amue

séminaire

Améliorer la qualité des données

Bernard GAILLOT
Consultant QUADEM

Sophie ANDRIAM
Consultante domaine RH
DSER AMUE

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Séminaire

1 – Présentation de la société QUADEM

2 - Données et indicateurs

3 – Leviers pour améliorer la qualité des données

4 - Paramètres de qualité des données

5 - Qualité des données : état de l'art

6 - Conclusion

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Présentation de la société QUADEM

un métier : la qualité des données

- Apporter aux entreprises et aux organisations du conseil, des démarches, des outils pour la production de données ;
- Appliquer ces démarches et ses outils sur des projets d'initialisation, migration et fiabilisation de données ;
- Société créée en 1999 ;
- Clients dans tous les secteurs : industrie, télécommunications, banques, collectivités, organisations internationales, etc…

Quadem CTS - 17 avenue Charles de Gaulle

69370 Saint-Didier-au-Mont-d'Or

Tél. : 33 (0)4 78 64 34 79 - Fax : 33 (0)4 78 64 31 22

bgt@quadem.fr

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Données et indicateurs :** caractéristiques marquantes

Les données représentent le monde réel sous forme numérique pour des opérations plus ou moins complexes.

- Pas de traitement : donnée uniquement hébergée et véhiculée par le SI
- Peu complexe : sélection, tri, agrégation, cumul, moyenne

(niveaux croissants de complexité)

- Très complexe : donnée utilisée dans des traitements algorithmiques

==> *la non Qualité des Données frappe plus facilement les données utilisées pour des opérations peu complexes*

==> *la "promotion" des données pour l'utilisation dans des traitements plus complexes est rarement gérée et est source de dysfonctionnements*

Améliorer la qualité des données
…selon la société Quadem

Data Quality way

- **Données et indicateurs :** caractéristiques marquantes

Les données sont produites à partir de sources de données :

- Le plus souvent, les données ne sont pas créées à partir du monde réel, mais à partir de **Sources de substitution**, plus faciles à exploiter.
- Pour certaines utilisations, une source de données dûment qualifiée peut être labellisée « **réputée fiable » (ou certifiée)**.
  - *exemple* : pour les données d'identité

