STIC B 510 – « Qualité de l'information et des documents numériques »

13 mars 2019

« Data Quality Tools : concepts and practical lessons from a vast operational environment »

par

Gani Hamiti

Introduction

par

Isabelle Boydens

Smals ICT for society

fnrs LA LIBERTÉ DE CHERCHER

ULB

« Qualité de l'information et des documents numériques »

- Qualité des données :
  - « *fitness for use* » : adéquation des données à leurs objectifs (sociaux, industriels, juridiques, scientifiques, …), sous *contrainte de budget*
  - pas de « *qualité totale* »
  - approche continue et pluridisciplinaire

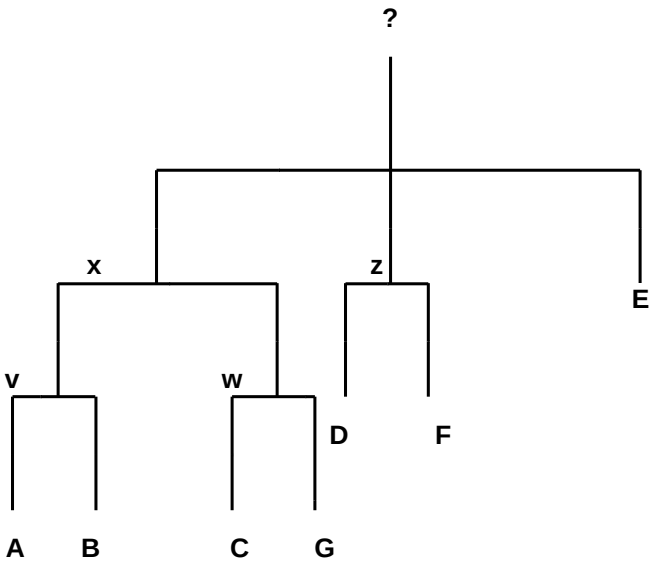
- Groupe de contact FNRS créé en 1994 (25 ans en 2019)
  - http://www.fnrs.be/index.php/sciences-appliquees
  - http://www.fnrs.be/index.php/sciences-humaines-et-politiques

- Centre de compétence en qualité des données, Smals
  - https://www.smals.be/fr/content/data-quality
  - https://www.smals.be/nl/content/data-quality

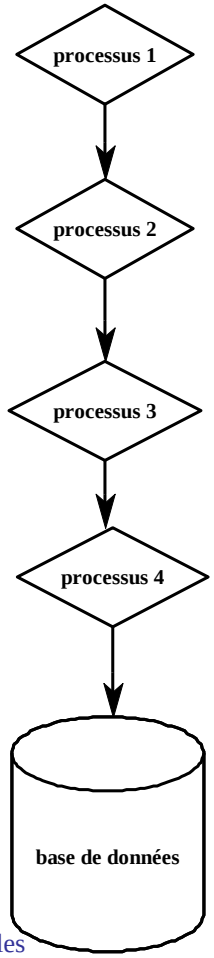
- Cours de Master en STIC à l' Université libre de Bruxelles : « Qualité de l'information et des documents numériques » http://www.ulb.ac.be/programme/cours/2019/STIC-B510/index.html

