Groupe de contact FNRS
« Analyse critique et amélioration de la qualité de l'information numérique »

30 janvier 2014

Du stemma codicum au « data tracking »

Isabelle Boydens

Smals ICT for society
fnrs LA LIBERTÉ DE CHERCHER
ULB

Groupe de contact FNRS
« Analyse critique et amélioration de la qualité de l'information numérique »

- Qualité de l'information : définition et enjeux
- Groupe FNRS : objectifs, réalisations et défis
- Application de la critique historique aux sources informatiques à des fins opérationnelles : actualité
  - Herméneutique des bases de données
  - Email address reliability (introduction)

Qualité de l'information : définitions et enjeux

- *ποιος, qualis*, “quel ?”, "welk ?“, “which ?”, …

- Au sens appréciatif, origines industrielles :
  - Taylorisme, production en série
  - Concept de « one best »
  - Perfection : « non-valeur »
  - Arbitrage « coûts-bénéfices »

- Qualité de l’information : adéquation de l'information aux usages , "*fitness for use*"

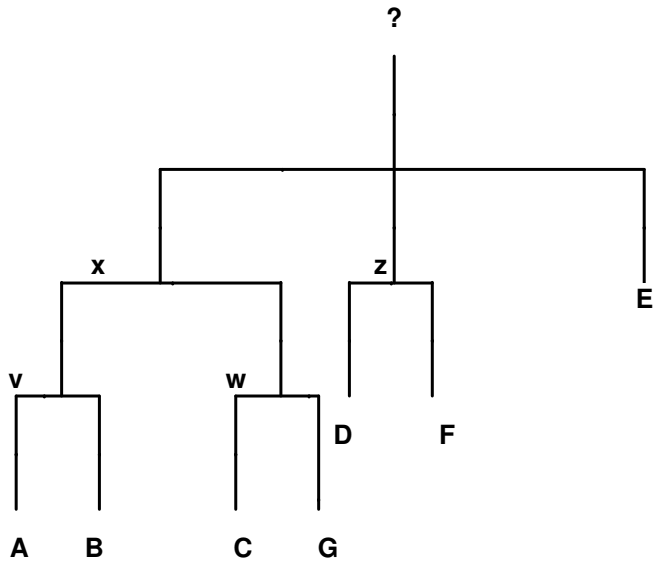
Enjeux stratégiques lorsque l'information est un instrument d'action sur le réel

Quelques exemples :

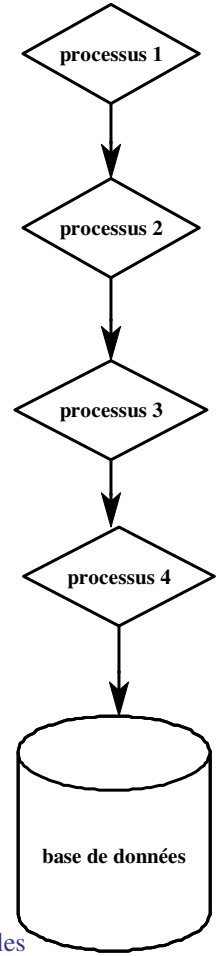
- 1442, Laurent Valla démontre que la « *Donation de Constantin* » est un faux antidaté de 4 ou 5 siècles
- Étude du réchauffement climatique (couche d'ozone, 1980)
- Première guerre du Golfe (1990-1991)
- 1999, guerre du Kosovo, bombardement de l'ambassade de Chine à Belgrade
- 2008, interactions entre données et marché financier : Google Finance, Yahoo! Finance
- 2013, secteur administratif
- …

Du stemma codicum au « data tracking »

Stemma codicum : "le Lai de l'ombre", poème français du 13ème siècle



"Data tracking", AT&T Bell Laboratories, 1992



Critique historique et bases de données

- Critique externe :

  étude formelle et syntaxique, analyse du lieu, de la date, de l'auteur, …

- Critique interne :

  interprétation des phénomènes empiriques en termes d'interactions par rapport à un cadre conceptuel plus général construit afin de leur conférer un sens

→ Reconstruction conjecturale argumentée de l'information (cercle herméneutique)

→ Communauté IT : rejoint le « data quality research »

Groupe de contact FNRS : objectifs, réalisations et défis (1)

