



[Blog](#) [Talks](#) [Publications](#) ▾ [Tools](#) ▾ [Radar / Plan](#) ▾ [Team](#) [About](#)

Anomalies & Transactions Management System (ATMS) : enjeux, concepts, réalisations et travail en cours

Posted on 2019-04-23 by [Isabelle Boydens](#)

Cet article de blog a pour objet d'introduire le concept d'ATMS (*Anomalies & Transactions Management System*) :



- après en avoir montré l'importance fondamentale dans le cadre du « back tracking » récemment évoqué dans un [article de blog](#) de mai 2018, nous en rappelons les principales références ;
- nous en évoquons ensuite les concepts généralisables, le ROI, l'originalité ainsi que les premières implémentations d'envergure déjà appliquées à un DBMS hiérarchique dans le domaine de la sécurité sociale, modélisées et appliquées à d'autres domaines d'application empiriques ;
- pour terminer, nous introduisons une nouvelle initiative (*Proof of Concept*) en cours, en vue de généraliser l'approche aux standards et DBMS relationnels supportés par Smals dont la présentation fera l'objet d'un prochain blog du [Centre de Compétence en Qualité de Données](#) de Smals, reposant sur une synergie entre les équipes Databases et Recherche.

Contexte et enjeux, le « back tracking » : rappel

Dans le post « Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal » publié en mai 2018, nous faisons état d'une *approche préventive* dans le domaine de la qualité des données qui, en complément des *approches curatives*, tels les « Data Quality Tools », permet d'agir à la source et de résoudre structurellement un grand nombre problèmes de qualité (Figure 1).

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

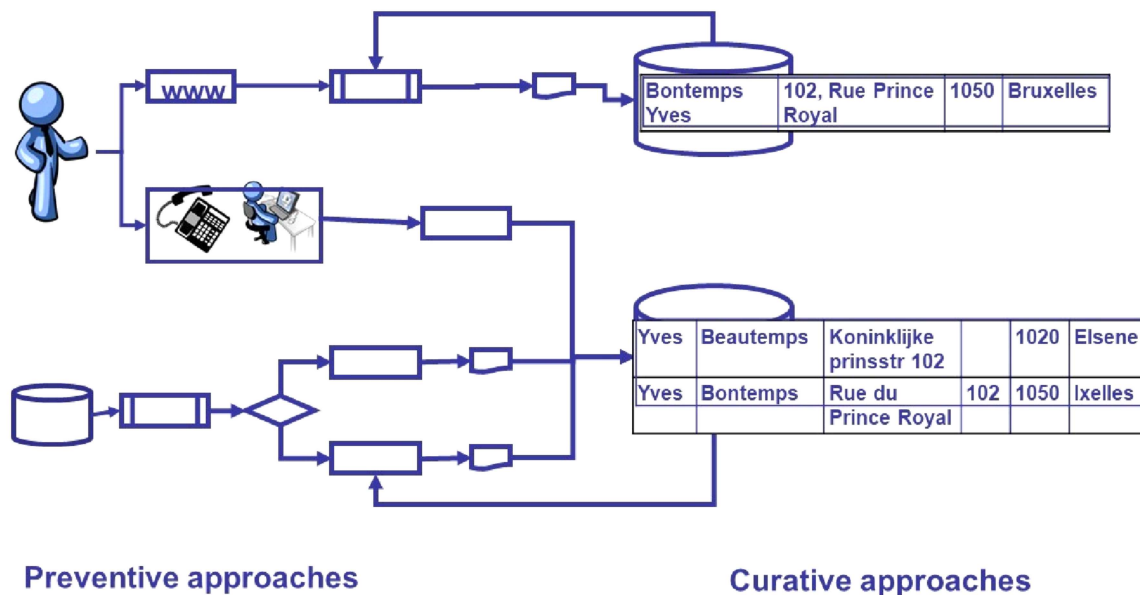


Figure 1. « Data Quality » : approches préventives et curatives

En raison du ROI de l'opération, celle-ci a été reconnue à grande échelle et a donné le jour à un Arrêté Royal paru le 2 février 2017. Celui-ci en généralise l'application à toute la Belgique dans le cadre des baromètres de qualité appliqués aux secrétariats sociaux agréés dans le cadre de la sécurité sociale.