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

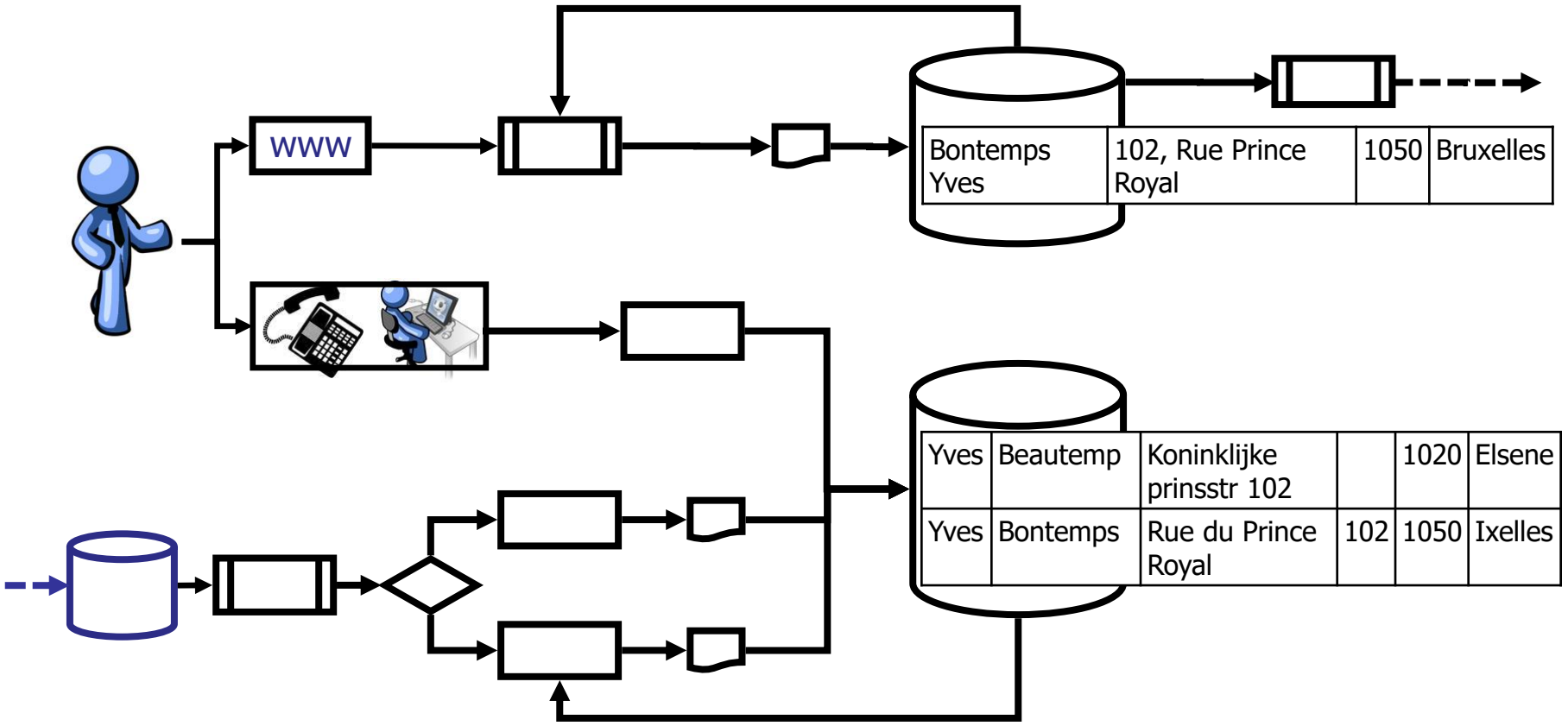


Enjeux stratégiques lorsque l'information est un instrument d'action sur le réel

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, 2017 interactions entre données et marché financier : Google Finance, Yahoo! Finance, World-Check (2017)
- 2013, secteur administratif (Obamacare)
- 2016, domaine des pipelines, de l'hydrologie, …
- 2018-2019, bases de données médicales, en matière de terrorisme, Registre National, …

Evaluer et améliorer la qualité des données : deux approches complémentaires



Approches préventives

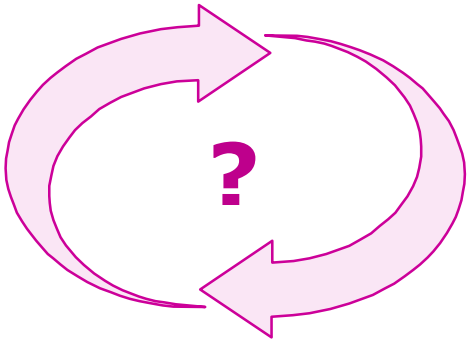
Approches curatives

Qu'est-ce qu'une donnée correcte ?