- Interuniversitaire (1994 - …)
  - ULg, ULB (Infodoc, STIC), Unamur, UCL, Umons, Fusl, KUL, UGent, …
- Objectifs : rassembler tous les acteurs impliqués
  - Utilisateurs et concepteurs IT
  - Sciences humaines et sciences appliquées
  - Monde scientifique et industriel
- Quelques réalisations :
  - Smals, « Data Quality Competence Center » au sein de l'egovernment (2001 - …)
  - Création d'un cours universitaire dédié : « qualité des données et de l'information numérique » (ULB, créé en 2006)
  - Recherches confrontant théorie et pratique (voir orientation bibliographique en annexe, 1993-2014)

Groupe de contact FNRS : objectifs, réalisations et défis (2)

- Thématiques traitées lors des rencontres du groupe :
  - Méthodes statistiques et quantitatives
  - Archivage
  - Qualité logicielle
  - Systèmes de méta-information
  - …
- Quelques résultats opérationnels novateurs :
  - Herméneutique des bases de données et gestion intégrée des anomalies
  - Meta-information systems (glossaires DmfA)
  - Data quality tools and technics (profiling, standardization, matching)
  - Email address reliability
  - …
- Défis :
  - Établir un dialogue entre disciplines et acteurs distincts
  - Vocabulaire commun

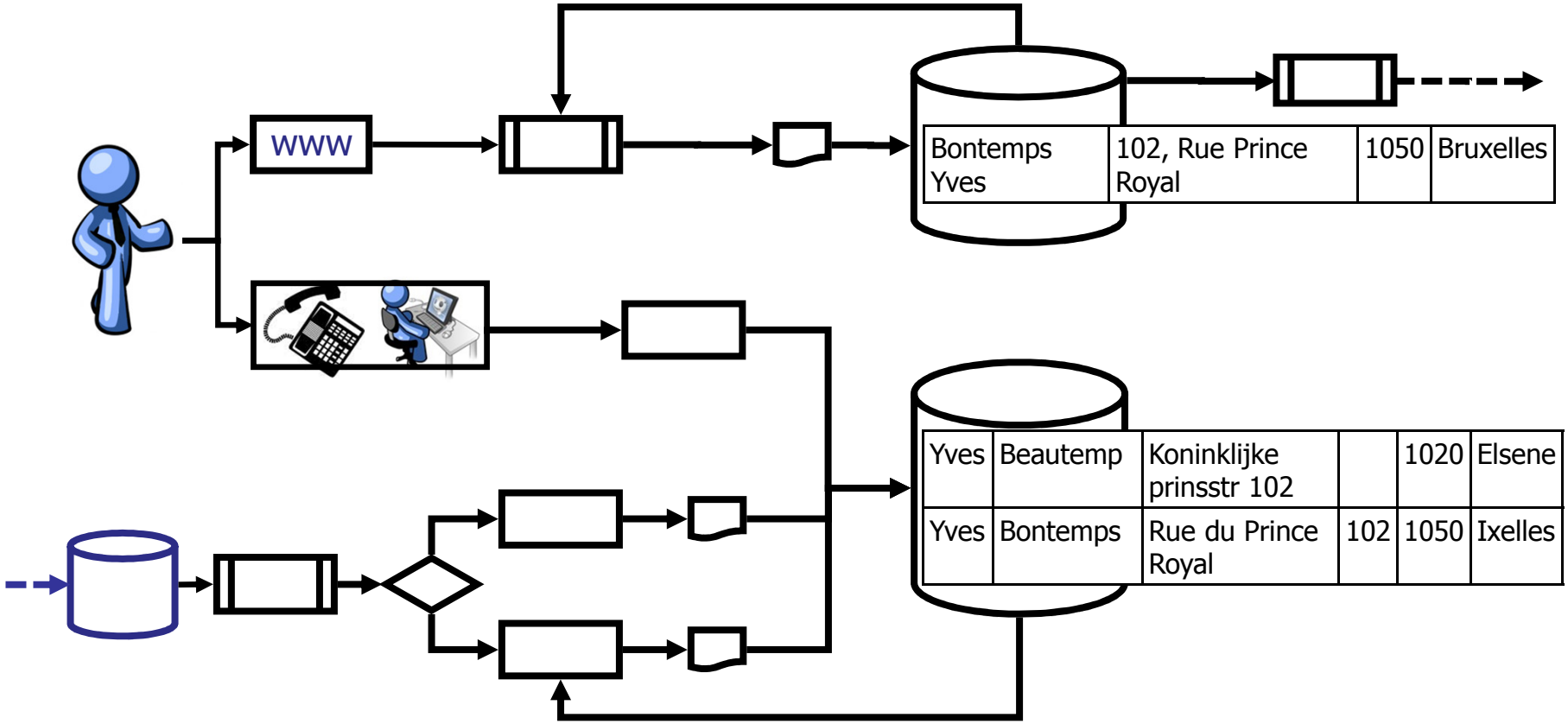
Herméneutique des bases de données

- Qu'est-ce qu'une donnée ?
- Qu'est-ce qu'une donnée "correcte" ?
- Comment les données se construisent-elles progressivement ?

