

Groupe de contact FNRS

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

29 janvier 2020

“I want to talk to a HUMAN!” : impact de la qualité des bases de connaissances sur les agents conversationnels

par

Max De Wilde

Introduction

Isabelle Boydens



Smals
ICT for society

fnrs
LA LIBERTÉ DE CHERCHER



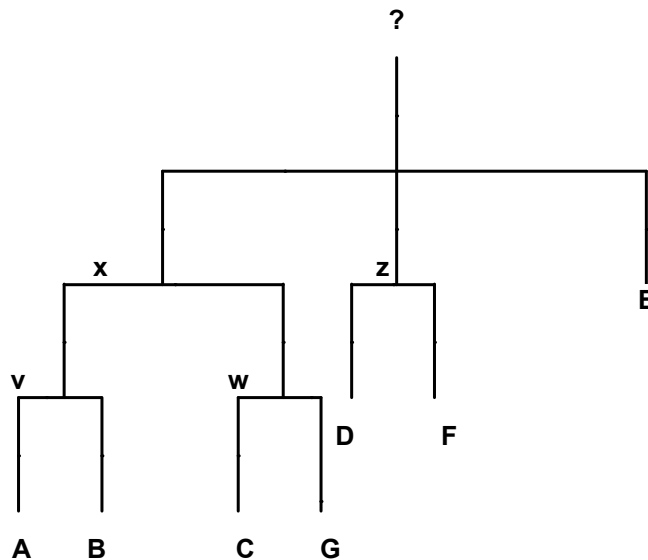
Groupe de contact FNRS

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

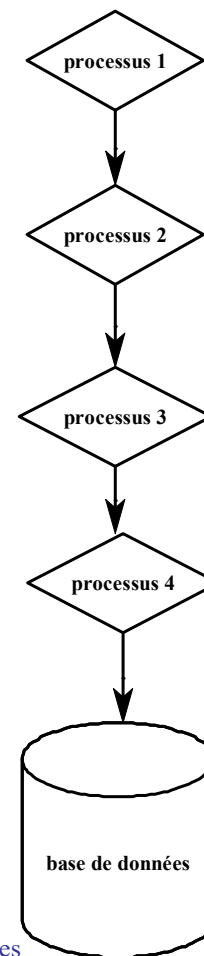
- Qualité des données :
 - « *fitness for use* » : adéquation des données à leurs objectifs (sociaux, industriels, juridiques, scientifiques, ...), sous *contrainte de budget*
 - pas de « *qualité totale* »
 - approche continue et pluridisciplinaire
- Groupe de contact FNRS créé en 1994 (25 ans en 2019)
 - <https://www.frs-fnrs.be/fr/financements/mobilite-fnrs/groupe-de-contact/91-nos-financements/mobilite-fnrs/groupe-de-contact/103-sciences-appliquees>
 - <https://www.frs-fnrs.be/fr/financements/mobilite-fnrs/groupe-de-contact/91-nos-financements/mobilite-fnrs/groupe-de-contact/104-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 »
<https://www.ulb.be/fr/programme/stic-b510-1>

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



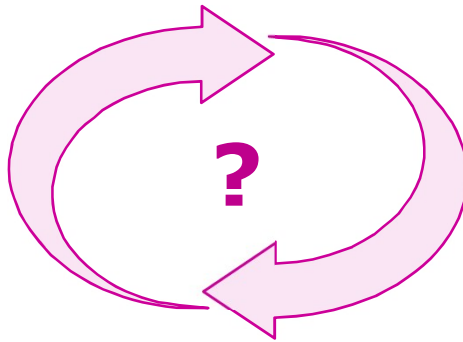
Enjeux stratégiques lorsque l'information est un instrument d'action sur le réel

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, USA : coûts de la « non-qualité » (3.100 milliards \$, 20% PIB USA, T. Redman)
- 2017, domaine des pipelines, de l'hydrologie, ...
- 2018-2019, bases de données médicales, en matière de terrorisme, Registre National, ...

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

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, ...)

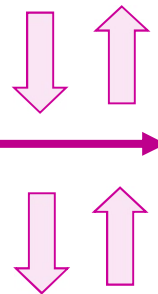
Comment les données se construisent-elles progressivement ?