Améliorer la qualité des données

…selon la société Quadem

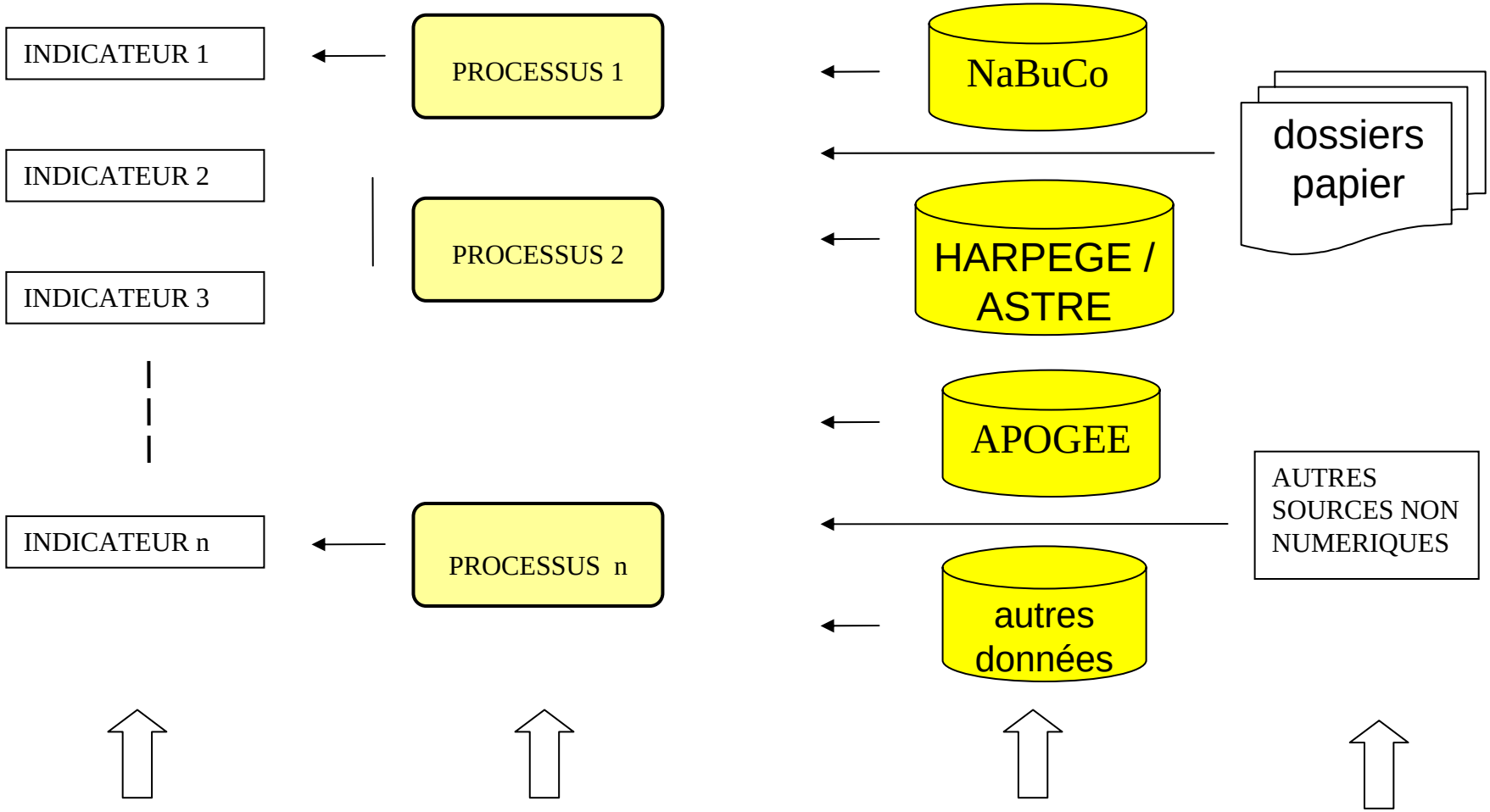
Data Quality way

- **Données et indicateurs :** production des indicateurs
  - schéma

séminaire

Améliorer la qualité des données

…selon la société



Séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Données et indicateurs**
  - **INDICATEUR = PRODUIT FINI**
    - Les indicateurs répondent à un cahier des charges et doivent satisfaire à certains critères, en particulier en matière de qualité.
  - **Processus de production**
    - La production fait l'objet de processus, ses caractéristiques principales :
      - production par le ministère / par l'établissement
      - production automatique / semi automatique / manuelle
      - processus dédié à un seul indicateur ou à plusieurs
    - L'établissement doit formaliser les processus non automatiques, et matérialiser des règles dans une approche **industrialisée**.

séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Données et indicateurs

Sources de données informatiques

- Les établissements en sont responsables ; c'est sur les données que portera l'essentiel du travail de fiabilisation.

Autres sources de données

- Les établissements en sont responsables ; elles doivent être gérées dans une relation client / fournisseur (interne ou externe) et fiabilisées.

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- Leviers pour améliorer la qualité des données
  - **Projet de fiabilisation des données**
    - Ce type d'action a pour but de traiter les anomalies résultant du passé.
  - **Amélioration des processus existants**
    - Ce type d'action a pour but de garantir le futur en améliorant les mécanismes ; **ces actions constituent le meilleur investissement et doivent être lancées en priorité.**
  - **Actions sur l'organisation**
    - Ce sont les actions de sensibilisation, formation, définition des responsabilités.

