



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

Au coeur de la "data governance": les "data catalogs" ou systèmes de méta-information

Posted on 2025-03-19 by [Isabelle Boydens](#)

Nederlandstalige versie

Avec l'émergence et la complexité croissantes des applications informatiques, la documentation des données et des programmes est plus que jamais vitale, pour une bonne « **data governance** » quel que soit le secteur d'activité concerné.

Au seuil des années 2000, nous avons contribué à la mise en place des glossaires de la sécurité sociale et avons suivi leur développement par la suite. Pour cette raison, les concepts de cet article de blog nous sont familiers car certaines fonctionnalités n'ont pas changé depuis les années 2000.

Après une définition du concept de « **Data Catalog** » ou « **système de méta-information** ». nous en présentons dans les grandes lignes l'organisation, les fonctions principales [1] et les bonnes pratiques. En conclusion, nous dégageons un ensemble de recommandations méthodologiques généralisables.

Les systèmes de méta-information ou « Data Catalog » : définition et stratégie de gestion

« Méta-information » est souvent défini comme « information sur l'information ». Nous retenons ici la définition suivante : un système de méta-information est un système documentaire automatisé destiné à écrire un ensemble d'informations ou de données et ce faisant, à les

interpréter en vue d'en faciliter la gestion. Le recours à de tels systèmes est stratégique lorsque l'information est un instrument d'action sur le réel[2].

La conception d'un système de méta-information s'inscrit dans le cadre d'une stratégie de gestion. Les coûts correspondants émanent des opérations d'analyse, de conception, de développement ou d'acquisition de software et de maintenance. Les bénéfices escomptés tiennent à une meilleure interprétation de l'information, à une réutilisation plus aisée des applications préexistantes, à une crédibilité accrue du système et à une diminution des coûts de gestion (corrections a posteriori de la base de données, réparation des préjudices dus à la diffusion de données incorrectes, ...)[3].

Les systèmes de méta-information ou « Data Catalog » : fonctions

Data Ingestion, Rôles, IAM, gestion des règles

Nous présentons successivement les fonctionnalités suivantes : rôles et impact, gestion de champs multilingues, gestion des versions, mise en place de mécanismes d'héritage, application du concept de WOPM (*Write Once Publish Many*), standards, Graph Databases, publication en tant que REST API, système de recherche multibases, déploiement d'un workflow de validation documentaire (incluant éventuellement du Machine Learning supervisé dans les Data Catalogs) et quelques mots sur les softwares.

Un Data Catalog doit être alimenté ou croisé automatiquement avec d'autres systèmes connexes : on appelle cela "data ingestion". Ainsi, quand les glossaires de la sécurité sociale documentant les échanges d'information entre l'ONSS et les organismes prestataires, d'une part et les employeurs ou secrétariats sociaux agréés, d'autre part, créés au seuil des années 2000, l'alimentation des premières informations de base alors structurées en Word, fut réalisée via un programme *PERL*. D'autres méthodes plus modernes existent à cette fin en fonction du contexte.

Un Data Catalog s'adresse à la fois aux informaticiens et responsables business en charge de la gestion des bases de données, par exemple via un portail accessible aux citoyens en charge de l'envoi des déclarations

électroniques à l'administration, l'objectif étant que tous travaillent sur une base commune. Il s'agit que les droits d'accès soient gérés via un IAM.

Ce système de méta-information vise à automatiser partiellement les procédures ultérieures de saisie, de traduction et de validation de la documentation, à en renforcer l'intégrité et à en gérer les versions au fil des modifications législatives. Il s'agit de modéliser *la connaissance et les processus qui l'engendrent* : ainsi, le dictionnaire inclut à la fois des informations descriptives (par exemple, le domaine de définition d'un champ) et fonctionnelles (par exemple, la spécification formelle des contrôles destinés à tester les déclarations entrantes). Par ailleurs, les schémas des messages échangés entre les citoyens et l'administration ou toute autre partie peuvent être générés à partir du Data Catalog.

Gestion de champs multilingues

La documentation technique doit être diffusée dans les différentes langues nationales. Il en est de même dans tout contexte supranational. Des tables contrôlées multilingues (validées par les traducteurs, les juristes et l'IT) permettent, lors de la saisie des définitions, d'intégrer l'information dans une langue et d'obtenir ses contreparties dans les autres langues. L'ensemble pourra être complété au niveau spécifique si nécessaire (voir plus loin : héritage). Ceci permet de minimiser la charge de travail manuel, d'accélérer le processus de saisie et de renforcer la cohérence de l'ensemble.

Gestion des versions

La gestion des versions est fondamentale dans le domaine administratif[3]. En effet, la législation évolue fréquemment et toutes ses versions successives doivent être conservées au moins durant la période de prescription (par exemple, lorsqu'il s'agit de traiter des arriérés, il est fondamental de pouvoir retrouver les définitions antérieures de la base de données, les déclarations enregistrées ayant un statut légal de « force probante », c'est-à-dire qu'elles peuvent faire office de « preuve » lors d'un litige devant les tribunaux). Il est dès lors crucial d'identifier précisément les modifications apportées à chaque nouvelle version par rapport à la précédente. Ce « delta » est par ailleurs diffusé sous format standard, de telle sorte que les modifications puissent être intégrées de

façon semi-structurée dans les applicatifs encadrant les bases de données. Chaque item spécifiant la description d'une donnée pour une version considérée renvoie à la fiche correspondante (dans la langue choisie par l'utilisateur) avec la mention des champs modifiés par rapport à la version immédiatement antérieure, en ce compris l'historique des documents supprimés.

