

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



- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
-  • Pause et questions
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
-  • Conclusion et questions





Table des matières

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion

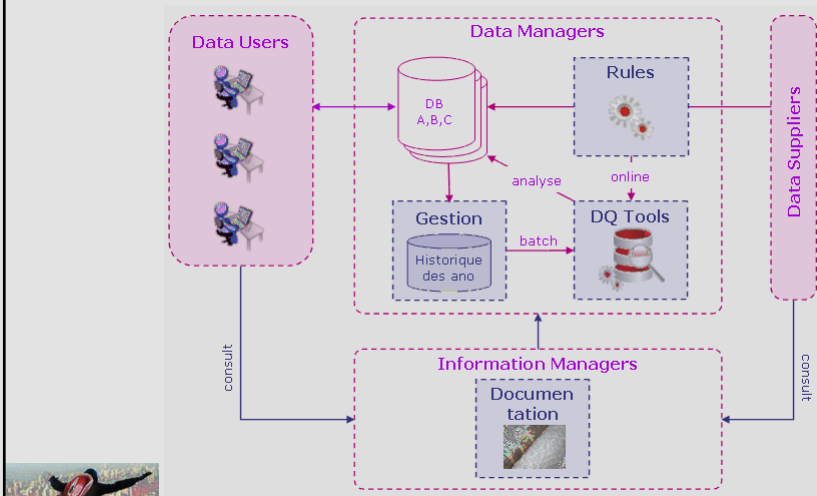


Contexte de l'étude

- 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 »



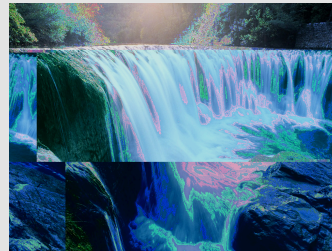
Contexte de l'étude



5

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)



6

Contexte de l'étude

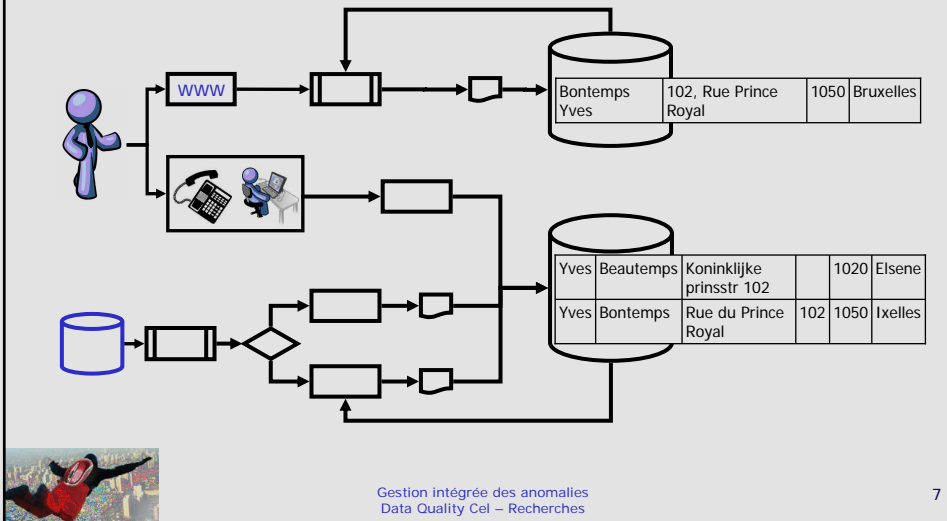


Table des matières

- Contexte de l'étude
- **Définitions**
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion

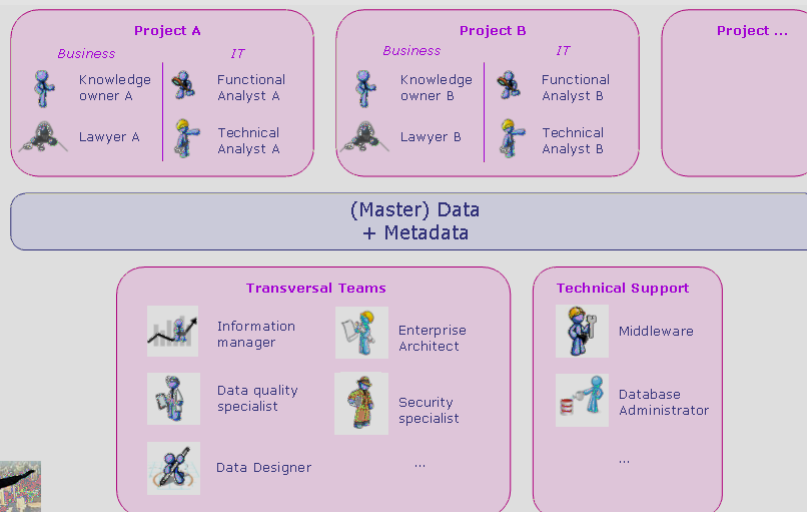


Définitions (rappels) Qualité d'une base de données

- « 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



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 » ?