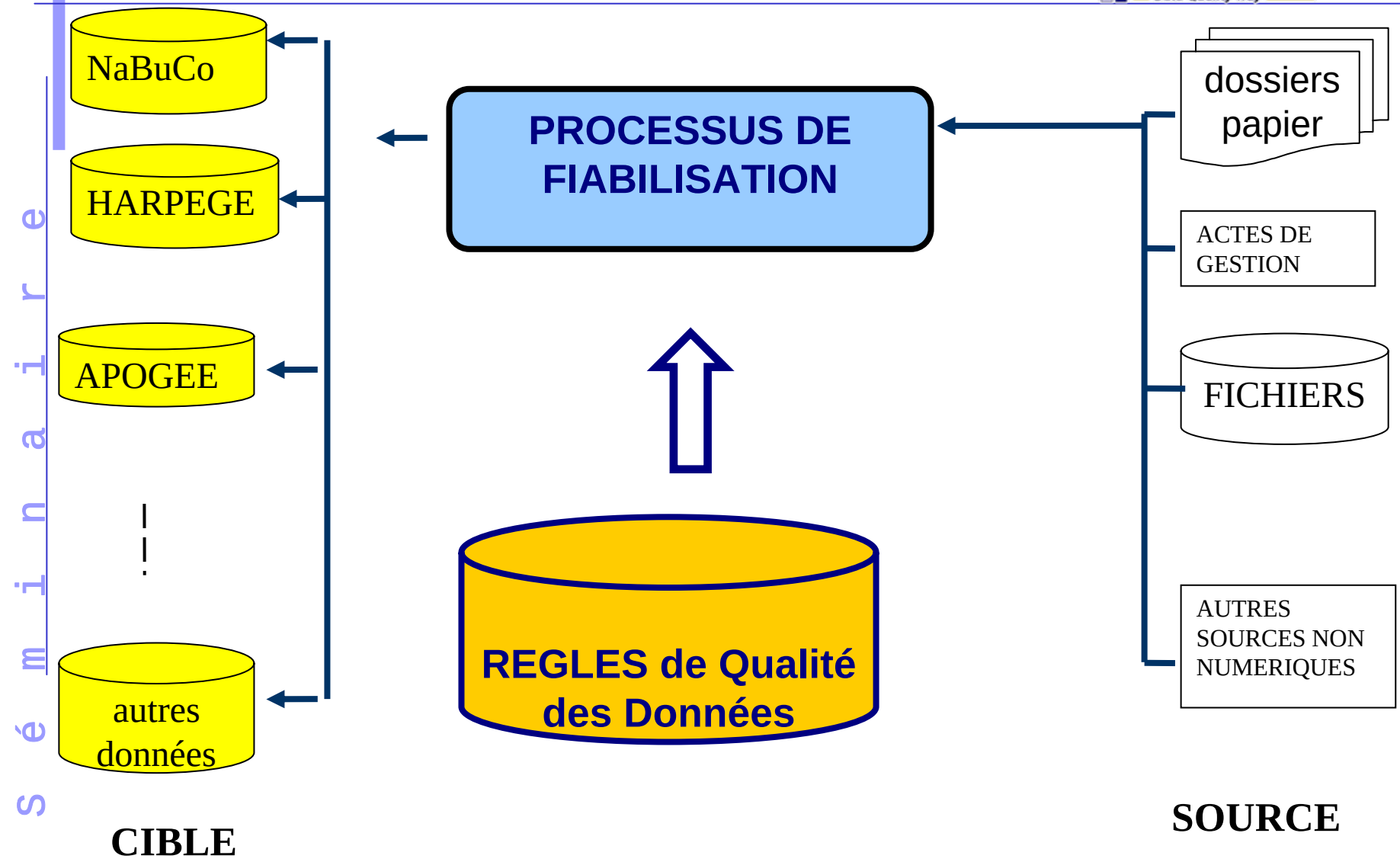
Améliorer la qualité des données

…selon la société

- Leviers pour améliorer la qualité des données
  - Projet de fiabilisation des données
    - Schéma

Améliorer la qualité des données

…selon la société



Améliorer la qualité des données

…selon la société

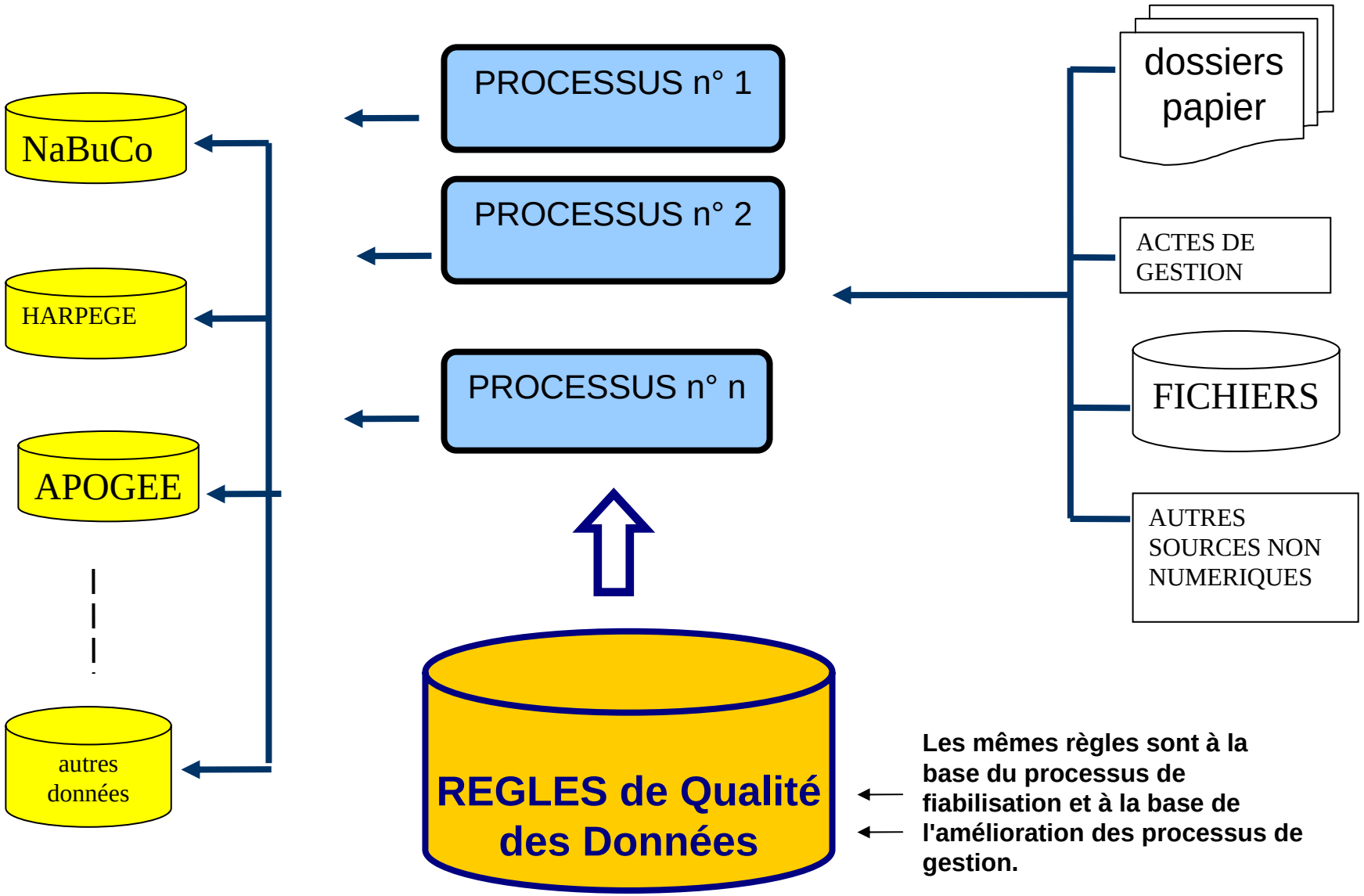
- Leviers pour améliorer la qualité des données
  - Amélioration des processus existants
    - Schéma

