge de travail des institutions

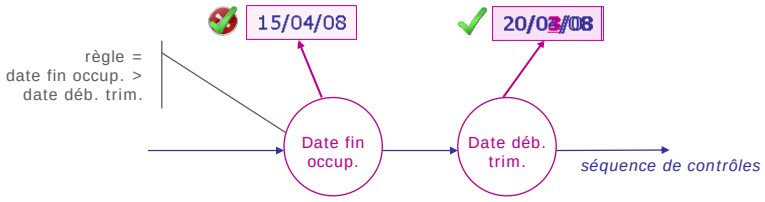
Modélisation de la séquence des contrôles
Critères de modélisation

- Dépendance entre données
  - Contrainte d'intégrité référentielle
  - Dépendance fonctionnelle
  - Dépendance/règle métier

- Rythmes de mise à jour
  - risque d'anomalies si les sources sont pas mises à jour en même temps

Modélisation de la séquence des contrôles
Conflits potentiels et arbitrage (1)

- Objectifs
  - séquence sans conflit n'existe pas !
  - déterminer ses besoins (*fitness for use*)
  - minimiser les conflits sur la base d'arbitrage
  - éviter autant que possible les anomalies fictives



Modélisation de la séquence des contrôles
Conflits potentiels et arbitrage (2)

- Arbitrage : qualité vs disponibilité des données
  - Pour des données stratégiques, aucune anomalie formelle ne peut être détectée
    - exemple : catégorie d'employeur dans la DmfA
  - Impossibilité de rejeter toutes les données
    - exemple : DmfA = 40 milliards € / an
  - Nécessité de trouver le juste équilibre (arbitrage type "coûts-bénéfices")
  - Mise en place d'indicateurs

Modélisation de la séquence des contrôles
Conflits potentiels et arbitrage (3)

- Arbitrage : contrôle des données vs besoins et ressources disponibles
  - ↗ contrôles = ↗ anomalies = ↗ ressources nécessaires
  - une augmentation du contrôle de la qualité peut se traduire par un rapport "coûts bénéfice" négatif en raison du temps liés à la correction des anomalies supplémentaires détectées

Modélisation de la séquence des contrôles
Conflits potentiels et arbitrage (4)

- Arbitrage : rapidité vs stabilité
  - contrôle effectué rapidement (données instables)
    - feedback rapide
    - présence d'un plus grand nombre d'anomalies fictives
  - contrôle effectué plus tard (données + stables)
    - feedback moins rapide
    - moins d'anomalies fictives

Modélisation de la séquence des contrôles
Recommandations

- Collaboration business - IT nécessaire
- Documenter les contrôles
- Si plusieurs sources impliquées dans un contrôle, analysez
  - les rythmes de mise à jour
  - les domaines de définition

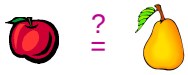


Table des matières

Smals
ICT for society

28/10/2010

Data Quality met Data Quality Tools
Constatatie

- Overal waar Data Quality een rol speelt
  - projecten, governance programma's
    - Data Integratie, Data Migratie , Re-engineeringen, dubbel- en incoherentiedetectie, Anomaliebeheer, …
  - drie 'fasen'
    - Discovery = analyse
    - Validation door business
    - } iteratieve analyse met business ← essentieel voor succes
    - Exploitatie
- Data Quality Tools doen dit beter, dankzij
  - Performantie
  - Productiviteit
  - Flexibiliteit
  - configureren ipv ad hoc programmeren
  - data matching: 1 miljoen records/uur
  - GUI/visualisatie/drill-down/export
    - vorm ideaal voor iteratief overleg

Smals
ICT for society

Anomaliebeheer met DQ-tools

DQ Tools

Business Logica
-definities (anomalie, …),
-business rules,
-correctiescenario's,
-standaardisatiekeuzes,
-algoritmen,
-Δ Business Process
-Δ Applications,
-Δ DB-constraints, …

Validation | Batch DQ Analyse | Discovery
iteratief overleg
result <!,?>
result ✓
extract
update
Business
Applications
DB A,B,C
Documentation
DQ Tools
DQTeam
Gestion
Historique des ano
Exploitatie batch / online
result
extract
Online DQ
Exploitatie
DQ Tools, J2EE, SaaS, …

Data Quality Tools
Functionaliteiten

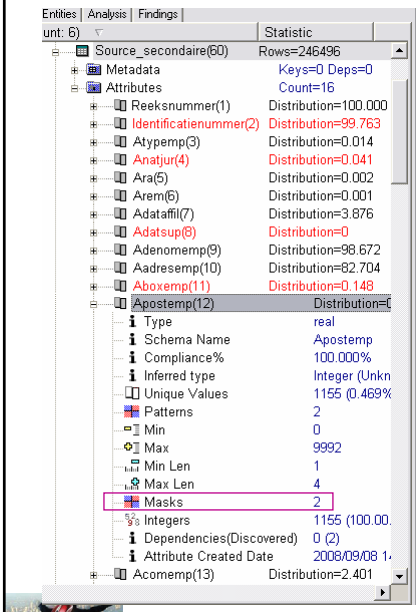
- Data Profiling
- Data Standaardisatie
- Data Matching / fuzzy matching / record linkage