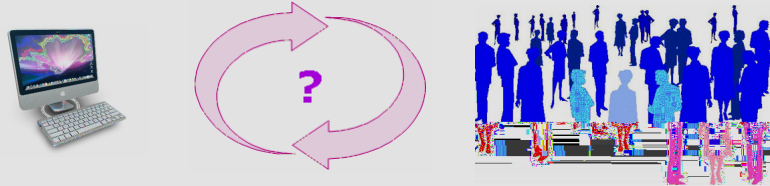
- Typologie des violations de contrainte d'intégrité
 - Erreur formelle
 - Présomption formelle d'erreur : anomalie
 - Erreur indétectable formellement



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 ?

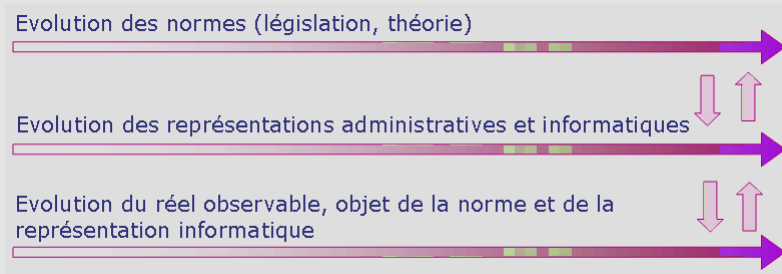




Table des matières

- Contexte de l'étude
- Définitions
- **Modélisation conceptuelle de l'historique des anomalies**
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion



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

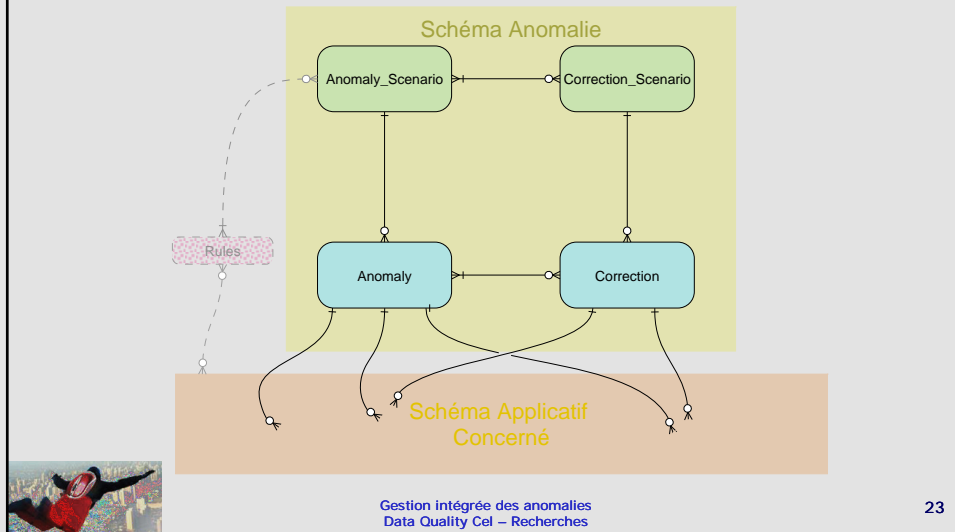
- L'implémentation doit pouvoir trouver sa place, de la même manière, dans **des DB existantes** comme au sein **des nouvelles architectures**
- La démarche doit pouvoir supporter la gestion **automatique** comme **manuelle**
- Le **couplage** entre l'application et la gestion des anomalies doit être **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



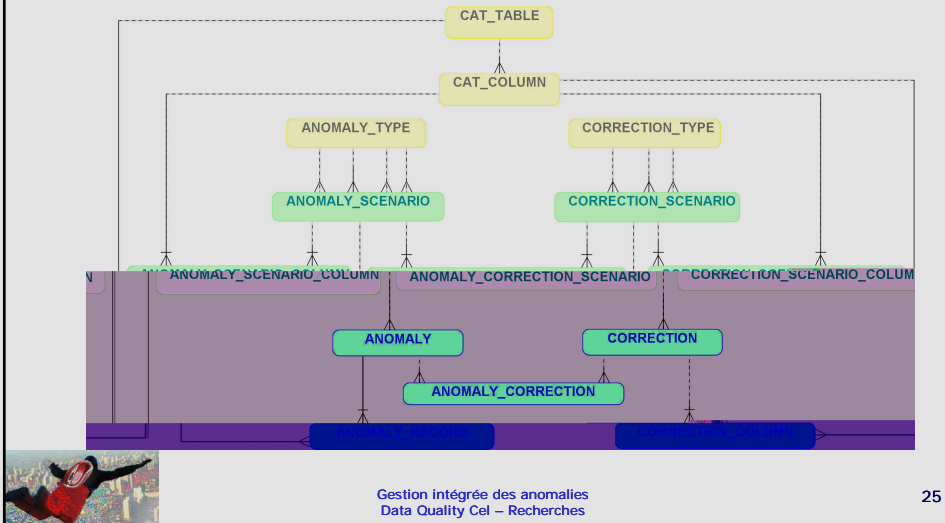
Modélisation de l'historique des cas d'anomalies Structuration en trois couches