La méthode est généralisable à tout domaine d'application empirique et revêt un intérêt d'autant plus important quand on connaît les enjeux de la « non qualité » des données, évalués ainsi par Thomas Redman en 2016 à l'échelle des USA dans son livre "Getting in front on Data" (p. 25) : "*\$3,1 Trillions/year in the US, which is about 20 percent of the Gross Domestic Product*".

A l'heure actuelle, l'approche innovante du « back tracking » développée sein du Centre de Compétence en Qualité de Données de Smals, continue de susciter l'intérêt à l'étranger, ainsi qu'en témoigne par

exemple un article publié par « Le Courrier des Statistiques » à Paris en décembre 2018 (1).

Comme nous l'avions noté dans le post mentionné plus haut, le « back tracking » repose sur l'extension de la base de données à un historique des présomptions de violations de contraintes d'intégrité (anomalies) et de leur traitement que nous désignons ici sous l'acronyme « ATMS », pour « *Anomalies & Transactions Management System* ».

En effet, seule une telle extension permet de:

- partir d'un monitoring complet des anomalies et traitements stratégiques, sans être soumis à l'inévitable marge d'erreur d'un échantillonnage ;
- recourir à la « stratégie du saumon », remontant à contre-courant les flux d'information (d'où, l'appellation « back tracking »), et ce faisant, réalisant un gain substantiel en temps et en manpower car :
 - les recherches des origines structurelles des anomalies prennent fin dès que toutes leurs causes par type ont été détectées, sans que tous les flux ne soient inutilement parcourus ;
 - l'ensemble des flux d'information, qui peuvent inclure des centaines de processus alimentant la base, ne doivent dès lors pas être examinés exhaustivement.

Nous proposons maintenant de présenter les concepts principaux d'un ATMS, dont nous avons initié et publié les fondations, assorties d'une application à grande échelle aux bases de données LATG et ensuite DmfA, au plan national, dont un livre primé par l'Académie Royale des sciences, des lettres et des beaux-arts de Belgique (2) et international, dont un livre co-édité chez Springer à New York (voir par exemple : 3, 4). David Bade, de l'Université de Chicago, a par ailleurs entièrement dédié un article référant à nos travaux en 2011 (5). Nous avons également produit un rapport d'étude sur la question au sein de la section Recherche, « gestion intégrée des anomalies » (6) et plusieurs mémoires de fin d'études dont nous avons assuré la direction à l'Université libre de Bruxelles ont dérivé de ce modèle généralisable des prototypes opérationnels applicables à divers domaines d'application empiriques (7).

ATMS : principaux concepts originaux, applications et travail en cours

Hypothèse du monde clos et catégories empiriques volatiles

Pour des raisons opérationnelles évidentes, le fonctionnement d'une base de données repose sur *l'hypothèse du monde clos*, en vertu de laquelle toute valeur non incluse dans le domaine de définition de la base est considérée comme fausse.

Toutefois, s'agissant des données empiriques, si l'on sort de ce cadre formel, il se peut qu'entre le moment où le schéma de la base de données a été formalisé et celui où l'information a été saisie, de nouvelles caractéristiques soient apparues au sein du domaine traité (contrairement à ce qu'affirment certaines théories postulant une relation bi-univoque permanente entre les données et le réel empirique correspondant).

Avec la mondialisation, de nouveaux cas non prévus initialement dans les tables de références pas plus que dans la législation peuvent par exemple apparaître dans nos bases de données nationales. Cela peut se produire dans le domaine énergétique, à un instant donné, pour ne prendre qu'un exemple parmi des milliers d'autres, à travers des unités d'exploitation d'entreprises étrangères (énergie géothermique s'agissant de l'énergie renouvelable, ...), par exemple (Figure 2).