Qu'est-ce qu'une donnée ?

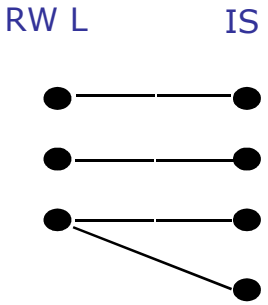
- Sur le plan formel, triplet :
  - Intitulé (ex : salaire mensuel)
  - Domaine de définition (ex : « valeur numérique incluse entre 1000 € et 100.000 € »)
  - Valeur à un instant t : 3000 €
- Sur le plan conceptuel, différence entre données :
  - Déterministes : définition immuable
  - Empiriques : définition évolutive avec l'interprétation humaine du réel (« concepts mobiles »)
- « Closed world assumption » (hypothèse du monde clos)

Donnée : formalisation du réel observable inscrite dans un système d'information

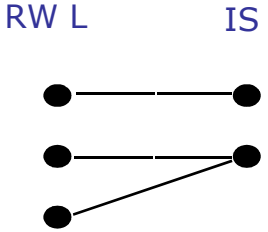


Qu'est-ce qu'une donnée correcte ? Isomorphisme entre le réel observable et ses représentations ?

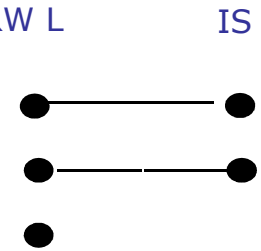
Représentation correcte



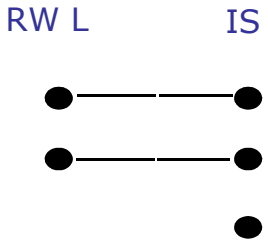
Représentation ambiguë



Représentation incomplète



État non significatif



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

Etude de cas : bases de données de la sécurité sociale belge (ordres de grandeur)

- Nombre enregistrements saisis chaque trimestre : 4.000.000
- Montants en jeu : 45 milliards d'euros annuels
- Plusieurs centaines de champs
- Nombre d'anomalies formelles : 10 % environ (voir aussi secteur bancaire)
- Service affecté au traitement des anomalies : environ 300 personnes
- Modifications de schémas fréquentes et complexes (évolutions législatives)

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

Test d'intégrité au moment de la saisie d'une déclaration sociale

Employeur (déclaration)

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

Anomalie formelle

Catégorie_taux (table de vérification interne)

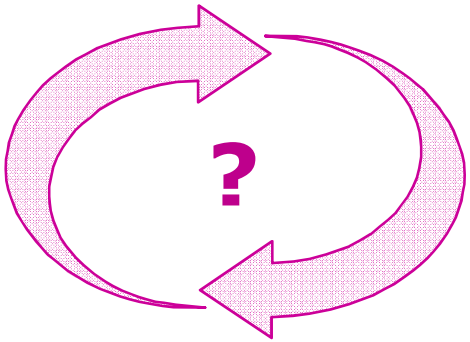
Catégorie	Taux-cotisation
énergie solaire	0.28%
énergie éolienne	0.27%
énergie biomasse	0.29%

Qu'est-ce qu'une donnée correcte ?

- Typologie des violations de contraintes d'intégrité :
  - Erreur formelle
  - Présomption formelle d'erreur (anomalie)
    - A priori
    - A posteriori
  - Situation indétectable formellement (faux actifs, …)

Les « données » ne sont pas « données »