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

Les « données » ne sont pas « données »

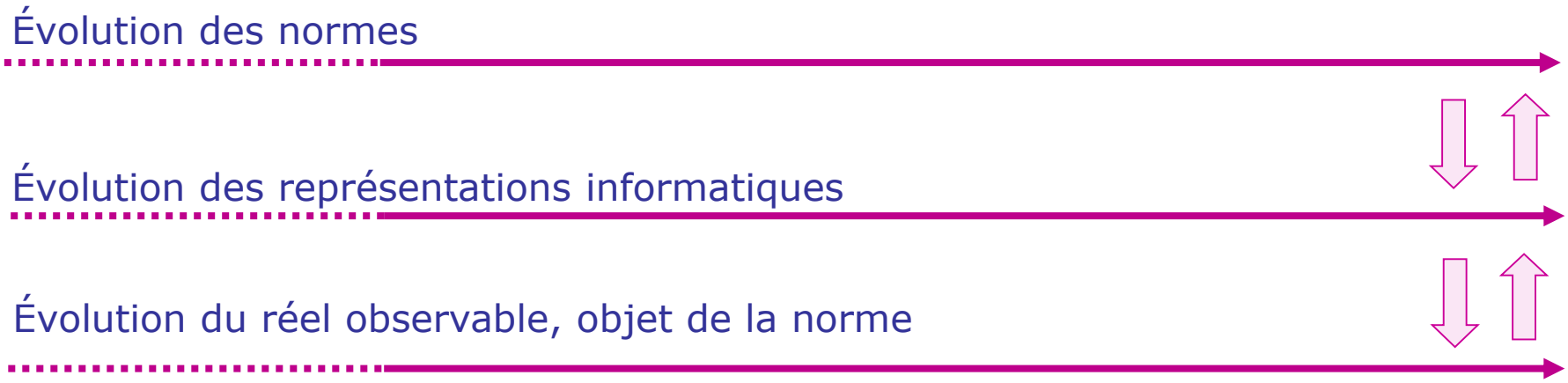
On ne dispose d'aucun référentiel "absolu" en vue de tester la correction d'une vaste base de données empiriques



Étude des anomalies à des fins opérationnelles

Comment les données se construisent-elles progressivement ?