Integrity test before entering a data: comparison with reference tables

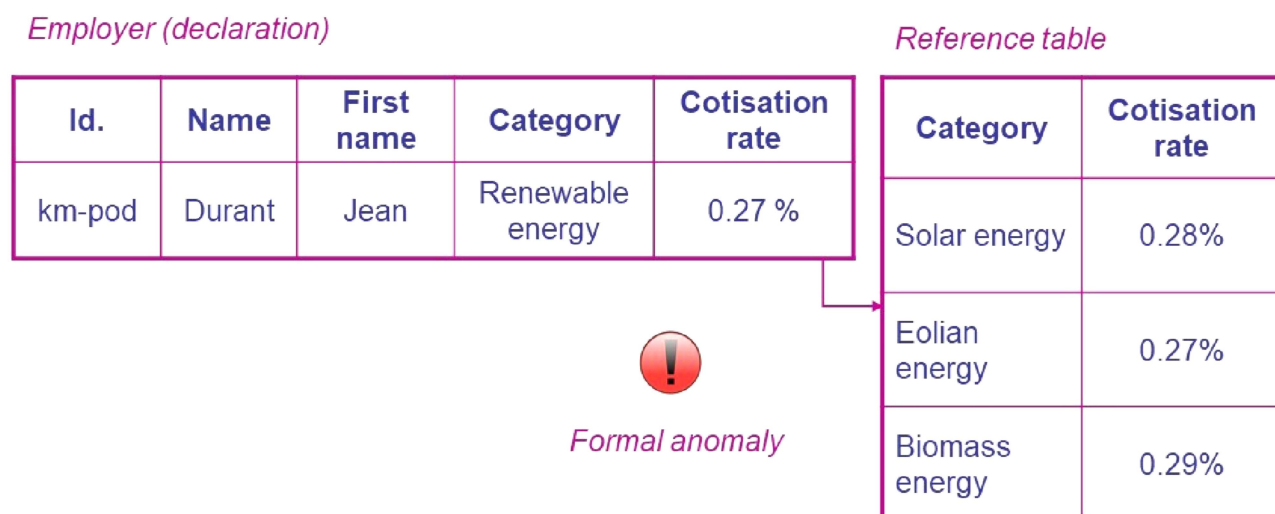


Figure 2. Violation de « l'hypothèse du monde clos » dans un domaine empirique

Dans ce cas, il est impossible de vérifier la correction des valeurs de la base de données de manière automatique. En effet, lorsqu'une incohérence apparaît entre une valeur saisie au sein de la base et les tables de référence permettant d'en tester la validité, il peut s'avérer indispensable, selon les enjeux, de procéder à une vérification manuelle, en contactant le citoyen ou l'entreprise concernée, par exemple.

En raison de cette volatilité et de la nécessité d'interpréter l'information, dans des contextes où les enjeux financiers sont importants (la DmfA permet le prélèvement et la redistribution de 65 milliards d'euros annuellement) ou au sein desquels la vie humaine peut être impactée, on ne dispose dès lors d'aucun référentiel formel "absolu" en vue de tester la correction d'une vaste base de données empirique.

Ainsi, en 2005, la catastrophe de l'ouragan Katrina a fait plus de 1.800 morts aux USA. A posteriori, on s'est rendu compte que les bases de données qui alimentaient les instruments de mesure censés alerter les citoyens afin qu'ils s'enfuient à temps de cette zone à risque n'avaient pas pris en considération l'évolution de phénomènes empiriques importants : la montée des eaux dans les océans suite au réchauffement climatique et la sur-construction, ne permettant plus l'écoulement rapide de l'eau dans les sols (8). De tels phénomènes sont encore fréquents dans le domaine hydrologique ou climatique.

Au fil de la prise en compte de l'évolution des réalités empiriques, qu'elles soient juridiques, empirico-formelles ou sociales, la signification en intensification des bases de données devrait idéalement évoluer dans le temps, si les enjeux le demandent. On peut conclure de ce mécanisme que les données ne sont pas données, les données se construisent progressivement avec l'interprétation des valeurs qu'elles permettent d'appréhender.