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Séminaire



Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - **Paramètres de modélisation (DOMAINE INFORMATIQUE)**
    - *2 exemples* **: cohérence de la modélisation ; composition** (source : Tom Redman)

Modélisation n°1

CODE	Libellé	Cohérence de la modélisation	Composition
IDENT	Nom et prénom		**NOK**
DAT_NAIS	Date de naiss. JJ/MM/AAAA		OK
DAT_BAC	Date BAC JJ/MM/AAAA	**NOK**	OK
AGE_BAC	Âge à l'obtention du BAC	**NOK**	

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Paramètres de qualité des données

Modélisation n°2

CODE	Libellé	Cohérence de la modélisation	Composition
NOM	Nom		OK
PREN	Prénom		OK
DAT_NAIS	Date de naiss. JJ/MM/AAAA		OK
DAT_BAC	Date BAC JJ/MM/AAAA		OK

Améliorer la qualité des données

…selon la société

- **Paramètres de qualité des données**
  - **Paramètres de présentation (DOMAINE INFORMATIQUE)**
    - 2 *exemples* **: interprétabilité ; précision**

    (source : Tom Redman)

séminaire

Paramètres de qualité des données

Présentation n°1

identité, origine ……..

Appréciation du séminaire (cocher) (*) : **1**
2
3

(*) du moins bon au meilleur

Présentation n°2

nom : …… …….
prénom ……
établissement ……
Appréciation du séminaire (cocher) : ☐ **bof**
☐ **bon**
☐ **excellent**

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - **Paramètres liés aux valeurs (DOMAINE UTILISATEUR)**

    (source : Quadem par simplification de la littérature existante)

    Exactitude (*)

    Complétude des valeurs

    Complétude des occurrences

    Non-duplication

    Actualité

    Cohérence

(*) la précision peut être vue comme paramètre annexe de l'exactitude

séminaire

Améliorer la qualité des données

…selon la société Quadem

- **Paramètres de qualité des données**
  - **Exactitude**
    - Ce paramètre s'applique aux attributs

      (un attribut = un champ)

Jean-Pierre MARTIN informatique

CODE	Libellé	Valeur
NOM	nom	MARTIN
PREN	prénom	Jean-Pierre
DAT_NAIS	date de naissance	**30/06/1983**
NATIONALITE	nationalité	F
DAT_BAC	date d'obtention BAC	22/06/2002

Jean-Pierre MARTIN réel

REPUBLIQUE FRANCAISE

Nom : **MARTIN**
Prénom(s) : **Jean-Pierre**
Né(e) le : **30/05/1983**
Nationalité : **Français**

Attention : Dossier d'inscription = source de substitution

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - **Complétude des valeurs**
    - Ce paramètre s'applique aux attributs

N°	Attribut	Etudiant 1	Etudiant 2	Etudiant 3
001	Nom	Martin	Durand	Dupont
002	Prénom	Jean-Claude	Jean-Pierre	J……
003	Date naissance	26/02/1984	30/04/1985	25/12/1984
004	Pays naissance	France	……	……
005	Ville naissance	Paris	Lyon	……
006	Nationalité	F	F	……
007				
008				

Renseigné / non renseigné

A traiter par une règle

Améliorer la qualité des données

…selon la société

- **Paramètres de qualité des données**
  - **Complétude des occurrences**
    - Ce paramètre s'applique aux entités

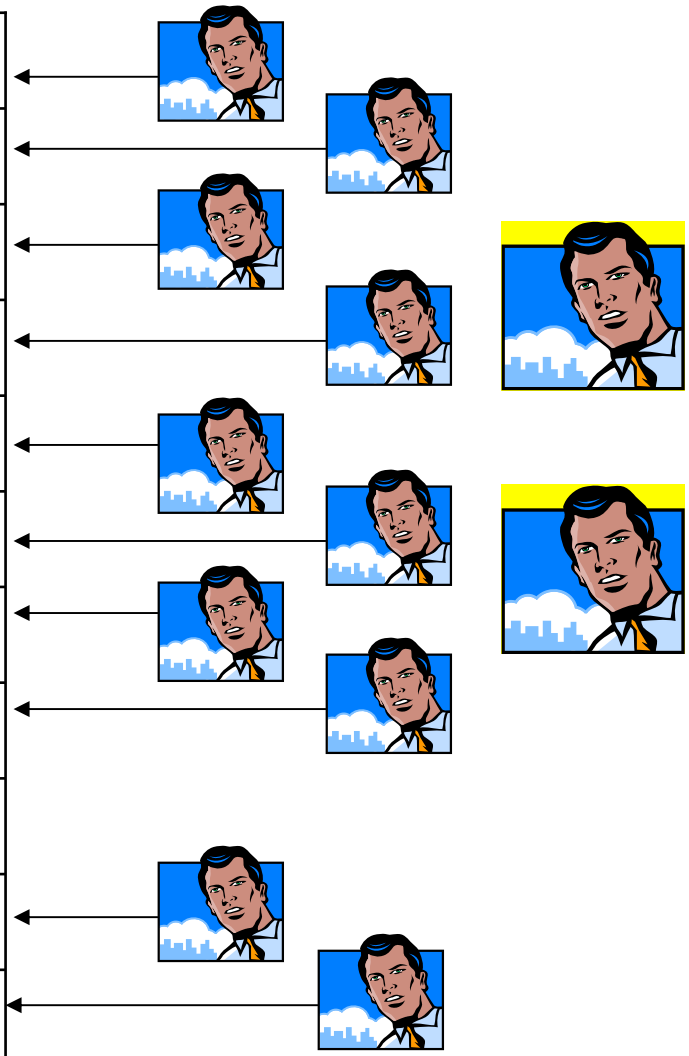
      (une entité = un objet, une occurrence)
      - Schéma