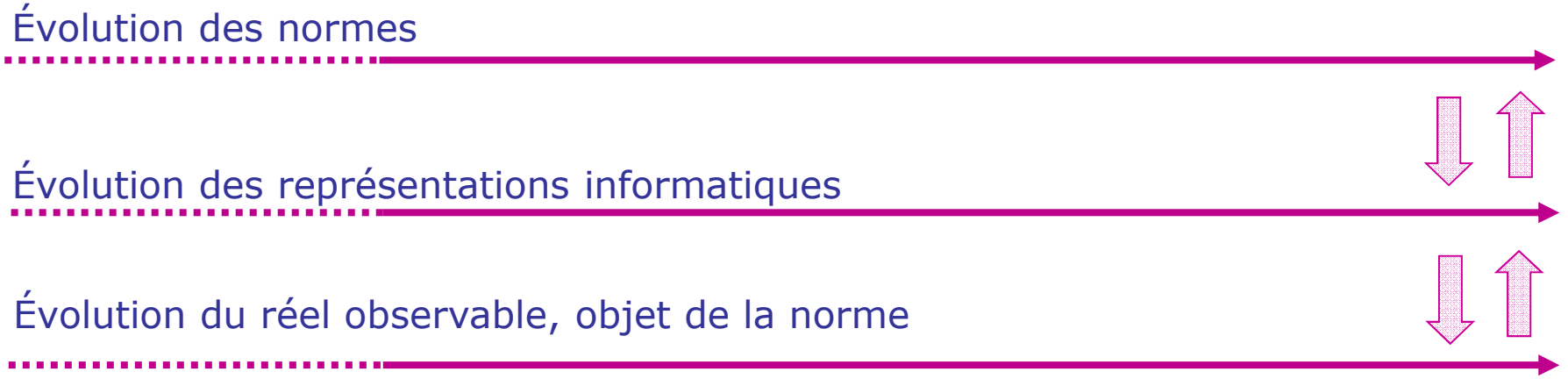
On ne dispose d'aucun référentiel "absolu" en vue de tester la correction d'une vaste base de données empiriques



Étude des anomalies à des fins opérationnelles

Comment les données se construisent-elles progressivement ?

Cadre d'analyse temporel

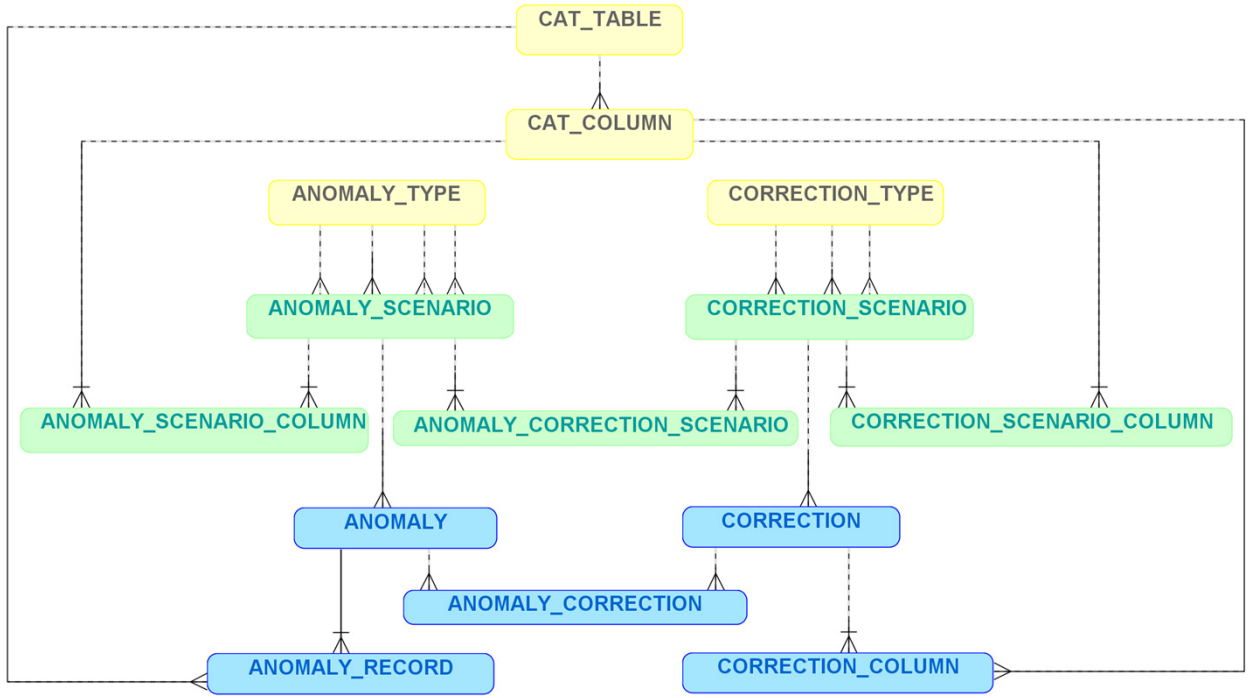


F. Braudel, « *temporalités étagées* » (1976)

N. Elias, « *continuum évolutif* » (1996)

Stratégie opérationnelle : passage d'un « monde clos » à un « monde ouvert sous contrôle »