L'évolution de la norme, les transformations opérées au sein des bases de données, et des catégories observables sur le terrain sont solidaires. Solidaires, mais asynchrones. Elles opèrent, suivant leur nature, au sein d'échelles de temps différentes. On distingue par exemple "le temps long" des normes juridiques, renouvelées d'un trimestre ou d'une année sur

l'autre, le "temps intermédiaire" de la gestion des bases de données et le "temps court" du réel observable, celui des citoyens ou des entreprises assujettis à l'administration, dont l'évolution est continue.

Régulièrement, en effet, des entreprises fusionnent, se scindent, d'autres disparaissent alors que de nouvelles professions ou catégories d'activité non prises en compte dans les nomenclatures officielles voient progressivement le jour, avec la diversification des métiers de l'informatique par exemple. D'un point de vue dynamique, une base de données idéale devrait donc calquer le rythme de ses mises à jour sur la répartition – imprévisible – en "temporalités étagées" des évolutions de la réalité qu'elle appréhende. A ce qui ressemble à une gageure s'ajoute la nécessité, toujours révélée a posteriori, d'intégrer des observations imprévues, interdites *a priori* par l'hypothèse du monde clos.

Le caractère opératoire d'un ATMS, ses spécifications et son ROI

C'est là qu'intervient la nécessité d'une mise en place d'un ATMS. Le suivi quantitatif des anomalies et de leur traitement permet en effet le déploiement de stratégies de gestion de la base (2, 3, 4, 5, 6). Il est par exemple possible d'évaluer la rapidité de traitement des anomalies afin de déterminer le moment le plus opportun d'exploitation de la base de données.

Le suivi statistique des violations de contraintes d'intégrité ("anomalies formelles") permet de détecter non seulement les augmentations "anormales" (en fonction d'un seuil donné) d'anomalies mais aussi les augmentations de "validations" d'anomalies stratégiques lors de la phase de traitement. Une opération de validation signifie qu'après examen, un agent a estimé que l'anomalie, qui est une présomption d'erreur formelle, correspondait (« en réalité » ?) à une valeur pertinente. L'opérateur peut en effet "forcer" le système à accepter la valeur, grâce au système d'ATMS présenté ci-dessous et dont les modalités de formalisation et d'implémentation pourront varier. Si le taux de telles validations d'anomalies est élevé et récurrent et/ou si l'anomalie validée est stratégique (un cas inexpliqué à ce jour de rémission totale de la maladie du sida, par exemple, comme cela s'est présenté pour la seconde fois au niveau mondial en ce début 2019), la probabilité est grande que la structure de la base elle-même ne soit plus pertinente. Un algorithme peut

alors émettre un "signal" destiné aux gestionnaires de la base (IT et « Business ») afin qu'ils examinent si une modification structurelle de son schéma, voire une révision de la norme correspondante (législation, théorie, ...) sont requises.

En effet, lorsque les cas de validations sont importants, il est intéressant d'approfondir le phénomène : comme nous l'avons vu, un cas de figure inédit (l'émergence d'une nouvelle catégorie d'activité, d'une nouvelle maladie ou l'évolution d'un phénomène lié au réchauffement climatique) est peut-être apparu, ce qui requiert une adaptation de la structure de la base. Si l'on n'adapte pas le schéma, les anomalies correspondant à ces cas vont continuer d'apparaître en masse, nécessitant un examen manuel potentiellement conséquent et ralentissant considérablement le traitement des dossiers, ce qui affecte la qualité des données avec des impacts financiers ou vitaux pour l'homme.

Afin de soutenir cette approche, plusieurs prérequis conceptuels indispensables ont été proposés (2, 3, 4) et sont applicables si les enjeux et les budgets disponibles le justifient :

- un système de détection et d'enregistrement daté des anomalies jugées prioritaires lors de la saisie mais aussi "ex post" (en raison des interactions potentielles entre données, des anomalies formelles peuvent survenir a posteriori suite à la correction d'autres anomalies, par exemple) doit être mis en place et documenté.
- des procédures claires, soutenues par une équipe d'agents spécialisés, quant à leur traitement dans le temps (correction, validation, interprétation d'une valeur en l'absence de toute violation de contrainte d'intégrité, ...) doivent être établies, documentées et enregistrées. Ceci est fondamental, a fortiori lorsque la base de données s'inscrit dans un environnement fédéré et que plusieurs institutions sont respectivement responsables d'un sous-ensemble de l'information. Il s'agit de définir clairement quelle instance traite quelle partie de la base, à quel moment et quelles sont les modalités de traitement autorisées.
- la base de données doit être structurée de telle manière que l'historique des valeurs, anomalies et traitements puisse être enregistré dans le temps.