Cadre d'analyse temporel

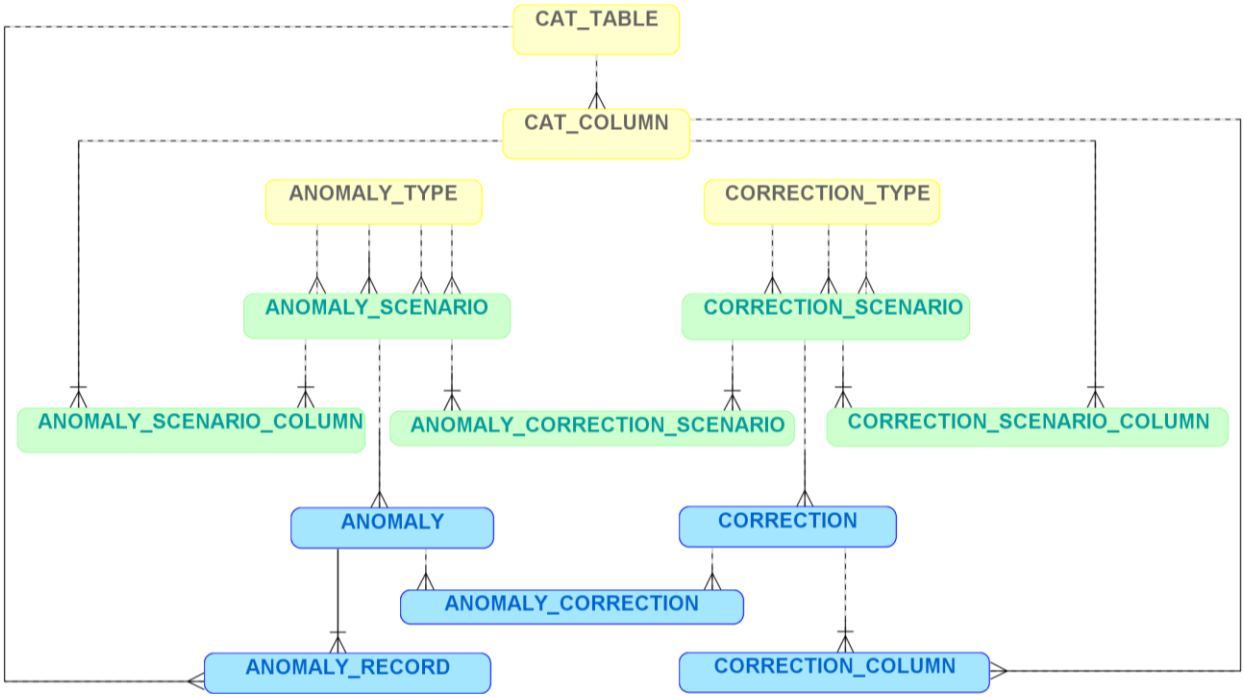


F. Braudel, « *temporalités étagées* » (1949)

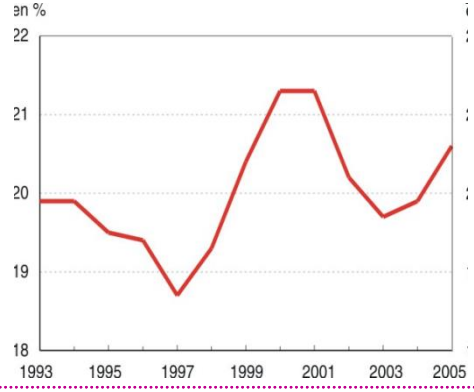
N. Elias, « *continuum évolutif* » (1996)