- 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



25

Modélisation de l'historique des cas d'anomalies

Déroulement pour implémenter la démarche

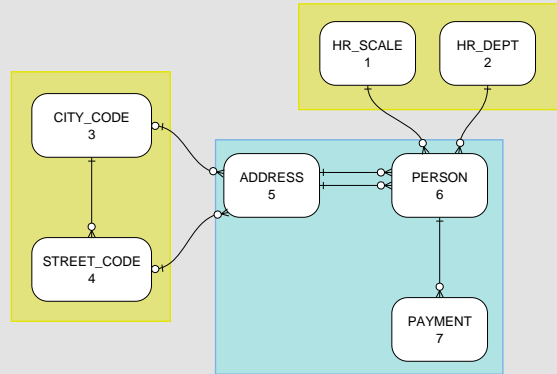
- Remplir le dictionnaire technique et business :
 - Le **catalogue** avec les tables et les colonnes concernées du schéma applicatif
 - Les **types** d'anomalie et de correction, dont la spécification et la granularité dépendent de la base de données originelle
 - La description des **scénarios** d'anomalies 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



26

Modélisation de l'historique des cas d'anomalies

Exemple : une base de données applicative



Exemple : Phase de détection des cas d'anomalie

PERSON (6) Schéma Applicatif

PK	CREATED_TMS	END_TMS	yyy	...	xxx	...	RECORD_ANOMALY_ID
536 814	Tms 1	NULL	yyy	...	xxx	...	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

PERSON (6) Schéma Applicatif

PK	CREATED_TMS	END_TMS	...	xxx (8)	...	RECORD_ANOMALY_ID
xxx	Tms 1	Tms 3	...	ααα	...	2564
xxx	Tms 3	NULL	...	0000	...	2564

ANOMALY_ID counter: 1433
 RECORD_ANOMALY_ID counter: 2564
 CORRECTION_ID counter: 870

ANOMALY_ID	CREATED_TMS	ANOMALY_SCENARIO_ID	RESOLVED_IND
1433	Tms 2	A101-0610	Y

Anomalie correction

CORRECTION_ID	CREATED_TMS	CORRECTION_SCENARIO_ID	ANOMALY_ID
870	Tms 3	C101-0610	1433

CORRECTION_ID	RECORD_ANOMALY_ID	COLUMN_ID	ORIGINAL_TMS
870	2564	8	Tms 1

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
 - Voir 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

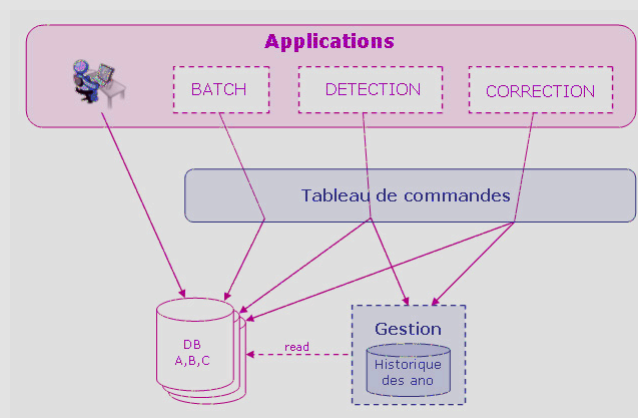
Seuls les programmes de gestion des anomalies ont les droits de modification sur les RECORD_ANOMALIE_ID applicatifs

Modélisation de l'historique des cas d'anomalies Démonstration



The slide features a large, stylized graphic of the word 'Maquette' in 3D yellow and orange letters, arching over a golden 3D stick figure holding a laptop. Below this, the word 'MOHICAN' is written in a grey, textured font. In the bottom left corner, there is a small image of a person in a red and black suit flying through a cityscape.

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



Détails pour le NISS : 19650428662

CREATED	ENDED	NAME	FIRSTNAME	BIRTH	RECRUTE	SALARY	RECANO
11/10/2010 17h19.23		Moens	Nicolas	28/04/1965	01/01/2010	5731	2265
24/04/2010 12h36.28	11/10/2010 17h19.23	Moens	Nicolas	28/04/1965	01/01/2010	5731	--

MONTH	CREATED	ENDED	ACCOUNT	AMOUNT	RECANO
07.2010	26/04/2010 09h38.15	--	537-9396902-02	3175.00 €	--
06.2010	26/04/2010 09h38.15	--	537-9396902-02	3171.00 €	--
05.2010	26/04/2010 09h38.15	--	537-9396902-02	3167.00 €	--
04.2010	26/04/2010 09h38.15	--	537-9396902-02	3163.00 €	--
03.2010	26/04/2010 09h38.15	--	537-9396902-02	3159.00 €	--
02.2010	26/04/2010 09h38.15	--	537-9396902-02	3155.00 €	--
01.2010	26/04/2010 09h38.15	--	537-9396902-02	3152.00 €	--