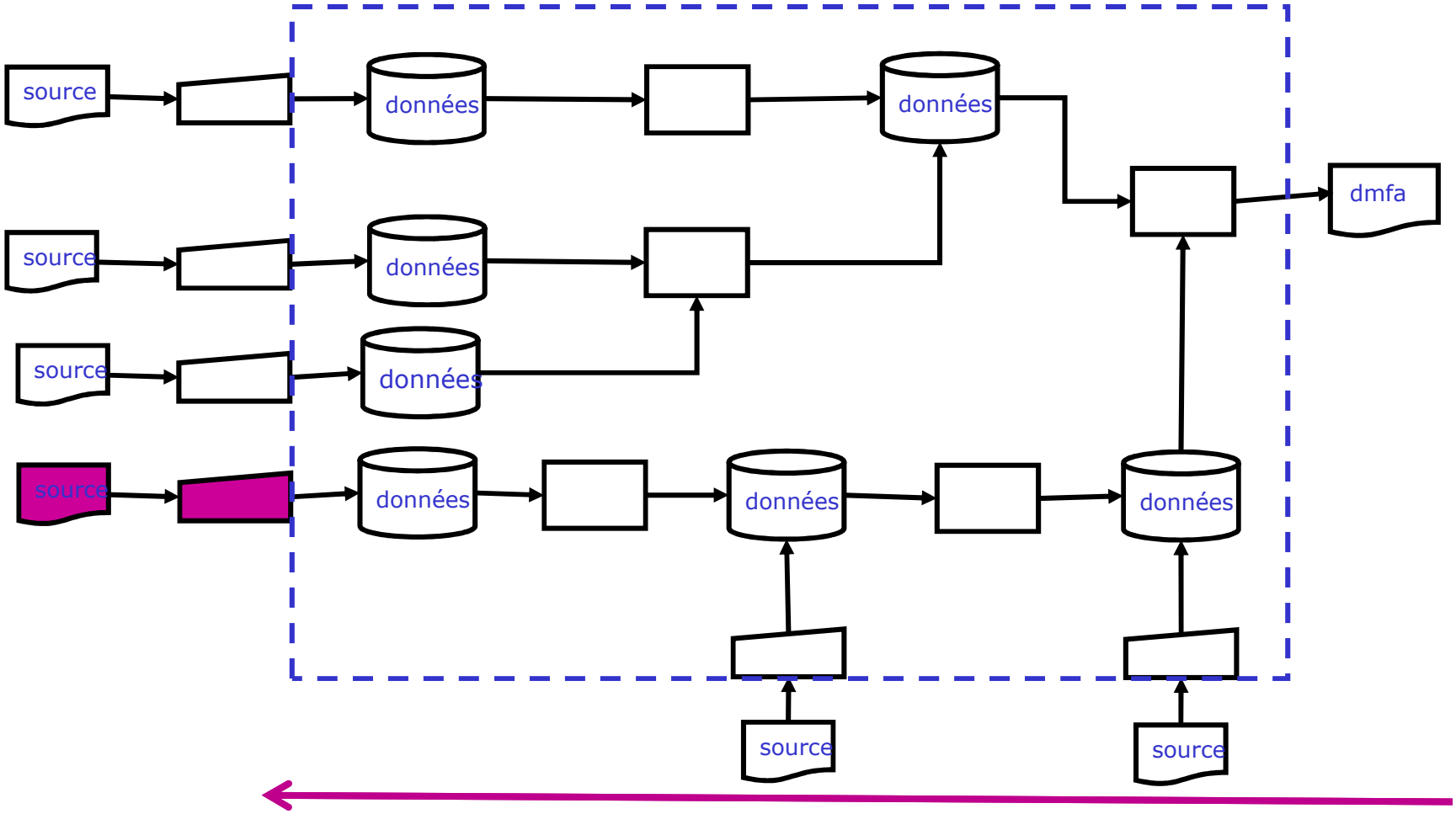
- Extension du modèle de la base données
- Intégration du traitement des anomalies et historique :
  - Typologie et suivi dans le temps
  - Détection / correction / validation …