Cadre d'analyse temporel

Évolution des normes

Évolution des représentations informatiques

Évolution du réel observable, objet de la norme



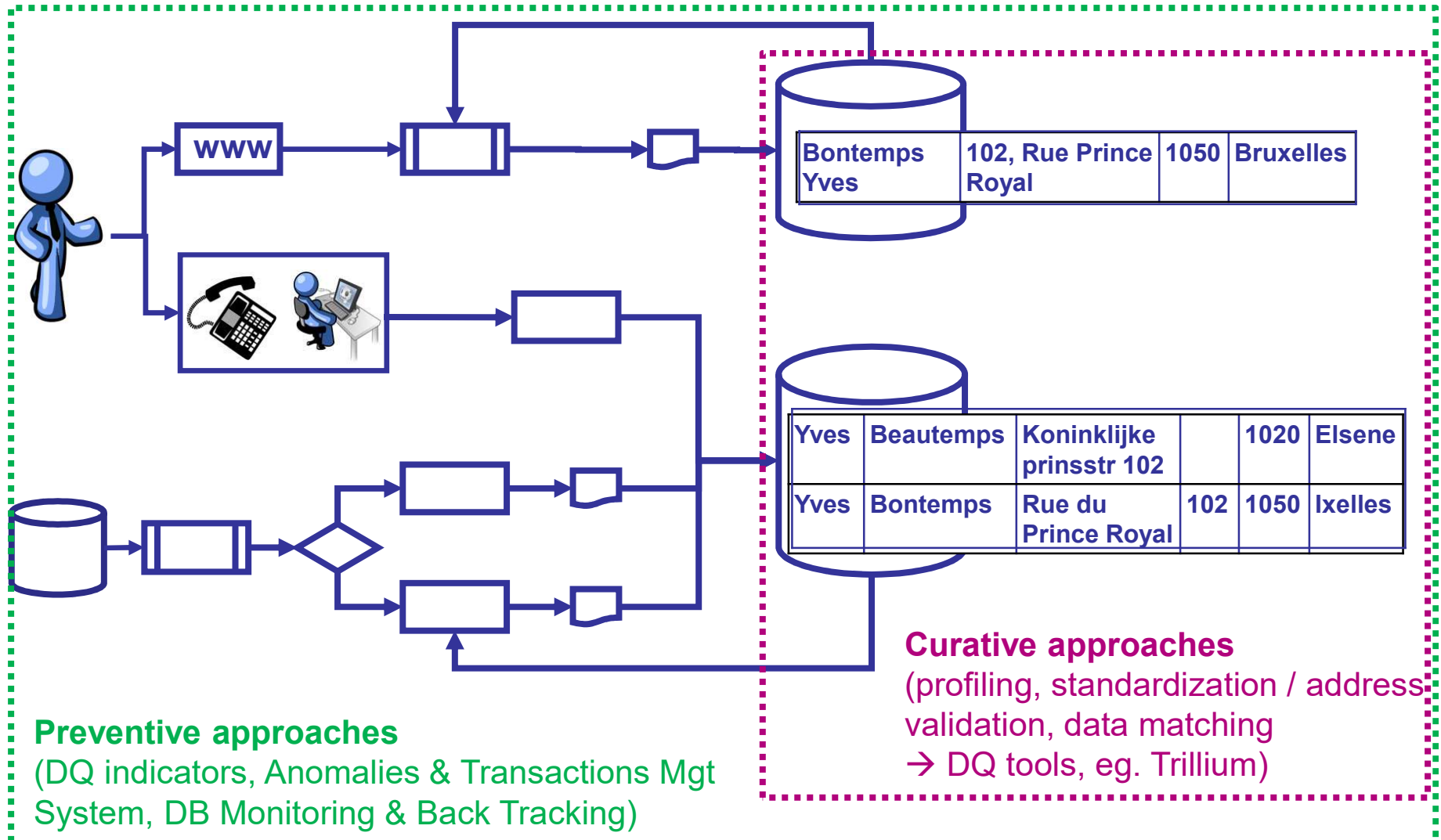
F. Braudel, « *temporalités étagées* » (1949)

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

*« Tout a été dit. Sans doute. Si les mots n'avaient changé
de sens; et les sens, de mots »*
Jean Paulhan (1966)

Evaluer et améliorer la qualité des données : deux approches complémentaires

First: identify business priorities, «fitness for use», budget and «cost-benefits»



Thématiques 2014 – 2019 (1)

- 2014

- « *Du stemma codicum au data tracking* » : vingt ans de recherche en matière d'évaluation et d'amélioration de la qualité des bases de données », par I. Boydens,
- « *Email Address Reliability*, » par V. Berten

Programme, slides et vidéo : <http://mastic.ulb.ac.be/2013/10/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique/>

- 2015 : « *Ouverture des données, gouvernance et engagement citoyen: questions de qualité* »

- « *Gouvernance et individus: réflexion sur un couple impossible* », Par F. de Smet
- « *Semantic Information Technology as a provider of tools for a productive and efficient organization of government* » par Agis Papantoniou

Programme et slides : <http://mastic.ulb.ac.be/2014/10/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-2/>