Toepassingsvoorbeelden
Link met anomalieën, documentatie
Link met Gestion Historique des Anomalies

Data Profiling
Definitie

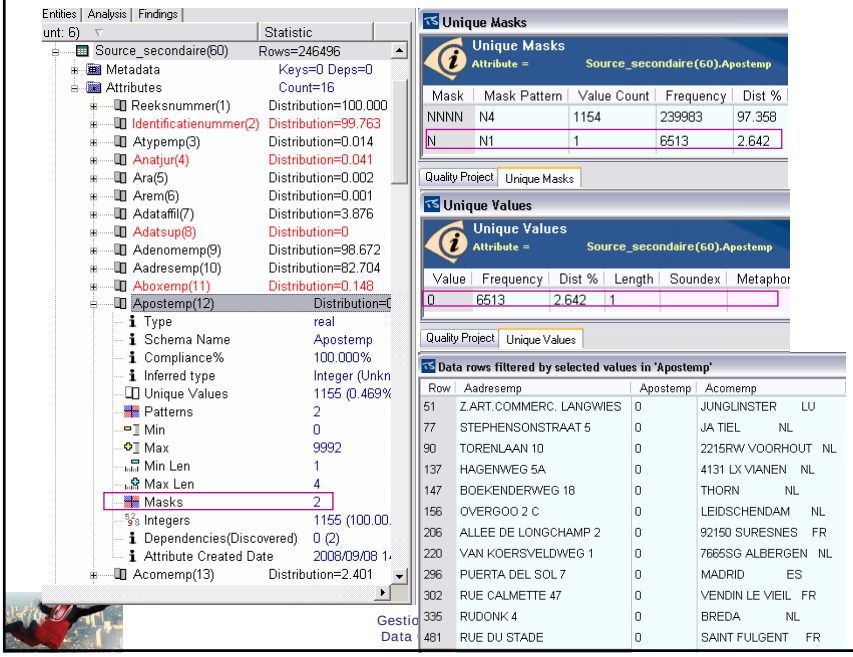
- Formele audit van gegevens en documentatie
  - *Jack E. Olson, "Data Quality – The Accuracy Dimension"*
  - input: werkelijke data + documentatie, metadata (kwaliteit onbekend)
  - output: Data Quality Issues (kwaliteit in kaart gebracht) + ondersteuning correctie documentatie
- Data Profiling met Data Quality Tools
  - Veld per veld: Column Property Analysis
  - Structuur: referentiële constraints
  - Consistency: business rules

Data Profiling – Column Property Analyse, drill-down (1)



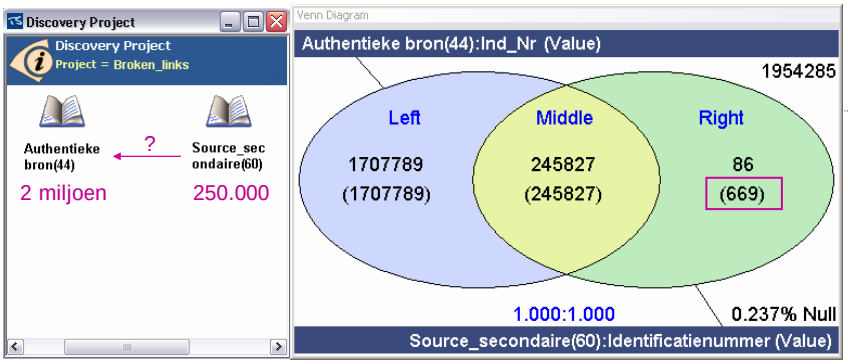
- (1)-(16) geanalyseerd tijdens load
  - gegenereerde metadata
    - Min, Max, Lengtes,
    - Null Values, Unique Values
    - Patronen,
    - Distributies,
    - Business Rules, …
- Apostemp(12): postcode werkgever
  - vb. patronen, a.h.v. Masks

Data Profiling – Column Property Analyse, drill-down (2)



Smals
ICT for society

Data Profiling – Referentiële integriteit, foreign keys (1)