Liste des employés

NISS	CREATED	ENDED	NAME	FIRSTNAME	BIRTH	RECRUTE	SALARY	RECANO	NP
19600314966	24/04/2010 12h36.28	--	Urban	Kody	14/03/1960	01/01/1995	6161 €	--	--
19600311922	24/04/2010 12h36.28	--	El Amrani	Jules	17/03/1960	01/01/1994	2650 €	--	--
19600402280	24/04/2010 12h36.28	--	Beckers	Corrado	03/04/1960	01/01/1992	5731 €	--	80
19600409940	11/10/2010 17h19.23	--	Willems	Marcus	09/04/1960	01/01/2001	5376 €	2489	--
19600411728	11/10/2010 17h19.23	--	Hajji	Gillaume	11/04/1960	11/04/1970	2534 €	2032	--
19600415264	24/04/2010 12h36.28	--	Delhaye	Nicolas	15/04/1960	01/01/1982	5376 €	--	--
19600512222	24/04/2010 12h36.28	--	Radari	Enzo	12/03/1960	01/01/1985	4509 €	--	--
19600520584	11/10/2010 17h19.23	--	Hardy	Johann	05/05/1965	03/05/1975	4649 €	2872	--
19600520532	24/04/2010 12h36.28	--	Piotte	Anna	10/05/1960	03/03/1986	3638 €	--	--

Tableau de commandes

228 Corrections réalisée !

Visualisation

Update

Detection

Valeur de Birth_Date dans PERSON

Incompatibilité entre Recruiting_Date et Birth_Date

Per.Sal-HR

Per.Sal - Pay.Amt

Correction

C201-0607 <-> Valeur de Birth_Date dans PERSON

C201-0610 <-> Incompatibilité entre Recruiting_Date et Birth_Date

Initialisation

Numéro d'Anomalie courant : 2120

Numéro de Record Anomalie courant : 3252

Numéro de Correction courant : 1288

Liste des Scénarios Anomalies

Count	Reference	Type	Name	Description
8	A101-0607	Valeur hors domaine	Valeur de Birth_Date dans PERSON	Valorisation du champ BIRTH_DATE (colonne 31 = 607) dans la table PERSON (Table_Nr = 6) - Les employés doivent avoir plus de 16 ans et moins de 65 ans
287	A107-0610	Erreur de contrôle de contenu	Incompatibilité entre Recruiting_Date et Birth_Date	La date de recrutement n'est pas dans le domaine de valeur par rapport à la date de naissance - Les employés ne peuvent pas être recrutés avant leur 16 ans ni après leur 65 ans
0	A110-1002	Erreur de contrôle de contenu	Par.Sal-HR	
1	A110-2001	Erreur de contrôle de contenu	Per.Sal - Pay.Amt	

Liste des Anomalies liées au RECANO : 2679

Ref.	Created	Scenario	Solved
1546	11/10/2010 17h15.49	Valeur de Birth_Date dans PERSON	1010
2679	19601009024 - Vidéom - Sarah	PERSON 19601009024 - Vidéom - Sarah	
1096	11/10/2010 17h15.53	Incompatibilité entre Recruiting_Date et Birth_Date	No



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

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- **Monitoring des anomalies et stratégies de gestion**
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion



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)



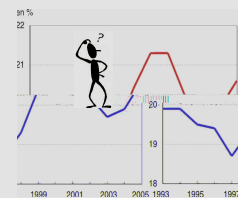
Stratégies de gestion (case studies) Suivi et amélioration de la structure de la DB (2)

- 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, ...)



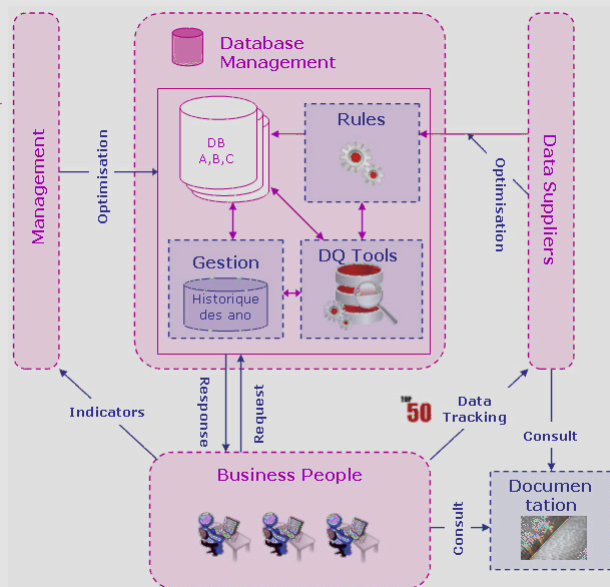
Stratégies de gestion (case studies) Data tracking

- 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

**TOP
50**



Aspects organisationnels



Pause

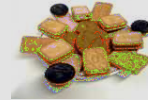


Table des matières

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- **Modélisation de la séquence de contrôles**
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion

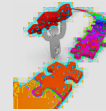


Modélisation de la séquence des contrôles

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

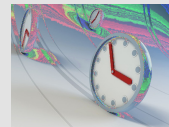
- **Dépendance** entre données

- Contrainte **d'intégrité** référentielle
- Dépendance **fonctionnelle**
- Dépendance/règle **métier**