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

001	Données étudiant 1
002	Données étudiant 2
003	Données étudiant 3
004	Données étudiant 5
005	Données étudiant 6
006	Données étudiant 7
007	Données étudiant 8
008	Données étudiant 10
⋮	⋮
n-3	Données étudiant n-1
n-2	Données étudiant n



séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - **Non-duplication**
    - Ce paramètre s'applique aux entités
      - Schéma

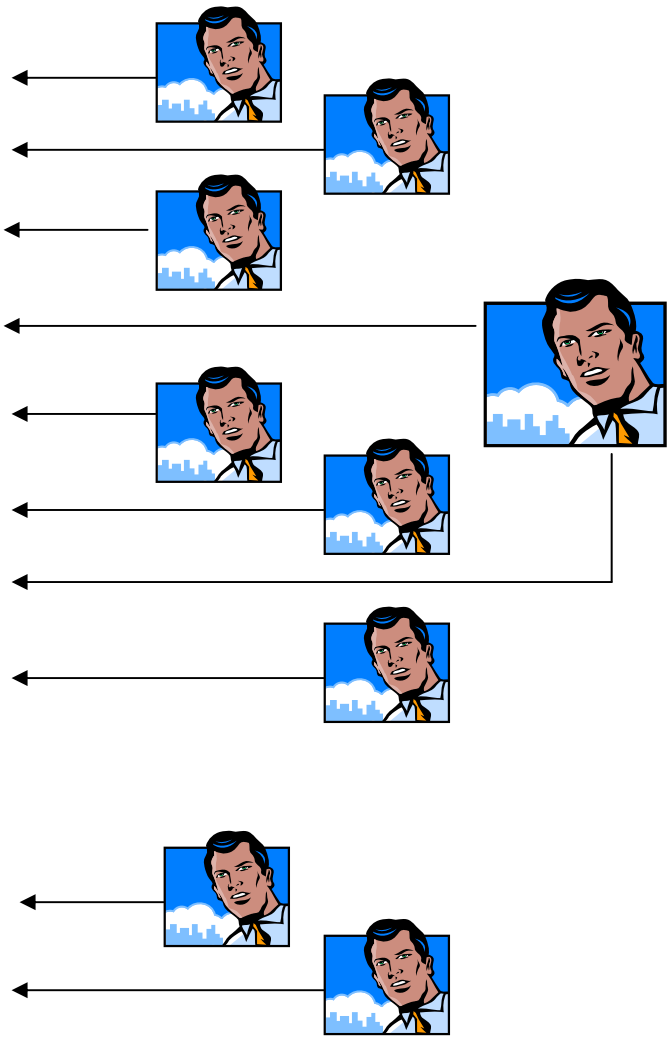
séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

001	Données étudiant 1
002	Données étudiant 2
003	Données étudiant 3
004	Données étudiant 5
005	Données étudiant 6
006	Données étudiant 7
007	Données étudiant 5
008	Données étudiant 8
⋮	⋮
n	Données étudiant n-1
n+1	Données étudiant n



Séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

séminaire

- **Paramètres de qualité des données**
  - **Actualité**
    - Ce paramètre s'applique aux attributs

N°	Attribut	Données	*Réel < 31/12*	*Réel >= 01/01*
001	Nom	Martin	*Martin*	*Durand*
002	Prénom	Sophie	*Sophie*	*Sophie*
003	N°	32	*32*	*64*
004	rue	rue des oliviers	*rue des oliviers*	*Bd des pins*
005	compl adresse			*Les pins*
006	CP	34000	*34000*	*34000*
007	Ville	Montpellier	*Montpellier*	*Montpellier*
008				

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - **Cohérence**
    - Ce paramètre s'applique à des occurrences d'ensembles de données
      - Traduit le lien algorithmique ou sémantique d'un ensemble de données avec d'autres données ou avec d'autres ensembles de données représentant les mêmes objets du monde réel dans des bases de données différentes et redondantes
      - Ses déclinaisons sont nombreuses, de l'algorithme le plus simple au plus complexe : listes / limites de valeurs, règles métier transcrites en algorithmes, croisement avec d'autres données
      - Les contrôles associés sont supportés par des règles et matérialisés par des algorithmes informatisés ; leur mise en œuvre est source de fiabilisation
    - *Exemples* :
      - Dernier établissement fréquenté / situation précédente ;
      - Les codes des destinations des activités, sous destinations et détails de sous destination sont au format alphanumérique sur 5 positions dans *Harpège* afin d'être en cohérence avec les nomenclatures NABuCo.