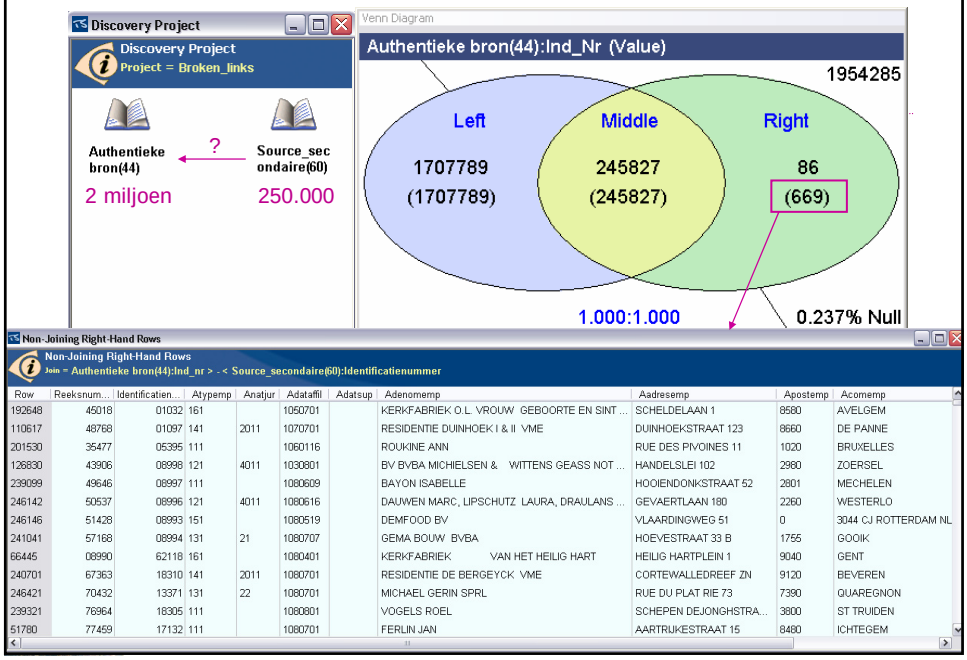


- bron 1: Authentieke bron(44)
  sleutel: Ind_nr
- bron 2: Source_secondaire(60)
  Identificatienummer verwijst naar Ind_nr's

- **Confrontatie bron 1 – bron 2**
  86 waarden niet in authentieke bron
  in 669 records (er zijn dus dubbels)

Smals
ICT for society

Data Profiling – Referentiële integriteit, foreign keys (2)



Row	Reeksnum...	Identificatien...	Atypemp	Anatjur	Adatafil	Adatsup	Adenomemp	Aadresemp	Apostemp	Acomemp
192648	45018	01032	161		1050701		KERKFABRIEK O.L. VROUW GEBOORTE EN SINT ...	SCHELDELAAN 1	8580	AVELGEM
110617	48768	01097	141	2011	1070701		RESIDENTIE DUINHOEK I & II VME	DUINHOEKSTRAAT 123	8660	DE PANNE
201530	35477	05395	111		1060116		ROUKINE ANN	RUE DES PIVOINES 11	1020	BRUXELLES
126830	43906	08998	121	4011	1030801		BV BVBA MICHIELSEN & WITTENS GEASS NOT ...	HANDELSLEI 102	2980	ZOERSEL
239099	49646	08997	111		1080609		BAYON ISABELLE	HOOIENDONKSTRAAT 52	2801	MECHELEN
246142	50537	08996	121	4011	1080616		DAUWEN MARC, LIPSCHUTZ LAURA, DRAULANS ...	GEVAERTLAAN 180	2260	WESTERLO
246146	51428	08993	151		1080519		DEMFOOD BV	VLAARDINGWEG 51	0	3044 CJ ROTTERDAM NL
241041	57168	08994	131	21	1080707		GEMA BOUW BVBA	HOEVESTRAAT 33 B	1755	GOOIK
66445	08990	62118	161		1080401		KERKFABRIEK VAN HET HEILIG HART	HEILIG HARTPLEIN 1	9040	GENT
240701	67363	18310	141	2011	1080701		RESIDENTIE DE BERGEYCK VME	CORTEWALLEDREEF ZN	9120	BEVEREN
246421	70432	13371	131	22	1080701		MICHAEL GERIN SPRL	RUE DU PLAT RIE 73	7390	QUAREGNON
239321	76964	18305	111		1080801		VOGELS ROEL	SCHEPEN DEJONGHSTRA...	3800	ST TRUIDEN
51780	77459	17132	111		1080701		FERLIN JAN	AARTRIJKESTRAAT 15	8480	ICHTEGEM

Smals ICT for society

Data Profiling – Functional Dependencies

Dependencies (Discovered) — Entity Adres Communicatie(350) — Detectie, tijdens load — quality | sample 10.000 | conflicten

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verif
C Taalcode,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
C Taalregime,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
D Begindatum,C Postcode	C Nis Gemeentecode	Discovered	No	60	98.760	9876	62	129	
D Begindatum,Gemeentenaam	C Nis Gemeentecode	Discovered	No	60	99.990	9999	1	2	
D Ts Lwijz	D Ts Creatie	Discovered	No	60	98.450	9845	109	232	
Gemeentenaam,Landnaam	C Nis Gemeentecode	Discovered	No	60	99.800	9980	3	6	
Straatnaam 21	Straatnaam Voll	Discovered	No	60	99.350	9935	29	65	
Straatnaam Voll	Straatnaam 21	Discovered	No	60	99.140	9914	68	136	

Dependencies (Verified) — Entity Adres Communicatie(350) — Verificatie, on demand — exhaustief, 2.9 miljoen

Lh Attrs	Rh Attr	Status	Verified	Job	Quality %	Confirming LR Values	Conflicting LH Values	Conflicting Rows	Verified Date	Verified By
Landnaam	C Landcode	Permanent	Yes	-	99.960	2935363	5	10	2010/04/06 16:17:19	smals

drill-down naar overzicht van conflicten indien Quality < 100%

Dependency Conflicts — Adres Communicatie(350) Dependency Landnaam -> {C Landcode}