Stratégies de gestion

- Conception d'indicateurs de qualité pour :
  - Quantifier le temps et la nature des traitements (monitoring)
  - Identifier les cas d'anomalies fréquemment validées (anomalies « fictives »)
  - En évaluer les causes (obsolescence des contrôles ?)
  - Adapter progressivement les contrôles pour les rendre adéquats à l'évolution des réalités et diminuer le nombre d'anomalies fictives
  - Exemples d'application concrète (LATG – DmfA) :
    - Déduction de cotisation pour les "bas salaires"
    - Baisse structurelle du nombre d'anomalies de 50 % (14.000/7.000)

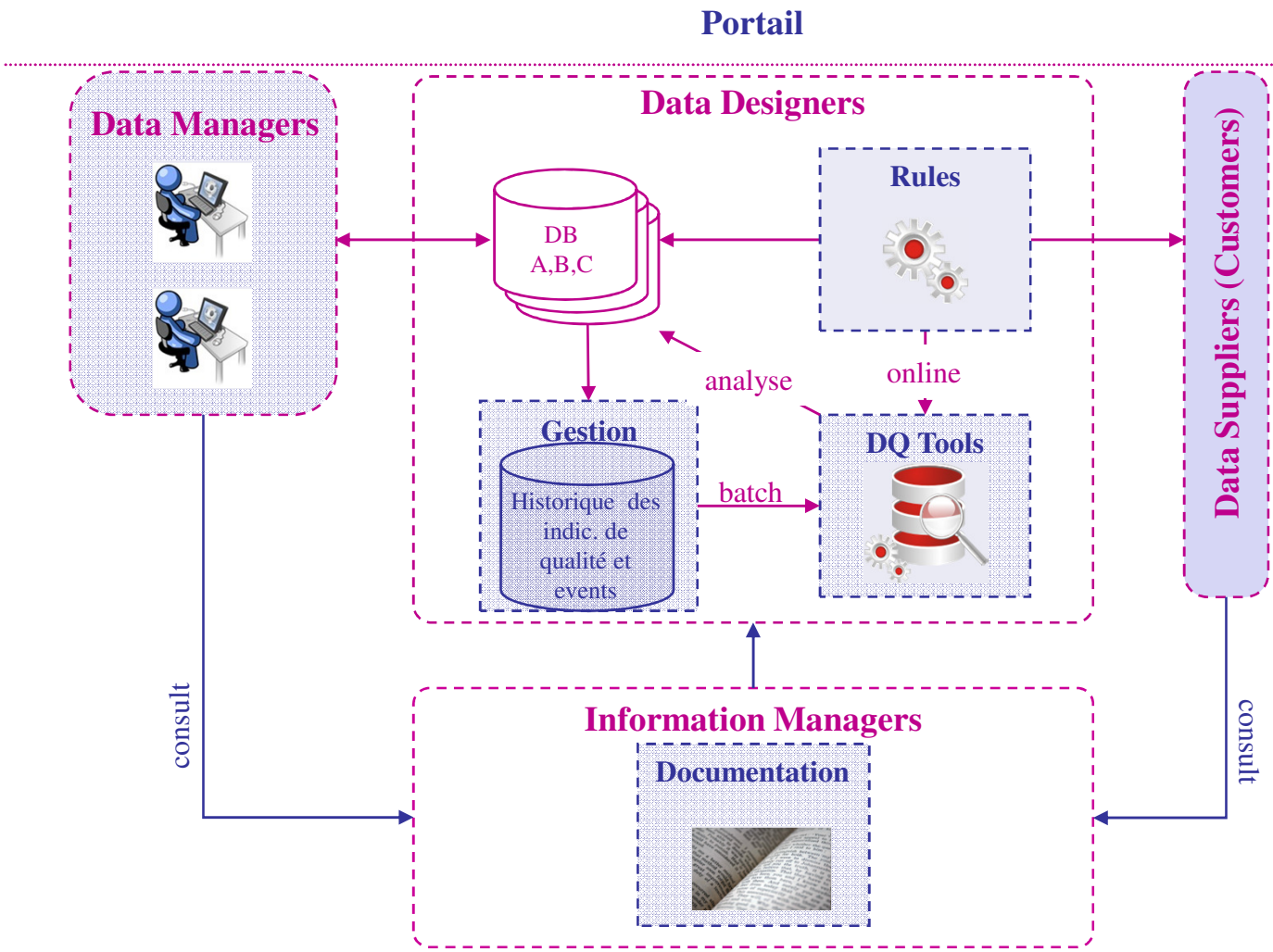
Approche complémentaire : « Back tracking »