Améliorer la qualité des données

…selon la société

- **Paramètres de qualité des données**
  - Mise en cohérence pour fiabilisation
    - Schéma

Améliorer la qualité des données

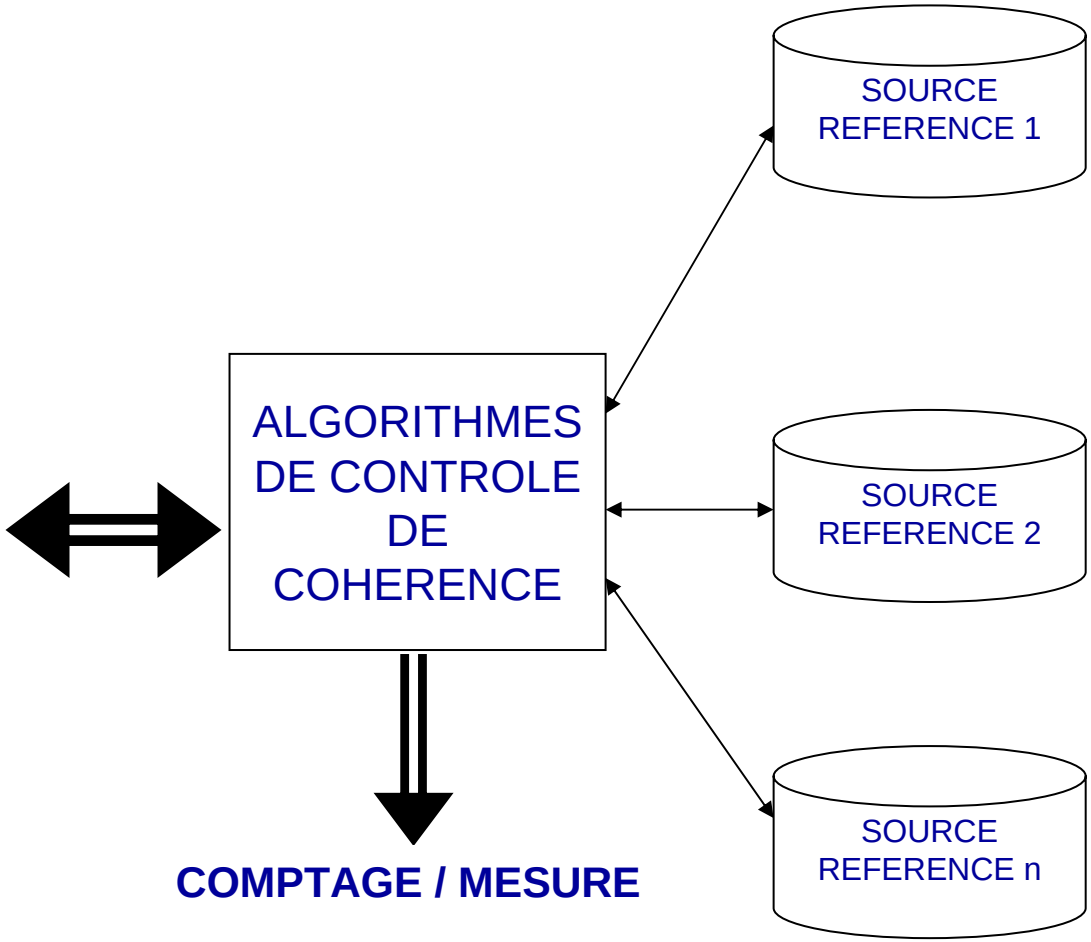
…selon la société Quadem

Data Quality way

séminaire

DONNEES A FIABILISER

001	Données étudiant 1
002	Données étudiant 2
003	Données étudiant 3
004	Données étudiant 4
005	Données étudiant 5
006	Données étudiant 6
007	Données étudiant 7
008	Données étudiant 8
⋮	⋮
n-1	Données étudiant n-1
n	Données étudiant n



Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - Utilisation des paramètres : mesure
    - Indispensable dans tout projet d'amélioration

Taux de Conformité : f (paramètres de base, pondérations)

$TC_{paramètre\ de\ base}$ = nb occurrence OK / nb occurrences total

$$TC = a1 \times TC_{exact} + a2 \times TC_{cpl\_v} + a3 \times TC_{cpl\_o} + a4 \times TC_{n\text{-}dupl} + a5 \times TC_{act+} + a6 \times TC_{coh}$$

séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Paramètres de qualité des données**
  - Utilisation des paramètres : mesure
    - ***Exemple* : diagnostic avec 3 paramètres sur un échantillon de 500 étudiants :**
      - **exactitude** : mesurée sur 5 attributs (nom, prénom, date naissance, série baccalauréat, catégorie socioprofessionnelle) : 45 champs erronés ou non renseignés *(complétude des valeurs intégrée à l'exactitude : non renseigné = erroné)*
      - **complétude des occurrences** : 3 étudiants ne sont pas identifiés ou pas trouvés
      - **non-duplication** : 8 étudiants existent en double dans les fichiers
      - **Pondérations** : a1 (TCexact)= 0,2 ; a2 (TCcompl)= 0,6 ; a3 (TCnon-dup) = 0,2