- enfin, une procédure claire de production d'indicateurs de qualité selon une périodicité donnée doit être définie et documentée sur cette base afin de réaliser du monitoring. Ce monitoring est destiné à supporter la gestion de la base de données face à l'évolution continue des réalités qu'elle traite et de son environnement conceptuel et humain à des fins opérationnelles.

Il va de soi que les paramètres des quatre points qui précèdent doivent pouvoir être maintenus dans le temps et sont susceptibles d'évoluer. La gestion des versions de schémas dans le temps est dès lors stratégique.

Anomalies & Transactions Management System : Organization

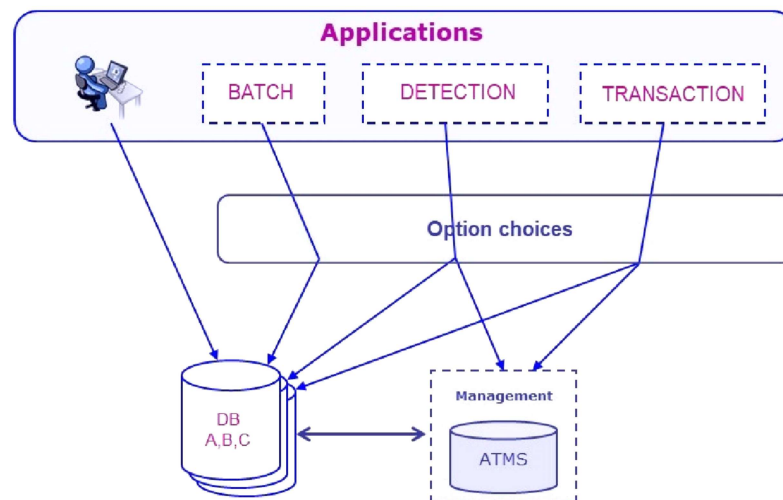


Figure 3. Fonctionnement schématisé d'un ATMS

La figure 3 illustre un schéma simplifié permettant de générer de tels indicateurs, tel qu'il a déjà pu être modélisé sous forme généralisable (2, 3, 4, 5, 6) et voir le jour en ce qui concerne le LATG ou la DmfA, en vue d'effectuer du monitoring de la base de données, de mener une opération de « back tracking » (2, 3, 4, 5) ou en tant que prototypes dans d'autres domaines d'application (7).

Dans le cadre de la sécurité sociale belge, dès la mise en place de ce système et, plus tard, avec la généralisation du « back tracking », cette méthode a permis d'améliorer la précision et la rapidité du traitement des réalisations sociales, réduisant potentiellement de plus de 50% le volume d'anomalies formelles prioritaires qui représentent chaque trimestre

environ 500.000 occurrences à gérer manuellement. La méthode est par ailleurs généralisable à tout domaine d'application empirique (7) : environnement, énergie,

Travaux en cours (Proof of concept)

Si les concepts constitutifs d'un ATMS que nous avons résumés ici ont été posés, modélisés et appliqués à grande échelle, plusieurs méthodes de modélisation conceptuelle et d'implémentation se présentent en vue d'en assurer la mise en oeuvre.

A ce jour, ces concepts ont été modélisés dans un DBMS hiérarchique très performant, s'agissant du LATG et de la DmfA, impliquant la génération de code applicatif externe à la base de données, en vue de maintenir l'intégrité des données. Les modélisations relationnelles précitées reposent sur ce même principe implicite (2, 3, 4, 5, 6, 7).