ROI & généralisation

- Bénéfices :
  - Diminution des anomalies et du manpower en terme de temps de correction (de 50 % à un facteur 20 dans des opérations stratégiques et ciblées)
  - Rapidité et précision des traitements financiers
  - Aide à la documentation du système
  - Crédibilité du système d'information
  - Renforcement du partenariat avec les utilisateurs

Organisation



Email address reliability (1)

- Contexte : étude menée à la section « recherches » de Smals, « Data Quality Competence Center » (2013)
- Enjeux stratégiques :
  - Opérationnels dans les secteurs privés et publics (dématérialisation de l'information, egovernment)
  - Sur le plan de l'histoire des techniques et de leurs usages :
    - Adresses e-mail (1971 - …)
    - Aléas et difficultés d'une approche prévisionnelle (fax, …)

Email address reliability (2)

- Cumul d'incertitudes et volatilité :
  - Standards (double système de standards évolutifs)
  - Noms de domaine
  - Usages
- Dégressivité potentielle de la validité dans le temps
- Temporalités étagées, continuum évolutif

Email address reliability (3)

- Critique externe :
  - Analyse syntaxique
- Critique interne
  - Tests d'existence
  - Matching interne
- Suivi dans le temps des présomptions d'erreurs (incertitude, anomalies, …) & des indicateurs de qualité

Un système d'information se transforme avec l'interprétation des valeurs qu'il permet d'appréhender

Orientation bibliographique (1)

- Base D., « It's about Time!: Temporal Aspects of Metadata Management in the Work of Isabelle Boydens ». In *Cataloging & Classification Quarterly* (The International Observer), volume 49, n° 4, 2011, pp. 328-338 (Université de Chicago, recension des recherches et publications d'Isabelle Boydens, période 1993-2011).
- Bade D., *Responsible Librarianship, Library policies for unreliable systems*. Library Juice Press, 2007.
- Berten V. et Boydens I., *Email Address Reliability*, Deliverable, Section Recherches, Bruxelles, Smals, 2013 (à paraître).
- Berti Equille L. éd., *La qualité et la gouvernance des données au service de la performance des entreprises*. Paris, Hermès, 2012.
- Bizingre J., Paumier J. et Rivère P., *Les référentiels du système d'information*. Paris, Dunod, 2013.
- Bloch L. *Système d'information : obstacles et succès*. Paris : Vuibert, 2005.
- Bontemps Y., Boydens I. et Van Dromme D., *Data Quality : tools.* Deliverable, section recherches, Bruxelles, Smals, 2007.

Orientation bibliographique (2)

- Boydens I., « Informatique et qualité de l'information. Application de la critique historique à l'étude des informations issues de bases de données ». In *Belgisch Tijdschrift voor Nieuwste Geschiedenis. Revue belge d'histoire contemporaine*, vol. 3-4, 1993, p. 399-439.
- Boydens I., *Informatique, normes et temps*. Bruxelles : Bruylant, 1999.
- Boydens I., « Les bases de données sont-elles solubles dans le temps? ». In *La Recherche hors série* ("Ordre et désordre"). Hors série n° 9, novembre-décembre 2002, p. 32-34.
- Boydens I., « Déploiement coopératif d'un dictionnaire électronique de données administratives ». In *Revue Document Numérique,* vol. 5, n°3-4, 2001, Paris, Hermès, p. 27-43.
- Boydens I., « La conservation numérique des données de gestion ». In *Revue Document Numérique,* septembre 2004, Paris, Hermès, p. 13-22.
- Boydens I., "Qualité de l'information et administration électronique : enjeux et perspectives". In Assar S. et Boughazala I., éds., Administration électronique. Constats et perspectives. Paris : Lavoisier - Hermès Sciences, 2007, p. 103-120 (chapitre 5).
- Boydens I., "Hiérarchie et anarchie : dépasser l'opposition entre organisation centralisée et distribuée ?" In Hudon M. et El Hadi W. M., éds, Les cahiers du numérique (Numéro thématique « Organisation des connaissances et Web 2.0 »). Paris : Editions Hermès Sciences, 2010, vol. 6, n°3, p. 77-101.