TC = 0,2 x (2455/2500) + 0,6 x (497/500) + 0,2 x (492/500) = 98,96%

Séminaire

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Qualité des données : l'état de l'art**
  - Synthèse de l'état de l'art
    - Les démarches performantes sont inspirées des démarches qualité utilisées dans les autres domaine
      - *exemple* : **Six Sigma**
    - Les anglo-saxons sont en avance de plusieurs années
    - Normes : il n'en existe que pour les données géographiques

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

- **Qualité des données : l'état de l'art**
  - Un *exemple* de méthode : Six Sigma
    - *objectif : taux de pièces défectueuses inférieur à 3,4 / 1.000.000*
  - Application de Six Sigma : DMAIC
    - **D**éfinir : définir le contour du projet
    - **M**esurer : mettre en place et mesurer la Qualité des Données sur **un échantillon**
    - **A**nalyser : analyser les résultats, identifier les causes de non qualité
    - **I**mprove : améliorer (action sur les processus, rattrapage du passé)
    - **C**ontrôler : pérenniser les acquis, faire vivre, améliorer

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Sources de documentation

- International Association for Information and Data Quality (IAIDQ)

  www.iaidq.org

- Ouvrages de référence :
  - **La qualité des données à l'âge de l'information (Thomas Redman)**
  - **Improving Data Warehouse and Business Information Quality (Larry English)**
- Séminaires et conférences :
  - **2006 IAIDQ Information and Data Quality Conference 16-19 October 2006, San Francisco USA**
  - **Data Management and Information Quality Conference 30 October - 2 November 2006 London, UK**

Améliorer la qualité des données

…selon la société Quadem

Data Quality way

Conclusion

- Améliorer la QDD = projet d'établissement
  - Changer la culture et les mentalités à propos des données
  - Impliquer le management
  - Mettre en place des relations du type Client / Fournisseur entres les contributeurs
  - Mettre en place une organisation et des démarches de certification…
- Améliorer la QDD = privilégier l'avenir
  - D'abord garantir la qualité des nouvelles données en agissant sur les processus
  - Ensuite rattraper le passé en corrigeant les données déjà existantes
- Améliorer la QDD = un projet outillé
  - Un projet, une démarche industrielle, des outils modernes, un Plan Qualité
  - Un outil essentiel : la mesure de la Qualité des Données
  - Travailler avec des règles exhaustives et documentées

Améliorer la qualité des données

Donnée/information : des définitions concurrentes

- Dès lors que l'on s'intéresse à la qualité des données, on peut trouver un intérêt similaire pour la **qualité de l'information**.
  - Définir et comprendre ce qu'est l'information est un véritable casse-tête tant les différences en signification sont considérables et les approches développées variées et parfois incomplètes.
  - De nombreux spécialistes définissent les données comme « la matière première à partir de laquelle on développe l'information ».
    - En d'autres termes, les données constituent les entrées brutes des processus dont les sorties sont l'information « raffinée » (Dorn P.H.)

Améliorer la qualité des données

Donnée/information : des définitions concurrentes

- Il convient de se concentrer sur les données et de considérer où l'information apparaît dans le cycle de vie des données.
- Le cœur de cycle d'acquisition est la donnée ; le cœur du cycle d'usage est l'information.
- Les deux étapes de l'acquisition, création d'un modèle de données et obtention des valeurs, sont les plus importantes pour la qualité des données.

Améliorer la qualité des données

Aide méthodologique

- **A la mise en place d'un projet de qualité des données**
  - Démarrer à petite échelle sur un domaine d'activité afin de :
    - pouvoir mesurer les progrès réalisés rapidement,
    - inciter les personnels à poursuivre l'effort engagé,
    - établir un contrôle statistique et la vérification à la conformité aux exigences fixées.
  - Constituer une équipe projet transversale
    - 4 à 5 volontaires : gestionnaires du domaine concerné, informaticiens, chargés de mission pilotage ou statistiques par exemple

Améliorer la qualité des données

Aide méthodologique

- **A la mise en place d'un projet de qualité des données**
  - Le projet doit être :
    - porté par un responsable de l'établissement (ex. : chargé de mission à l'Informatique) ou d'une cellule qualité si elle existe.
    - expliqué aux personnels concernés (communication claire sur les objectifs poursuivis, la politique des données, les enjeux pour l'établissement…).
  - Les personnels doivent savoir quoi faire et à quel moment, pour développer et maintenir des données fiables
    - prévoir des formations en interne sur :
      - Les évolutions des applications de gestion (meilleure maîtrise des fonctionnalités, minimisation des erreurs),
      - Les processus qualité des données définis,
      - Le SI de l'établissement, etc…

Améliorer la qualité des données

Aide méthodologique

- **A la mise en place d'un projet de qualité des données**
  - La qualité des données peut devenir une partie du travail du responsable de la cellule qualité ou du chargé de mission à l'Informatique.
    - La présence d'un meneur bien visible et engagé est essentielle face aux résistances au changement ;
    - …mais aussi une partie du travail quotidien de chaque personne manipulant des données.
  - Ce premier projet peut ensuite servir de modèle et être étendu à d'autres domaines.