Frequency	Landnaam	C Landcode
2854		150
233		
1292	Allemagne	173
945	Allemagne	134
3	Angola	341
1	Angola	381
20	Bahamas	425
5	Bahamas	484
18	Tchécoslovaquie	130
5	Tchécoslovaquie	171

Drill down to Matching Rows
Resolve Conflicts...
List Corrections...
Filter...
Bookmark
Convert view to Entity...
Back
Forward
Change View
Export
Export to Server
Copy

Smals ICT for society

Data Profiling – Business Rules (1)

Sub Contractor5(490) Rows=104930
Metadata Keys=0 Deps=0
Attributes 12
Rows Loaded 104930
Business Rules(Passed) 2(0)
Keys(Discovered) 0 (3)
Dependencies(Discover... 0 (5)
Date Created 2010/07/
Entity State Fully Loa
Attributes Count=12
Denomination(1) Distribut
Id_1(2) Distribut
Id_2(3) Distribut
Id_3(4) Distribut
Rue(5) Distribut
Rue Num(6) Distribut
Rue Box(7) Distribut
Commune(8) Distribut
Postcode(9) Distribut
Country Code(10) Distribut
Creation Date(11) Distribut
Max Work End Date(12) Distribut

Entity Business Rule
Enabled
Name: Présence_Identifiant
Description: Au minimum 1 des 3 types d'identifiant doit être rempli
NOT (Id_1 = "" AND Id_2 = "" AND Id_3 = "")

Entity Business Rule
Enabled
Name: Enregistrement à temps
Description: Date de déclaration < date de fin des derniers travaux
[Creation Date] < [Max Work End Date]
Choose expression elements from the lists below
Attributes: Functions, Literals, Operators
Bce: Commune, Country Code, Creation Date, Denomination
Set the threshold to be 100 % of
The test passed on 67.785% of
Save Cancel

Business Rules — Entity = Sub Contractor5(490)

Name	Description	Threshold	Result	Fail Count	Passing Fraction	Pass Count
Enregistrement à temps	Date de déclaration < date de fin des derniers travaux	100	failed	33803	67.785	71127
Présence_Identifiant	Au minimum 1 des 3 types d'identifiant doit être rempli	100	failed	3063	97.081	101867

Data Profiling – Business Rules (2)

Failing Rows (Présence_Identifiant)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
ATMI SARL				R ...	37/A		59300VA...	0	111	2009-05-20 ...	2011-10-21 00:00:0...
ENTREPRIS...				RU...	9		ANDENNE	5300	150	2005-08-18 ...	2006-06-22 00:00:0...
CONCEPT ...				AV...	15		FOREST	1190	150	2004-02-19 ...	2003-12-01 00:00:0...
ABWW SPRL				ZO...	11	n/a	MONT D...	7750	150	2009-05-20 ...	2009-03-31 00:00:0...
UTGES PR...				KL...	6A		OUDEN...	4730 AE	129	2007-09-26 ...	2007-12-31 00:00:0...
MEMOLI B...				G...	37		HASSELT	3511	150	2007-03-20 ...	2011-07-31 00:00:0...
JOSSE JON...				RU...	80		CHARLE...	6031	150	2005-12-02 ...	2006-09-28 00:00:0...
C & S CHA				BE	10		ZOUTLE	3440	150	2006-02-09	2006-07-15 00:00:0

Failing Rows (Enregistrement à temps)
Entity = Sub Contractor5(490)

Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
VAN DIJCK ...	76...			BA...	1		WUUST...	2990	150	2007-10-09 ...	2007-08-31 00:00:0...
COZIER LO...	46...	124...		RU...	n/a	n/a	LIBRAM...	6800	150	2009-05-20 ...	2003-11-28 00:00:0...
CURNET SP...	41...	127...		R ...	63	n/a	ETTERB...	1040	150	2009-05-20 ...	2008-05-31 00:00:0...
DE LENG K...	73...		78297...	TO...	7	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
VAN SCHIJ...	73...		78297...	DA...	2	n/a	HOESELT	3730	150	2007-06-07 ...	2007-06-04 00:00:0...
WOOD-B B...	46...	171...		P...	74	n/a	NIEL	2845	150	2009-05-20 ...	2006-04-14 00:00:0...
DESMIDT D...	74...		78242...	DU...	n/a	n/a	ST KATE...	2860	150	2005-06-16 ...	2005-01-31 00:00:0...
RACOMI BV	44	160		G	27	n/a	BEER	3990	150	2009-05-20	2003-05-15 00:00:0

Data Standaardisatie
Definitie

- Oplossen van gebrek aan standaardisatie
  - binnen één databank
  - overheen verschillende databanken
    - transversaal datamanagement
    - doorbreken van silo's
    - oplossen van inconsistenties in het (her-)gebruik van data-concepten
  - overheen verschillende bedrijven of instellingen
- Interne verwerkingsstap fuzzy matching
  - best practice
  - matching-resultaten worden betrouwbaarder