- 2016

- « *Analyzing the evolution history of data-intensive systems in support to software maintenance*», par A. Clève

Programme, slides et vidéo : <http://mastic.ulb.ac.be/2016/02/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-3/>

Thématiques 2014 – 2019 (2)

- 2017

- « *Temps réel, déterminisme et ordonnancement : trois défis majeurs des systèmes embarqués* », par J. Goossens

Programme, slides et vidéo : <http://mastic.ulb.ac.be/2017/02/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-4/>

- 2018

- « *La préservation du patrimoine scientifique à l'heure du numérique* », par A. Leroy

Programme, slides et vidéo : <http://mastic.ulb.ac.be/2017/09/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-5/>

- 2019

- « *Méthodes quantitatives, analyse critique de données et histoire médiévale : des données contemporaines aux sources anciennes* », par S. de Valeriola

Programme, slides : <http://mastic.ulb.ac.be/2018/08/reunion-du-groupe-de-contact-fnrs-analyse-critique-et-amelioration-de-la-qualite-de-linformation-numerique-6/>

Thématique 2020

"I want to talk to a HUMAN!" : impact de la qualité des bases de connaissances sur les agents conversationnels

par Max De Wilde

Max De Wilde

Max De Wilde est docteur en information et communication avec une spécialisation en linguistique computationnelle (Université libre de Bruxelles, 2015). En parallèle de son activité principale de consultant IT pour la Commission européenne, il est Maître de conférences à l'ULB où il dispense les cours de Traitement automatique de corpus et d'Ingénierie linguistique, ainsi que Chargé d'enseignement à l'Université de Genève où il enseigne les TIC aux étudiants en traduction

Programme 2020

- 14h Accueil (S. Van Hooland)
- 14h05 Introduction groupe de contact FNRS (I. Boydens)
- 14h15 *"I want to talk to a HUMAN!" : impact de la qualité des bases de connaissances sur les agents conversationnels* (M. De Wilde)
- 15h45 Table ronde (modérateur : G. Hamiti)
- 16h15 Réception

Les slides des présentations seront disponibles en ligne sur le site de l'invitation à la fin de la journée d'étude FNRS.

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., éd., *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., eds., *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., eds., *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., éd., "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*, Livrable, 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., éd., *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. éd., *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., "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/>
- 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/>
- Boydens I., « *Anomalies & Transactions Management System (ATMS) : enjeux, concepts, réalisations et travail en cours* ». Bruxelles, Smals, Research Section, post de blog, 23/04/2019. <https://www.smalsresearch.be/anomalies-transactions-management-system-atms-enjeux-concepts-realisations-et-travail-en-cours/>
- 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.).
- Hamiti G., *Data Quality Tools : concepts and practical lessons from a vast operational environment*. Cours-conférence, Université libre de Bruxelles, 13/03/2019. <http://mastic.ulb.ac.be/2019/01/cours-conference-data-quality-tools-concepts-and-practical-lessons-from-a-vast-operational-environment-par-gani-hamiti-13-mars-2019-a-18-heures/>
- 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.

Orientation bibliographique (5)

- McCallum Q. Ethan, *Bad Data Handbook, Mapping the World of Data Problems*. Sebastopol, O'Reilly Media, 2012.
- 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. *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.

Orientation bibliographique (6)

- 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. 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)

Thèses de doctorat dirigées et mémoires d'Etudiants de fin de Master depuis 2010 (1)

Thèses de doctorat

- VAN HOOLAND Seth, « *Metadata Quality in the Cultural Heritage Sector : Stakes, Problems and Solutions* », Faculté de Philosophie et Lettres, Département sciences de l'Information et de la Communication, thèse de doctorat sous la direction d'Isabelle BOYDENS défendue le 10/03/2009 à l'Université libre de Bruxelles.
- DE WILDE Max, *From Information Extraction to Knowledge Discovery: Semantic Enrichment of Multilingual Content with Linked Open Data*, doctoral thesis under the supervision of Isabelle BOYDENS and Pierrette BOUILLON (defended on 23 October 2015, Université libre de Bruxelles).
http://homepages.ulb.ac.be/~madewild/thesis_final.pdf

Mémoires de fin de Master

- 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.

Orientation bibliographique (8)

Thèses de doctorat dirigées et mémoires d'Etudiants de fin de Master depuis 2010 (2)

-
- 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).
 - 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).

Orientation bibliographique (9)

Thèses de doctorat dirigées et mémoires d'Etudiants de fin de Master depuis 2010 (3)

-
- 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 Helena, « *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.
 - Demeulder A.-C., « *Arbitrage entre ordre et désordre dans les systèmes collaboratifs en entreprise : entre la liberté humaine et la nécessité d'un contrôle* ». Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2017-2018.