Améliorer la qualité des données

Aide méthodologique

- **A la fiabilisation et à la valorisation des données**
  - Vérifier la fiabilité des données passe par :
    - Un besoin de rapprochement des données
    - Des enquêtes régulières
    - La complétude des données :
      - saisie des données manquantes (mobilisation et responsabilisation des utilisateurs)
      - correction des erreurs de saisie
    - L'analyse des raisons sous-jacentes aux erreurs découvertes
      - recenser les sources d'information de référence (documentations, données informatiques, connaissances des utilisateurs) à l'origine de ces erreurs et mettre en place des solutions pérennes pour y remédier

Améliorer la qualité des données

Aide méthodologique

- **A la fiabilisation et à la valorisation des données**
  - La fiabilité des données s'inscrit essentiellement dans une démarche qualitative.
  - Le « qualitatif par applications »
    - NABuCo
      - Globalement fiable vu le grand nombre de pointages de l'Agence comptable pour fiabiliser ses données
    - HARPEGE
      - Relativement fiable vu le caractère sensible des données, les nomenclatures nationales, les champs obligatoires, les contrôles de cohérence

Améliorer la qualité des données

Aide méthodologique

- **A la fiabilisation et à la valorisation des données**
  - APOGEE
    - Très divers
    - Fiable pour les variables sensibles qui vont se retrouver sur le diplôme par exemple
    - De plus en plus approximatif si on s'éloigne de l'étudiant (CSP des parents, adresse parents…)

Améliorer la qualité des données

Aide méthodologique

A la fiabilisation et à la valorisation des données

- **Malgré tout le qualitatif reste très difficile à corriger, voire impossible.**
  - Des données dynamiques avec des taux de création, de stockage, de manipulation… très élevés.
  - La variable peut être renseignée mais non « correctement » renseignée.

    Exemple :
    - un étudiant aurait dû indiquer le code 34 (professeurs scientifiques) pour la CSP de ses parents, à la place, on le retrouve avec le code 35 (profession des arts et des spectacles).
    - erreur de saisie, mauvaise lecture de la grille des CSP,…??

Améliorer la qualité des données

- **Limites à la fiabilisation des données au sein de l'établissement**
  - **Comportement des utilisateurs en matière de saisie des données (sous-utilisation des logiciels de gestion, données manquantes, champs facultatifs non renseignés…)**
  - **Développement « sauvage » pour répondre aux besoins non couverts par l'application de référence**
  - **Mobilité et aptitude au changement des personnels**
  - **Implication des enseignants limitée**
  - **Intérêt variable des Directeurs d'UFR**

Améliorer la qualité des données

- **Risques inhérents à une dispersion des données au sein de l'établissement**
  - **Relations entre les composantes, les sites et les services centraux :** dispersion des données
    - Différences de situations entre les sites et les composantes (niveau d'information, diversité des lieux de décision, élaboration des budgets…)
    - Difficulté à collecter les données
    - Nécessité de mutualiser les ressources (mise en commun de certains moyens : tableaux de bord, entrepôt de données)
  - **Absence d'instruments de pilotage et d'instruments de contrôle**
  - **Niveaux différents de pilotage au sein de l'établissement**

Améliorer la qualité des données

- **Risques inhérents à une dispersion des données au sein de l'établissement**
  - **Appréciation imprécise des moyens de l'établissement** (occupation des salles de cours, des enseignements prévus, de la nature précise des dépenses effectuées…)
    - Un rapport de l'IGAEN de 1998 soulignait les « difficultés à maîtriser les données en matière de gestion financière ».
    - Nécessité de mettre en place des procédures d'évaluation des activités afin de contrôler les résultats des actions mises en œuvre

Améliorer la qualité des données

Emergence d'un contrôle de gestion

- Situé au cœur des acteurs et de leurs actions
  - *Exemple* : création de cellules de contrôle de gestion au sein des établissements
- Nécessité de développer le potentiel d'informations disponibles grâce aux bases de données, mais aussi capacité des acteurs de s'informer, d'analyser, d'apprécier et d'agir
- Nécessité d'une instrumentation de contrôle combinée à celle de pilotage
  - Construire ses propres indicateurs, c'est s'assurer de leur fiabilité et de leur pertinence.
- Le contrôleur de gestion aide au pilotage. Il collecte les données mais ne les produit pas.

Améliorer la qualité des données

- La **LOLF** exige une meilleure qualité, « certifiable » des comptes s'appuyant sur des outils et indicateurs de pilotage
  - Certifiable : démarche innovante de qualité comptable au niveau de l'établissement
  - Mise en place d'une comptabilité permettant de mesurer et d'analyser les coûts
  - Pression pour le changement exogène