Data Standaardisatie

Voorbeeld: landcodes

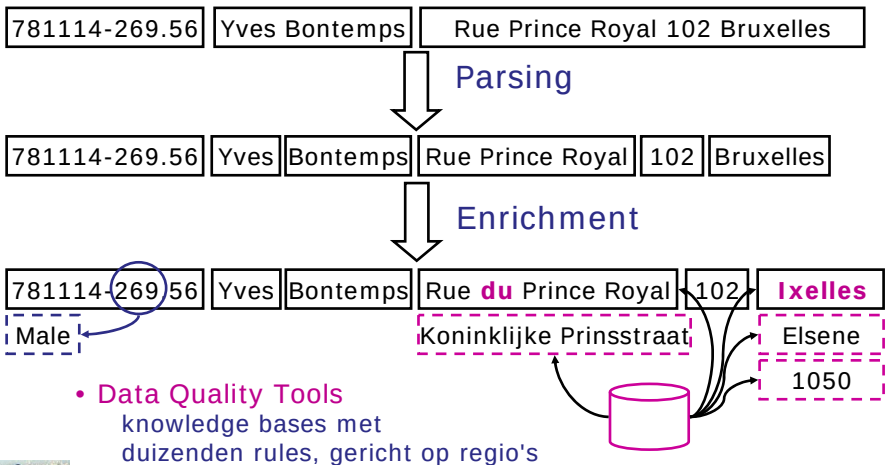
Data Rows (Dynamic)
Entity = pl tranfrmr p26(513)

origineel | toegevoegd m.b.v. data quality tool

Tsq Denom	Adres	Boite	Tsq Postcd	Tsq Commune	Orig Landcd	Iso3166 Cd	Iso3166 2l	Iso3166 3l	Iso3166 Nl	Ins Cd	Ins Nl
GREEFS ...					150	056	BE	BEL	België	150	België
JACOBS ...	MAH...			BLACKROC...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
SUIR EN...	OLD...			WATERSFO...	IRL	372	IE	IRL	Ierland		
BONCIK J...	268		067 82	MODRA NAD...	141	703	SK	SVK	Slowakije	141	Slovaakse Republiek
BONCIK J...	HIA...		067 82	MODRA NAD...	SK	703	SK	SVK	Slowakije		
BELROO...	Vizi...		1031	BUDAPEST	115	348	HU	HUN	Hongarije	115	Hongarije (Rep.)
BELROO...	VIZI...		1031	BUDAPEST	H	348	HU	HUN	Hongarije		
VOORBIJ ...	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
VOORBIJ ...	SICI...		1045	AMSTERDAM	NL	528	NL	NLD	Nederland		
DELAT KFT	MAD...		1131	BUDAPEST	H	348	HU	HUN	Hongarije		
DE WAAL...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
FUNDERI...	POS...		1440	AP PURMER...	NL	528	NL	NLD	Nederland		
SANDMA...	czer...	4	20 349	LUBLIN	122	616	PL	POL	Polen	122	Polen (Rep.)
SANDRA...	CZE...		20 349	LUBLIN	PL	616	PL	POL	Polen		
TPA EDIL ...	Via ...		24129	BERGAMO	128	380	IT	ITA	Italië	128	Italië
TPA GRO...	VIA ...		24129	BERGAMO	I	380	IT	ITA	Italië		
VLASMA...	Stee...		2407 BD	ALPHEN AA...	129	528	NL	NLD	Nederland	129	Nederland
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		
VLASMAN	STE...		2407	BD ALPHEN ...	NL	528	NL	NLD	Nederland		

Data Standaardisatie

Voorbeeld: naam, voornaam, adres, ...



- Data Quality Tools
  knowledge bases met
  duizenden rules, gericht op regio's

Data matching
Definitie

- Omgaan met schrijffouten, onnauwkeurigheden, gebrek aan standaardisatie
- om links te kunnen leggen tussen records
  - van één bron (dubbeldetectie)
  - van meerdere bronnen (dubbel- of incoherentie-detectie)
  - ook waar de datamodellen van de te vergelijken bronnen niet overeenstemmen (data integratie)
- met hoge performantie (1 miljoen/uur)
- en hoge flexibiliteit
  - definitie vaak initieel onduidelijk
  - wat mag als 'dubbel' of 'incoherent' beschouwd worden
  - link anomaliebeheer: slechts indien definities gevalideerd → formele anomalie, detectie mogelijk

Data matching
Voorbeeld op reële data: confrontatie 2 DB

Source	match type	Denomination	Adres	Boite	Postcd	Commune	Cdpays
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LITTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
R	115	PROJEKT SERWIS (LUTY WANDA)	UL BOHATEROW MODLINA 63	42	05-100	NOWY DROW MZAOWIE	PL
R	115	PROJEKT SERWIS LUTY WANDA NOWY DWOR	UL BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZOWIE	PL
R	135	PROJEKT SERWIS LUTY WANDA	BOHATEROW MODLINA 63/43		05-100	NOWY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATREROW MODLINA 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATERROW MODLINA 63/43	N/A	05-100	NOWY DWOR MAZOWIE	PL

2 databanken, L en R

- ✓ er bestaat geen 'vreemde sleutel'-relatie tussen L en R
- ✓ DQTool detecteert dubbels en organiseert in clusters
- ✓ DQTool legt link tussen beide databanken
- ✓ mogelijke fraude ?