A l'heure actuelle, le Centre de Compétence en Qualité de Données de Smals qui repose sur une synergie entre la section Databases (9) et la section Recherche a entrepris la mise en place d'un POC appliqué aux DBMS relationnels et aux standards associés de Smals.

Celui-ci repose sur une séparation « dynamique » entre la base de données en production et la base de données correspondante supportant l'ATMS, de sorte que l'intégrité des données ne doive pas être gérée via du code externe principalement mais le soit essentiellement via les standards en matière de RDBMS. Les travaux sont en cours et ont pour objet de permettre à la section Databases de livrer un nouveau service aux clients de Smals potentiellement intéressés sur la base des enjeux qu'ils rencontrent et du budget dont ils disposent. Le "*proof of concept*" inclura des données réelles issues de bases de données stratégiques. Les résultats de ces travaux seront présentés lors d'un prochain blog complémentaire à celui-ci.

REFERENCES

1. RIVIERE P., Utiliser les déclarations administratives à des fins statistiques. In *Le Courrier des statistiques*, Paris, INSEE, décembre 2018, n°1, p. 14-23.

2. BOYDENS I., *Informatique, normes et temps*. Bruxelles : Bruylant, 1999, 570 p. (Prix de l'Académie Royale des sciences, des lettres et des beaux-arts de Belgique, 1999).
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., édés., « Practical Studies in E-Government : Best Practices from Around the World », New York, Springer, 2011, p. 113-130 (chapitre 7).
4. BOYDENS I., « L'océan des données et le canal des normes ». In Carrieu-Costa M.-J., Bryden A. et Couveinhes P. édés, 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.
5. 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, p. 328-338.
6. BOYDENS I., HULSTAERT A. et VAN DROMME D., « Gestion intégrée des anomalies – Evaluer et améliorer la qualité des données », Rapport d'étude, Section Recherches, Bruxelles, Smals, 2011.
7. Voir par exemple : DUPONT N., « Les bases de données intra-site utilisant un système d'information géographique en archéologie : évolution et possibilités future s ». Bruxelles, ULB, mémoire Master en STIC (dir. I. Boydens), année académique 2010-2011. 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 ». Bruxelles, ULB, mémoire Master en STIC (dir. I. Boydens), année académique 2010-2011. 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. ». Bruxelles, ULB, mémoire Master 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 ». Bruxelles, ULB, mémoire Master en STIC (dir. I. Boydens), année académique 2014-2015. ZOMBEEK L., « La qualité des méta-données dans le domaine environnemental : les bases de données relatives à l'eau. Enjeux, éta

de l'art critique, étude de cas et recommandations ». Bruxelles, ULB, Mémoire Master STIC (dir : I. Boydens), année académique 2015-2016. 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.

8. 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.
9. HAMITI G., « Data Quality Tools : concepts and practical lessons from a vast operational environment ». Cours-conférence, Université libre de Bruxelles, 13/03/2019.

Ce post est une contribution individuelle d'Isabelle Boydens, Data Quality Expert chez Smals Research. Cet article est écrit en son nom propre et n'impacte en rien le point de vue de Smals.

data quality

information management

MORE POSTS

Je data beschermen tegen beheerders: 'on-premise' Confidential Computing

2026-03

Protéger ses données des administrateurs : l'informatique confidentielle « on-premise »

2026-03

De performance van LLM's: Een vergelijkende analyse tussen Frans en Nederlands

2026-03

Made by Smals Research – Privacyvriendelijk Kruisen van Persoonsgegevens

2026-02

Search

Search

Newsletter & webinars:

Dutch French

Subscribe

Keywords:

[analytics](#) [artificial intelligence](#) [big data](#) [blockchain](#) [bpm](#) [chatbot](#)
[cloud computing](#) [cost cutting](#) [cryptography](#) [data center](#)
[data quality](#) [development](#) [eda](#) [egov](#) [event](#) [gis](#)
[information management](#) [machine learning](#) [managing it costs](#)
[methodology](#) [mobile](#) [natural language processing](#) [open source](#) [privacy](#)
[productivity](#) [security](#) [social](#) [software design](#)
[software engineering](#) [standards](#)

nals Research