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 …



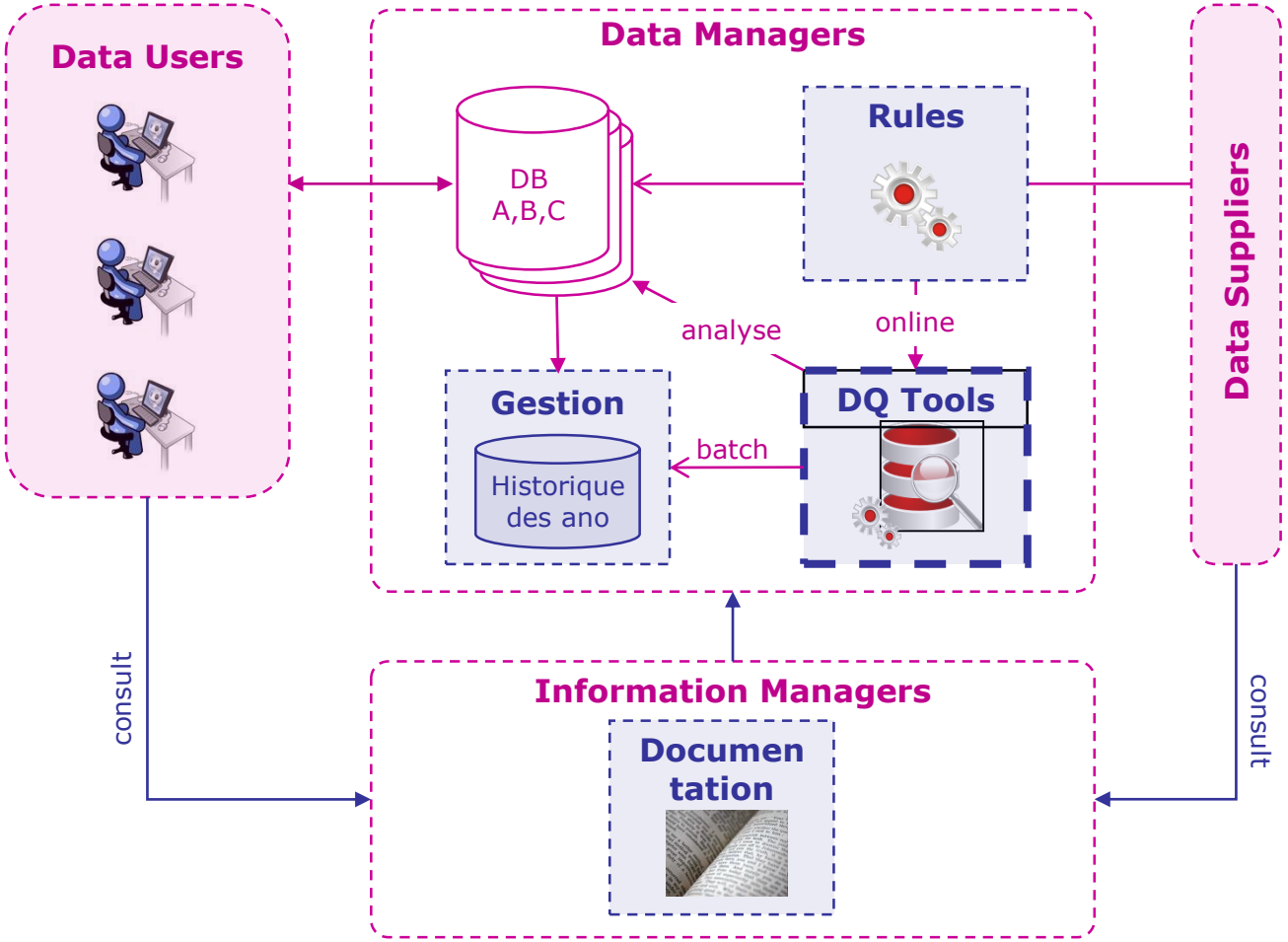
Stratégies de gestion



- Conception d'indicateurs de qualité :
  - Quantifier le temps et la nature des traitements (monitoring)
  - Identifier les cas d'anomalies fréquemment validées (anomalies « fictives ») et en évaluer les causes structurelles à travers les flux de traitement (obsolescence des contrôles ?) : "back tracking"
  - Adapter progressivement les contrôles à l'évolution des réalités et diminuer le nombre d'anomalies fictives

- Exemples d'applications concrètes :
  - DmfA : 65 milliards d'euros annuels de cotisations sociales prélevées et redistribuées
  - Déduction de cotisation pour les "bas salaires" et diminution structurelle du nombre d'anomalies de 50 % (14.000/7.000)
  - Le 2/2/2017: Arrêté Royal imposant l'application d'un "back tracking" à tous les SSA & prestataires de services en Belgique pour la sécurité sociale
  - Applications dans d'autres domaines empiriques : médecine, musées, corpus historiques, secteur bancaire, finances, …

Gani Hamiti,« Data Quality Tools : concepts and practical lessons from a vast operational environment »



Gani Hamiti, « Data Quality Tools : concepts and practical lessons from a vast operational environment »

After graduating from the Master in Information and Communication Science and Technology (Université libre de Bruxelles), Gani Hamiti has been working in the data quality field at Smals.

In this regard, he has been involved in various empirical data migration projects and in widely generalizable application areas such as social fraud detection or accounting information systems integration.

In parallel, a significant part of his work consists in spreading awareness and sharing knowledge about data quality.

Orientation bibliographique (1)

- Bade 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.
- Batini C. et Scannapieco M., éds., *Data and Information Quality. Dimensions, Principles and Techniques*. New York, Springer, 2016.
- Berten V. et Boydens I., *Email Address Reliability*, Section Recherches, Bruxelles, Smals, 2014.
- 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. (Prix de la Fondation L. Davin, conféré par l'Académie Royale des sciences, des lettres et des beaux-arts de Belgique, 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
- Boydens I., *Open Data et eGovernment*. Research Note, Bruxelles, Smals, n° 33, avril 2014, 23 pp. http://www.smalsresearch.be/download/research_reports/research_note/OpenDataRN.pdf

Orientation bibliographique (4)

- Boydens I., « *Data Quality & Back Tracking : depuis les premières expérimentations à la parution d'un Arrêté Royal* ». Bruxelles, Smals, Research Section, post de blog, 14/05/2018. https://www.smalsresearch.be/data-quality-back-tracking-depuis-les-premieres-experimentations-a-la-parution-dun-arrete-royal/
- 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 **:** Boydens I., *"Mapping the World of Data Problems" : la qualité des données vue par la communauté IT*. Bruxelles, Smals, Research Section, post de blog, 03/04/2013. http://www.smalsresearch.be/mapping-the-world-of-data-problems-la-qualite-des-donnees-vue-par-la-communaute-it-geek/
- Olson J., *Data Quality: The Accuracy Dimension*. Elsevier, The Morgan-Kaufmann Series in Database Management, 2003.