- **Rythmes de mise à jour**

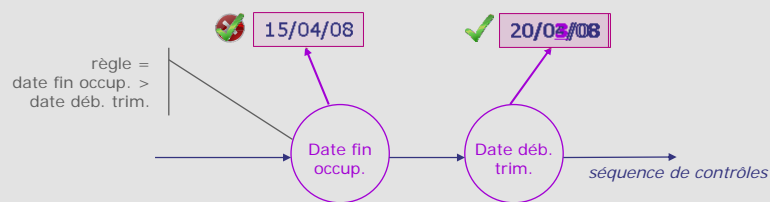
- risque d'anomalies si les sources sont pas mises à jour en même temps



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



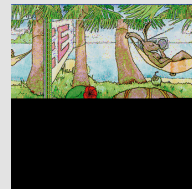
Modélisation de la séquence des contrôles Conflits potentiels et arbitrage (3)

- 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

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion

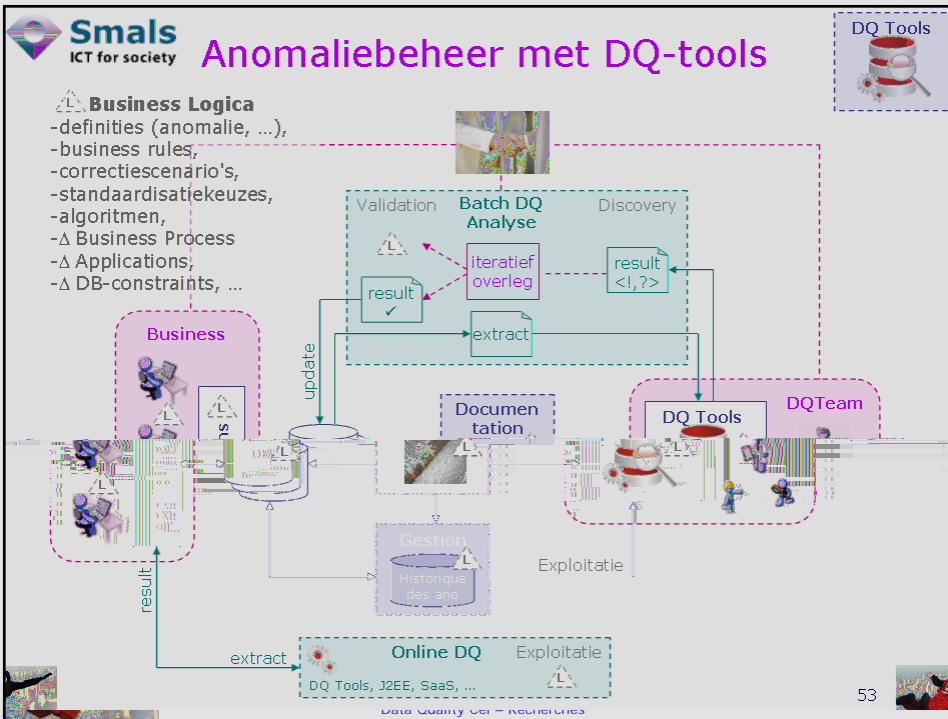


Data Quality met Data Quality Tools Constatatie

- Overall 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 ▪ configureren ipv ad hoc programmeren
 - Productiviteit ▪ data matching: 1 miljoen records/uur
 - Flexibiliteit ▪ GUI/visualisatie/drill-down/export
 - vorm ideaal voor iteratief overleg



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



Smals ICT for society Data Profiling – Column Property Analyse, drill-down (1) 28/10/2010

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

Gestion intégrée des anomalies
Data Quality Cel – Recherches 57

Smals ICT for society Data Profiling – Column Property Analyse, drill-down (2) 28/10/2010

Unique Masks

Mask	Mask Pattern	Value Count	Frequency	Dist %
NNNN	N4	1154	239983	97.358
N	N1	1	6513	2.642

Unique Values

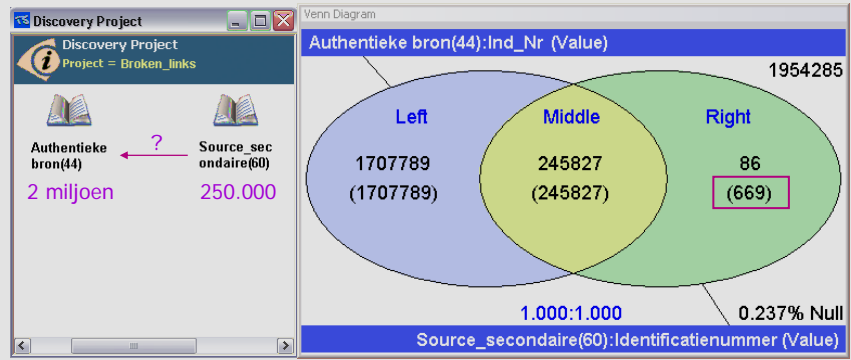
Value	Frequency	Dist %	Length	Soundex	Metaphor
0	6513	2.642	1		

Data rows filtered by selected values in 'Apostemp'