Data matching

Voorbeeld op reële data, met interne standaardisatiestap

C Postcode	Tq Gout Postal Code	Straatnaam Voll	Tq Gout Street Name	Huisnummer	Pr House N...	Gemeentenaam	Tq Gout Postal City
1020	1020	RUE E VANDER AA	RUE ERNEST VANDER AA	1	1	Brussel	BRUSSEL
1020	1020	rue Vander Aa	RUE ERNEST VANDER AA	3	3	Bruxelles	BRUXELLES
1050	1050	91 R VAN AA	RUE VAN AA	_	91	Elsene	ELSENE
1050	1050	27 R.VAN AA	RUE VAN AA	_	27	Elsene	ELSENE
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Ixelles	IXELLES
1050	1050	RUE VAN AA	RUE VAN AA	94	94	Elsene	ELSENE
1020	1050	rue Van Aa	RUE VAN AA	2	2	Bruxelles	IXELLES
1050	1050	2 R VAN AA	RUE VAN AA		2	Ixelles	BRUXELLES
1000	1000	R JOSEPH II 40	RUE JOSEPH II		40	Bruxelles	BRUXELLES
1000	1000	rue Joseph II 71 (...	RUE JOSEPH II		71	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II	RUE JOSEPH II	71	71	Brussel	BRUSSEL
1040	1000	Rue Joseph II 5-7	RUE JOSEPH II		5-7	Bruxelles	BRUXELLES
1040	1000	Rue Joseph II 67A	RUE JOSEPH II		67A	Bruxelles	BRUXELLES
1030	1000	rue JOSEPH II, 114 -	RUE JOSEPH II	116	114 - 116	Schaarbeek	BRUXELLES

Resultaat (adresvalidatie, adrescleansing)

- ✓ gecorrigeerde postcode
- ✓ gestandaardiseerde straatnaam
- ✓ correct ingedeelde adreselementen (parsing)
- ✓ gecorrigeerde gemeentenaam
- ✓ dubbels gedetecteerd en georganiseerd in clusters

Data standaardisatie & matching

Niet alleen namen en adressen

P/N	DESCRIPTION
1774-5674	TUBE, CENTRIFUGE POLY S 15ML (CS/500)CONICAL-BOTTOM
1774-5675	TUBE, CENTRIFUGE PPL 15ML (CS/500)CONICAL-BOTTOM
1774-4532	TUBE, CENTRIFUGE PPL 50ML (CS/500)CONICAL-BTTMPCK 25/RACK
1774-4538	TUBE, CENTRIFUGE POLY S 50ML (CS/500)CONICAL-BTMPK 25/RACK
645-4556	PIPET, CLEAR SEROLOGICAL 2ML (CASE/500)
195-7934	NUT, LOCK RH,11"
3324-7955	VIAL, WHEATON 33* CLEAR 4ML (CS/144)

P/N	ITEM NAME	MATERIAL	SIZE	UOM	DESCRIPTOR	PACKAGE	PACK METHOD
1774-5674	CENTRIFUGE TUBE	POLYSTERENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-5675	CENTRIFUGE TUBE	POLYPROPILENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-4532	CENTRIFUGE TUBE	POLYPROPILENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
1774-4538	CENTRIFUGE TUBE	POLYSTERENE	50	ML	CONICAL	CASE/500	BOTTOM PACKED 25/RACK
0645-4556	SEROLOGICAL PIPET		2	ML	CLEAR	CASE/500	
0195-7934	LOCK NUT		11	IN	RIGHT HAND		
3324-7955	WHEATON VIAL		4	ML	CLEAR	CASE/144	

Data standaardisatie & matching

Link met anomaliebeheer (vb. dubbeldetectie)

Match P.	Fst Denom Lnm	Street Lnm	Street Num	City Lnm	Postcode
	inter connector zeebrugge terminal sc/cv	8th floor, aldwych	61	london	0000
402	interconnector zeebrugge terminal sc/cv	8th floor aldwych	61	wc24ae london	0000
	HAAN TECHNIEK	SAAL VAN ZWANE...	2	TILBURG	0000
402	Haan techniek □/n	Saal Van Zwanenbe...	2	Tilburg	0000
	Brixx Sp. z.o.o.	Ul. Algierska	8	Warszawa	03-977
402	BRIXX sp.z.0.0.	Ul. Algierska	8	Warszawa	03*977
110	Brixx Sp.z.o.o.	Ul Algierska	8	Warszawa	03-977
110	Brixx sp z.o.o.	Ul. Algierska	8	Warszawa	03-977
135	BRIXX SP.Z.O.O. □/nond.nr. PL - 113-2...	ul. ALGIERSKA	8	WARSZAWA	03-977
110	Brixx Sp.z.oo	ul.Algierska	8	Warszawa	03977
135	BRIXX SP ZOO □/nPL 1132760927	UL ALGIERSKA	8	WARSZAWA	03-977