Orientation bibliographique (5)

- Redman T. C., *Data Quality for the Information Age*. Boston-London : Artech House Publishers, 1996.
- Redman T. C. *Getting in front on Data. Who Does What*. N. J., Technics Publications, 2016.
- 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., « Utiliser les déclarations administratives à des fins statistiques ». In *Le Courrier des statistiques*, Paris, INSEE, décembre 2018, n°1, p. 14-23.
  https://www.insee.fr/fr/information/3647013?sommaire=3647035
- 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.

Orientation bibliographique (6)

- 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, 2014.
- Volle M., *De l'informatique*. Paris, Economica, 2006.
- 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/

Orientation bibliographique (7)

Mémoires d'Etudiants de fin de Master depuis 2010 (1)

- DUPONT N., « *Les bases de données intra-site utilisant un système d'information géographique en archéologie : évolution et possibilités futures*» (ULB, mémoire de fin d'étude MA en STIC 2010-2011, dir. I. Boydens).
- SALLETS J., « *La problématique des adresses spatiales dans les bases de données administratives*» (ULB, mémoire de fin d'étude MA en STIC 2010-2011, dir. I Boydens). Prix de l'Association Belge de Documentation.
- DUTOIT F., « *La représentation de l'information empirique au sein des bases de données bibliographiques. Cas de l'apparat critique évolutif des œuvres d'attribution incertaine, avec exemplification sur la base d'un corpus de textes médiévaux* » (ULB, mémoire de fin d'étude MA en STIC, dir I. Boydens, année académique 2010-2011).
- Maroye L., « *La nouvelle norme ISO 25964-1 pour les thesaurus multilingues: apports et contraintes. Le cas particulier de TESE*. » (ULB, mémoire de fin d'étude MA en STIC, dir I. Boydens, année académique 2011-2012).
- De Roeck C., « *Analyse de la qualité d'un DataWarehouse au sein d'une institution gouvernementale. Etat de l'art critique, étude de cas, solutions.* » (ULB, mémoire de fin d'étude MA en STIC, dir I. Boydens, année académique 2013-2014).
- Meunier V. « *ISO 25964 et introduction de la distinction entre terme et concept : présupposés conceptuels et implications opérationnelles* » (ULB, mémoire de fin d'étude MA en STIC, dir I. Boydens, année académique 2013-2014).

Orientation bibliographique (8)

Mémoires d'Etudiants de fin de Master depuis 2010 (2)

- Paquot F. « *Etat de l'art critique de l'usage des Open Data à des fins journalistiques* ». (mémoire de fin d'étude MA en STIC, dir I. Boydens, année académique 2014-2015).
- Jacobs A., "*Transmitting Information Through the Pipeline Network: Reevaluating the Gas Explosions of San Bruno, Engelhart and Ghislenghien from the Perspective of Organizational, Conceptual or Information Management-Related Elements in the Pipeline Business.*" Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2015-2016. – Prix ABD-BVD 2017.
- Zombek L., « *La qualité des méta-données dans le domaine environnemental : les bases de données relatives à l'eau. Enjeux, état de l'art critique, étude de cas et recommandations* ». Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2015-2016.
- Hamiti G., « *E-government : potentiel d'amélioration de la performance des applications en ligne par l'utilisabilité. Etat de l'art critique et enseignements généralisables sur la base d'une étude de cas* ». Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2016-2017.
- De Oliviera Kempf H., « *La qualité de l'information au sein des bases de données alimentaires. Etat de l'art des problèmes et solutions et confrontation au cas de l'AFSCA* ». Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2016-2017.