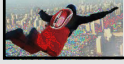
Row	Aadremp	Apostemp	Acomemp
51	Z ART COMMERC LANGWIES	0	JUNGLINSTER LU
77	STEPHENSONSTRAAT 5	0	JA TIEL NL
90	TORENLAAN 10	0	2215RW VOORHOUT NL
137	HAGENWEG 5A	0	4131 LX VIANEN NL
147	BOEKENDERWEG 18	0	THORN NL
156	OVERGOO 2 C	0	LEIDSCHEMAM NL
206	ALLEE DE LONGCHAMP 2	0	92150 SURESNES FR
220	VAN KOERSVELDWEG 1	0	7665SG ALBERGEN NL
296	PUERTA DEL SOL 7	0	MADRID ES
302	RUE CALMETTE 47	0	VENDIN LE VIEIL FR
335	RUDONK 4	0	BREDA NL
481	RUE DU STADE	0	SAINT FULGENT FR

Gestion intégrée des anomalies
Data Quality Cel – Recherches 58

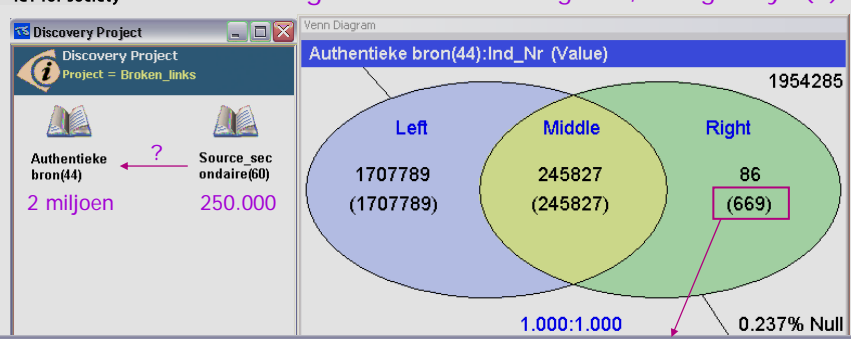
Data Profiling – Referentiële integriteit, foreign keys (1)



- 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)



Data Profiling – Referentiële integriteit, foreign keys (2)



Non-Joining Right-Hand Rows

Join = Authentieke bron(44):Ind_nr > < Source_secondaire(60):Identificatienummer

Row	Reeksnum...	Identificat...	Alytemp	Anajur	Adatalfi	Adatsup	Adenomemp	Adressamp	Apostemp	Acomemp
192648	45018	01032	161		1050701		KERK-FABRIEK 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
FRDAM NL	126830	45666	08998	121	1038001		BV BVBA MICHELSEN & WITTENS GEASS NOT...	HANDELSLEI 102	2980	ZOERSEL
	239099	49646	08997	111	1080609		BAYON ISABELLE	HOOIENDONKSTRAAT 52	2801	MECHELEN
	246142	50537	08996	121	1080616		DAUWEN MARC, LIPSCHUTZ LAURA, DRAULANS ...	GEVAERTLAAN 180	2260	WESTERLO
	246146	51428	08993	151	1080519		DEMFOOD BV	VLAARDINGWEG 51	0	3044 CJ ROT
	241041	57168	08994	131	21	1080707	GEMA BOUW BVBA	HOEVESTRAAT 33 B	1755	GOOIK
	66445	08990	62118	161	1080401		KERK-FABRIEK 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	7380	QUAREGNO
	239321	76964	18305	111	1080801		VOGELS ROEL	SCHEPEN DEJONGHSTRA...	3800	ST TRUIDEN
HTEGEM		51780	77459	17132	111	1080701	FERLUN JAN	AARTRUKSTRAAT 15	8480	

Smals ICT for society 28/10/2010

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 21	Straatnaam 21	Discovered	No	60	99.110	9911	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 28/10/2010

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 Loaded

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 NOT (id_1 == "" AND id_2 == "" AND id_3 == "")

Au minimum 1 des 3 types d'identifiant doit être rempli

Entity Business Rule

Enabled Name [Creation Date] < [Max Work End Date]

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

Gestion intégrée des anomalies
Data Quality Cel – Recherches

Failing Rows [Présence Identifiant]											
Entity = Sub Contractor(5490)											
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 & SCHA				BF	10		ZOLITE	3440	150	2006-02-08 ...	2006-07-15 00:00:0...

Failing Rows [Enregistrement à temps]											
Entity = Sub Contractor(5490)											
Denomination	Id_1	Id_2	Id_3	Rue	Rue Num	Rue Box	Commune	Postcode	Country...	Creation Date	Max Work End Date
VAN DUICK ...	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 SCHU...	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...	2660	150	2005-06-16 ...	2005-01-31 00:00:0...
BACOMBY	44	160		G...	27	n/a	DEEP	3000	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