Project | Data rows filtered by selected values in 'Highest Lev1matchpat' | rows: 26662

- Elk match pattern is een hulp voor de analist bij het finetunen van matching routines (Discovery)
- Iteratief overleg met business i.v.m. definitie en afhandeling van elk match pattern (Validation) → correctiescenario's
  - **110**: betrouwbaar
  - **402**: afwijking in de benaming + probleem met de postcode
  - **135**: grotere afwijking in de benaming
- Elk gevalideerd match pattern → anomalie (anomaly_type_id)

Hulp bij (anomalie-)afhandeling

Vb. Incoherentiedetectie overheen databanken

Bv.: 'commonize' telefoonnummer voor alle matched records

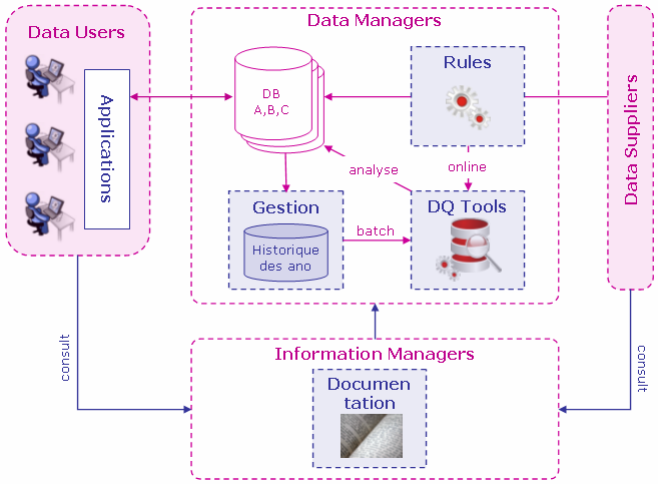
Date	First	Last	Phone	Email	Source
08/02/00	Art	Barrios	908-845-1234	bigwheels@hotmail.com	WEB
12/02/2005	A.	Barros	908-845-1234	abarrios@accen.com	CRM
6/17/2003	Arthur	Barrios	(902)-845-4417	abarrios@accen.com	SAP

Specific matching routines

Date	First Name	Last Name	Ignore Punctuation	Absolute

Table des matières

Documentation du système d'information
Importance d'une documentation



Documentation du système d'information
Fonctionnalités de base (rappel)

Gestion conjointe

- Des données structurées (échanges des messages XML entre l'administration et les citoyens et applications associées)
- Des codifications associées
- De la documentation non structurée associée aux bases de données

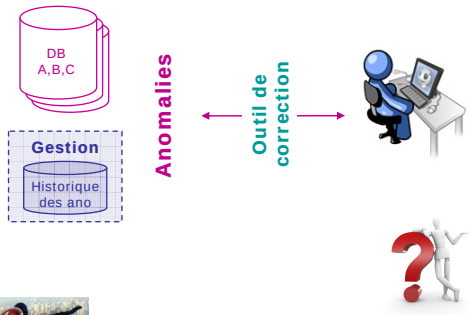
➔ impact financier et social stratégique

Exemple :
Les glossaires de la sécurité sociale
(en production depuis 2001, reengineering prévu)

- Workflow de validation
- Gestion des versions et de l'historique
- Structuration des champs multilingues
- Héritage et réutilisation (OO concept)
- WOPM (Write Once Publish Many)

Documentation du système d'information
Gestion des connaissances - Contexte

- Sur la base de consultances menées sur le terrain et d'un prototype développé à la demande du client
- Nécessité de traiter les anomalies détectées



Qu'est-ce que je corrige ?
00045-001 : Date de fin d'occupation non présent -> signification ?

Comment je corrige ?
Que dois-je faire pour résoudre l'anomalie ?

Pourquoi je corrige ?
Quel intérêt de corriger cette anomalie ?

KM System

Documentation du système d'information
Gestion des connaissances - Objectifs

1. centraliser les informations ;
2. les tenir à jour via un circuit de validation ;
3. homogénéiser le traitement des dossiers ;
4. partager des connaissances sur la correction des anomalies ;
5. disponibilité équivalente dans les deux langues ;
6. humaniser le travail des agents
   - démotivation partielle car ils ne perçoivent pas quel est l'impact et l'utilité de leur correction

Documentation du système d'information
Gestion des connaissances - Besoins

Besoins

Informationnels

- informations descriptives
  - identification de l'anomalie
  - impacts de l'anomalie
  - liens vers sources externes (?)
- informations de traitement
  - description de l'anomalie
  - causes possibles
  - étapes pour résolution

Fonctionnels

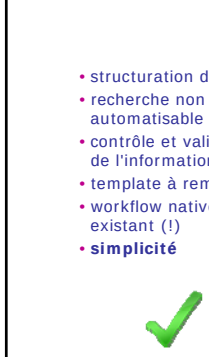
- interface d'édition
- workflow de validation
- gestion du multilinguisme
- moteur de recherche
- gestion des versions
- feedback binaire