Orientation bibliographique (3)

- Boydens I., "Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium". In Assar S., Boughzala I. et Boydens I., éds., "Practical Studies in E-Government : Best Practices from Around the World", New York, Springer, 2011, p. 113-130 .
- Boydens I. et Van Hooland S., "Hermeneutics applied to the quality of empirical databases". In *Journal of Documentation*, volume 67, issue 2, 2011, p. 279-289.
- Boydens I., Hulstaert A. et Van Dromme D., *Gestion intégrée des anomalies - Evaluer et améliorer la qualité des données*, Delivrable, Section Recherches, Bruxelles, Smals, 2011.
- Boydens I., Mendez E. et Van Hooland S., "Between commodification and sense-making. On the double-sided effect of user-generated metadata within the cultural heritage sector" In Marty P. F. et Kazmer M. M., éds, *Library Trends on "Involving Users in the Co-Construction of Digital Knowledge in Libraries, Archives, and Museums*, Library Trends, John Hopkins University Press, volume 59, n° 4, spring 2011, pp. 707-720.
- Boydens I., « L'océan des données et le canal des normes ». In Carrieu-Costa M.-J., Bryden A. et Couveinhes P. éds, Les Annales des Mines, Série "Responsabilité et Environnement" (numéro thématique : "La normalisation : principes, histoire, évolutions et perspectives"), Paris, n° 67, juillet 2012, pp. 22-29. http://www.ulb.ac.be/cours/iboydens/annales.pdf

Orientation bibliographique (4)

- De Wilde, M. et Verborgh, R., *Using OpenRefine*. Brimingham-Mumbai : Packt Publishing, 2013 (978-1-78328-908-0).
- Elmasri R. et Navathe S. B., *Fundamentals of Database Systems*. Addison Wesley, 2011 (6eme éd.).
- Loshin D., *The Practicioner's Guide to Data Quality Improvment*. Elsevier, Morgan-Kaufmann OMG Press, 2011.
- Madnick S. E. *et al.*., "Overview and Framework for Data and Information Quality Research"*. In Journal of Data and Information Quality*, Vol. 1, No. 1, 2009, p. 2-22.
- McCallum Q. Ethan, Bad Data Handbook, Mapping the World of Data Problems. Sebastopol, O'Reilly Media, 2012. (analyse critique : http://blogresearch.smalsrech.be/?p=5398 , I. Boydens, 3 avril 2013)
- Olson J., *Data Quality: The Accuracy Dimension*. Elsevier, The Morgan-Kaufmann Series in Database Management, 2003.
- Redman T. C., *Data Quality for the Information Age*. Boston-London : Artech House Publishers, 1996.
- Redman T. C., *Data Quality. The Field Guide*. Boston : Digital Press, 2001.

Orientation bibliographique (5)

- Rivière P., »Indicateurs de qualité en matière de production de données : quelques éléments de réflexion ». In *Courrier des statistiques*, septembre 2005, n°115, p. 35-40.
- Rivière P., *Les référentiels dans un système d'information. Quelques principes*, Paris, Assurance retraite, Direction Qualité, Méthodes et Urbanisation, 2011.
- Shazia S. ed., *Handbook of Data Quality. Research and Practice*. Berlin, Springer, 2013.
- Van Hooland S., "Spectator becomes annotator: possibilities offered by user-generated metadata for image databases". *Paper presented at Immaculate Catalogues: Taxonomy, Metadata and Resource Discovery in the 21st Century*, 13-15 September 2006, University of East Anglia, UK.
- Van Hooland, S., Kaufman, S. et Bontemps, Y., "Answering the call for more accountability: applying data-profiling to museum metadata", *Proceedings of the International conference on Dublin Core and metadata applications, 22- 26 September 2008, Berlin*, Dublin Core Metadata Initiative, Berlin, p. 93-103.
- Van Hooland, S., *Metadata quality in the cultural heritage sector: stakes, problems and solutions*, Thèse de doctorat sous la direction de Boydens I., Département Sciences de l'Information et de la Communication, Université Libre de Bruxelles., 2009.
- Van Hooland S. et Verborgh, R., *Linked data for libraries, archives and museums. How to clean, link and publish your metadata*. Brimingham-Mumbai : Facet Publishing (to be published in 2014).
- Volle M., *De l'informatique*. Paris, Economica, 2006.

Orientation bibliographique (6)

- Quelques ressources en ligne :
  - http://www.ulb.ac.be/cours/iboydens/
  - https://www.smals.be/fr/content/data-quality
  - http://www.smalsresearch.be
  - http://homepages.ulb.ac.be/~svhoolan/
  - http://homepages.ulb.ac.be/~madewild/
  - http://freeyourmetadata.org/
  - http://liliendahl.com/
  - http://iaidq.org/
  - http://www.ocdqblog.com/
  - http://www.dqa.be/
  - http://exqi.asso.fr/