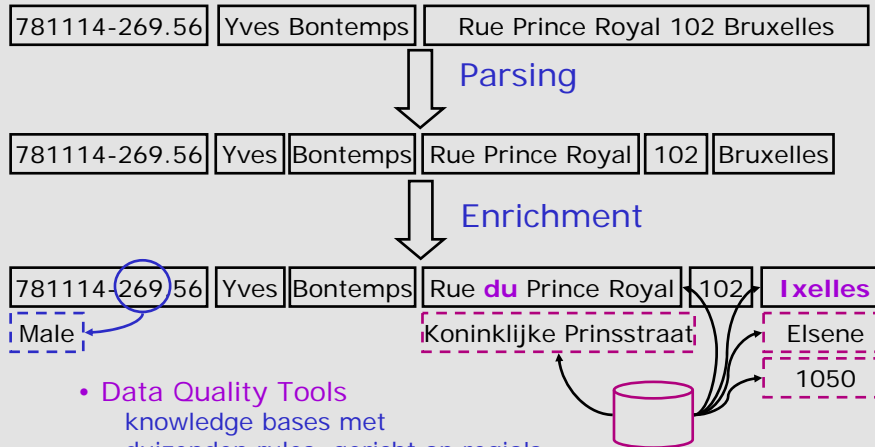
Data Rows (Dynamic)
Entity = pl_branfmr_p26(913)

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 NI	Ins Cd	Ins NI
GREEFS ...				BLACKROC...	150	056	BE	BEL	België	150	België
JACOBS ...	MAH...			WATERSFO...	116	372	IE	IRL	Ierland	116	Ierland /Eire/
SUIR EN...	OLD...			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		
VOORBU ...	SICI...		1045 AX	AMSTERDAM	129	528	NL	NLD	Nederland	129	Nederland
VOORBU ...	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 LUTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
L	100	PROJEKTSERWIS WANDA LUTY	BOHATEROW MODLINA 63	43	05-100	NOWY DWOR MAZ	122
R	115	PROJEKT SERWIS (LUTY WANDA)	UL BOHATEROW MODLINA 63	42	05-100	NOWY DWOR MAZOWIE	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	BOHATEROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	106	PROJEKT SERWIS WANDA LUTY	BOHATEROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATEROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATEROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	128	PROJEKT SERWIS WANDA LUTY	BOHATEROW 63LOK	43	05-100	NOKY DWOR MAZ	PL
R	138	SOCIETE PROJEKTSERWIS	BOHATEROW 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	POLYPROPYLENE	15	ML	CONICAL	CASE/500	BOTTOM PACKED
1774-4532	CENTRIFUGE TUBE	POLYPROPYLENE	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	



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 Q/n	Saal Van Zwanenbe...	2	Tilburg	0000
	Brixx Sp. z. o. o.	Ul. Algierska	8	Warszawa	03-977
402	BRIX sp. z.O.O.	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	BRIX SP.Z.O.O. Q/nond.nr. PL - 113-2...	ul. ALGIERSKA	8	WARSZAWA	03-977
110	Brixx Sp. z. oo	ul. Algierska	8	Warszawa	03977
135	BRIX SP ZOO Q/nPL 1132760927	UL ALGIERSKA	8	WARSZAWA	03-977

Project Data rows filtered by selected values in 'Highest Lev1matchpal' 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)
 - 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)

correctiescenario's



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
------	------------	-----------	--------------------	----------



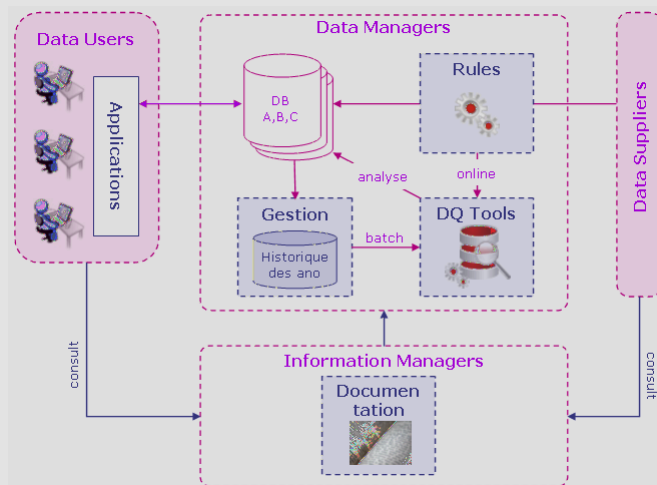


Table des matières

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion



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

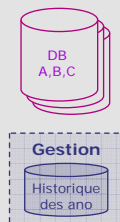
Exemple :
Les glossaires de la
sécurité sociale
(en production depuis 2001,
reengineering prévu)