Google

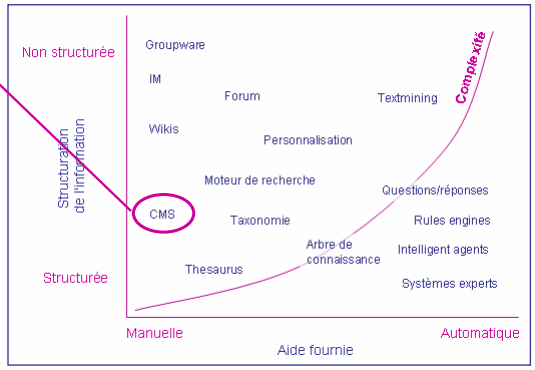
Ces informations vous ont-elles aidé ? ○ Oui ○ Non

Documentation du système d'information
Gestion des connaissances - Aspects technologiques

- structuration de l'info
- recherche non automatisable
- contrôle et validation de l'information
- template à remplir
- workflow nativement existant (!)
- **simplicité**

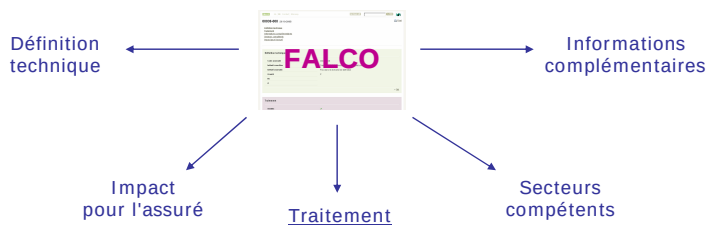


Knowledge Management System Matrix



Documentation du système d'information
Gestion des connaissances – Exemple

- Falco est un prototype de système de gestion des connaissances
- Documentation réalisée par et pour le service du contrôle (ONSS)



Smals
ICT for society

28/10/2010

Docume
Gestion d

Fiche anomalie
en édition

Définition technique

Code anomalie * 99999-999 Format : 99999-999

Intitulé zone/bloc (fr) intitulé zone FR

Intitulé anomalie (fr) intitulé anomalie FR

Gravité NP

Trimestre de validité (De) Format : 2004-1

Trimestre de validité (A) Format : 2007-3

Traitement

A résoudre

Soluble

Contact employeur

Description anomalie (fr) Start Edition

description de l'anomalie

Smals
ICT for society

Docume
Gestion de

Fiche anomalie
en consultation

FALCO NL | FR Contact | Sitemap Archive off

00045-001 15/01/2009 Print

Définition technique
Traitement
Informations complémentaires
Secteurs compétents
Impact pour l'assuré

Définition technique

Code anomalie	00045-001
Intitulé zone/bloc	Date de fin de l'occupation
Intitulé anomalie	Non présent
Gravité	P
De	
A	

« TOP

Traitement

Soluble	✔
A résoudre	✔
Contact employeur	✔

Description anomalie

Cette anomalie est signalée lorsque la date de fin de contrat n'est pas présente.

Causes possibles

Il y a une ligne rémunération avec une indemnité de rupture (code rémunération 3). Une date de fin de contrat doit toujours être indiquée.

Que faire ?

Prendre contact avec l'employeur ou son mandataire afin de déterminer une date de sortie correcte .
OU :
Sur base des jours déclarés et du régime de travail, déterminer soi-même la date de sortie.

Documentation du système d'information
Gestion des connaissances (7) - Recommandations

- Simplicité
  - Mise à jour de l'info par les gens du business
- Utilisateurs !
- Indiquer l'utilité du traitement (si possible)
- Guidelines de rédaction pour champ texte libre
  - homogénéité visuelle entre les deux langues
- Veiller de manière continue à la qualité du contenu

Documentation du système d'information
Organisation

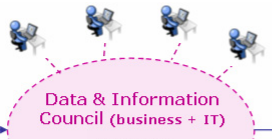


Table des matières

Conclusion

- Continuité des recommandations du DQ Competency Center (consultances et études)
- Quelques enseignements novateurs et opérationnels de l'étude :
  - Modélisation générique de l'historique des anomalies et stratégies de gestion
    - méthode conceptuelle originale
    - exemple de maquette opérationnelle
    - en fonction du contexte :
      - à compléter par un travail d'analyse spécifique
      - prendre en compte l'impact fonctionnel et organisationnel (ressources disponibles, enjeux …)

Conclusion

- Quelques enseignements novateurs et opérationnels de l'étude (suite)
  - Apports des « data quality tools »
    - software spécialisé
    - en batch : aide à l'analyse, à la détection et à la correction
    - on-line : aide à la détection et à la correction
  - Documentation des anomalies (Falco)
    - gestion des connaissances associée
  - Organisation globale composée de modules distincts et interagissants
  - Généralisable à toute base de données empirique

Conclusion

Questions

Isabelle Boydens
isabelle.boydens@smals.be
02.787.59.92

Marc Dessart
marc.dessart@smals.be
02.787.58.34

Arnaud Hulstaert
arnaud.hulstaert@smals.be
02.787.51.91

Dries Van Dromme
dries.vandromme@smals.be
02.787.55.11