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



Anomalies

Outil de
correction



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

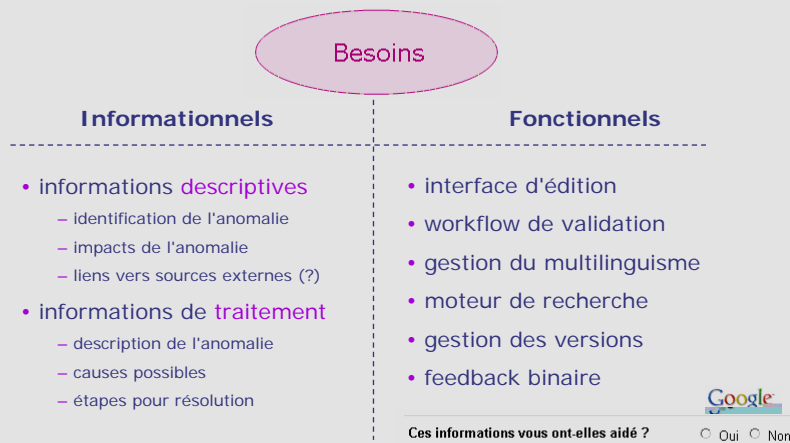


Documentation du système d'information Gestion des connaissances - Objectifs

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

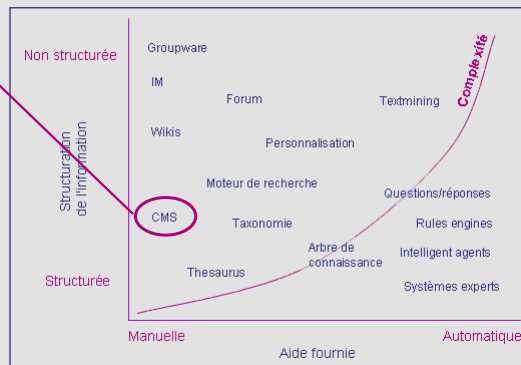


Documentation du système d'information Gestion des connaissances - Aspects technologiques

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

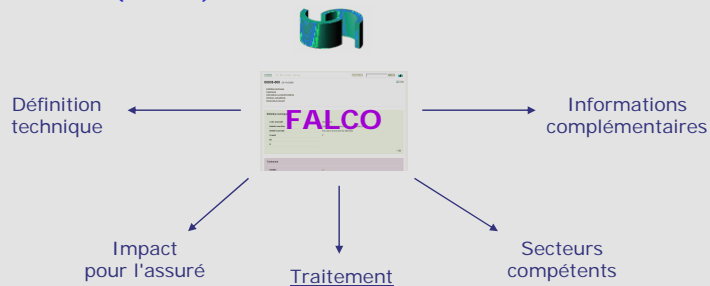


Knowledge Management System Matrix



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)



Document
Gestion de

Fiche anomalie
en édition



Définition technique

Code anomalie : 99999-999 Format: 99999-134

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

1

Document
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 complémentaires](#)
[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.

2



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

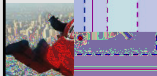
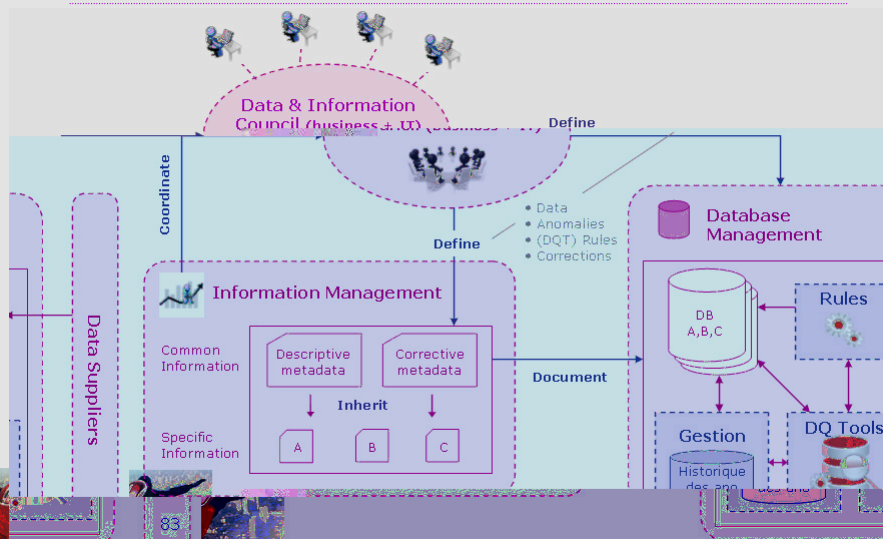




Table des matières

- Contexte de l'étude
- Définitions
- Modélisation conceptuelle de l'historique des anomalies
- Monitoring des anomalies et stratégies de gestion
- Pause
- Modélisation de la séquence de contrôles
- Data Quality Tools – retour d'expérience
- Documentation opérationnelle du système d'information
- Conclusion



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 ...)

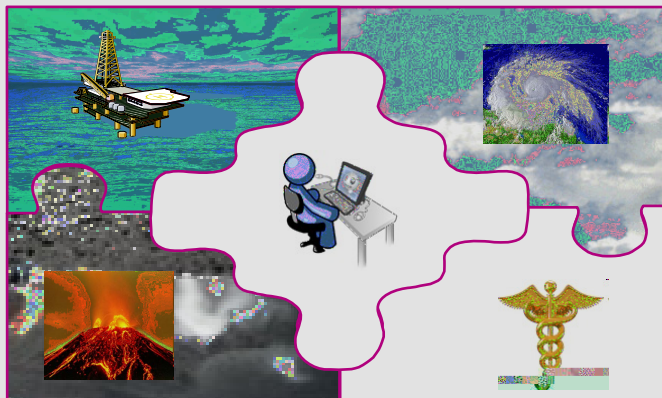


Conclusion

- 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