Workflow de validation (et ML supervisé)

En raison des enjeux légaux, sociaux et financiers correspondants, chaque nouvelle version doit être validée par les informaticiens et les juristes concernés par celle-ci. En vue de structurer cette validation, un système de *workflow* guide le déploiement du data catalog. Celui-ci s'inscrit dans le cadre d'un planning annuel de mise à jour, spécifiant de façon rigoureuse les périodes de mise à jour, de validation, de mise en acceptation et de mise en production. Le workflow est « piloté » de façon centralisée par une équipe dédiée à cette tâche et se déploie sur un mode décentralisé dans le cadre de l'extranet de la sécurité sociale, par exemple (Figure 1). Lors de la création de chaque nouvelle version, l'historique des échanges entre les différents responsables est conservé, de façon à garder un suivi du processus d'interprétation. Une vue permet aux gestionnaires de suivre le nombre de « *fiats* » requis pour la publication d'une nouvelle version. Ceci permet d'avoir une vue sur différents data catalogs interconnectés.

Validation Workflow

Demands gérées par gestionnaires Glossaires

Submitted by: Isabel Sastre Cantano Role: to All SMALS-MVM on 13/01/2009 at 17:24:29

Fiat necessary:
 Yes No

Type of concerned document: Annex

Nouveau code anomalie : Incompatibilité avec le nombre de titres-services

Sector: All

Réponse analystes/ juristes

12 Nouveau code anomalie : Incompatibilité avec le nombre de titres-services [Isabel Sastre Cantano 13/01/2009]

- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Alain Savatte 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Christian Decker 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Eric Boursin 15/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Jean-François Hubeit 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Jean-Sébastien Bastin 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Karen De Boeck 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Maryam Haghdad Mofrad 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Michel De Pauw 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Nadine Capelle 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Sofie Steurbaut 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Stéphane Dereppe 14/01/2009)
- Fiat to Nouveau code anomalie : Incompatibilité avec le nombre de titres-services (Vincent Depouillon 14/01/2009)

Figure 1. Documentation des glossaires de la sécurité sociale : workflow IT et Business

A cela s'ajoutent actuellement des fonctions de ML supervisé avec intervention humaine pour valider les modifications de méta-données à partir des modifications de data (à condition que celles-ci aient préalablement été validées par les business rules des bases de données correspondantes pour éviter de générer une méta-donnée à partir d'une donnée incorrecte).

Héritage et réutilisation dans un contexte multilingue

Le système de méta-information est éventuellement destiné documenter plusieurs dizaines de bases de données administratives répertoriant de nombreux champs communs dont certaines caractéristiques sont identiques (format, par exemple) et d'autres, distinctes (caractère obligatoire ou facultatif d'un champ, par exemple). Un mécanisme d'héritage doit dès lors être mis en place.

L'héritage (Figure 2) se définit comme la relation entre une classe A générique (que nous appelons ici « *stéréotype* » ou vocabulaire commun peu évolutif) et l'ensemble de ses instances $\{a_1, a_2, \dots, a_n\}$, où les propriétés (p_1, p_2, \dots, p_k) de la classe A constituent un *sous-ensemble* des propriétés de chaque objet instantié à partir de la classe A. Lors de l'instantiation, ce sous-ensemble de propriétés génériques peut être complété par un autre sous-ensemble de propriétés spécifiques à chaque instance $(p_{1+pa1}, p_{2+pa2}, \dots, p_{k+pan})$. Ce mécanisme est applicable à un nombre arbitraire de niveaux « méta ».

Héritage : propagation des modifications

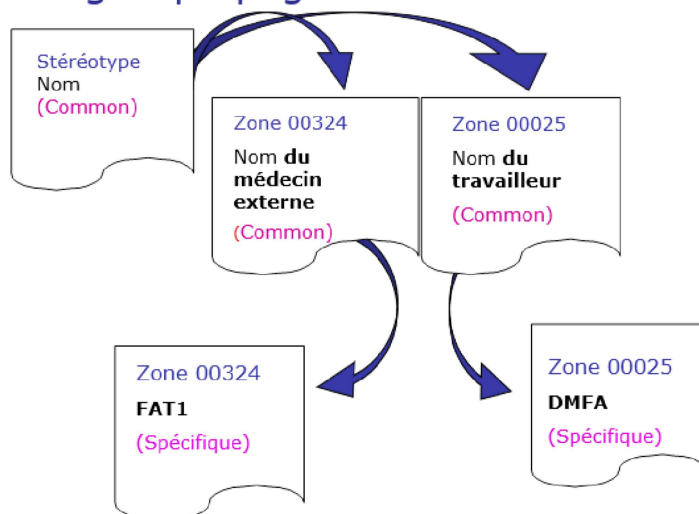


Figure 2. Documentation des glossaires de la sécurité sociale : principe de l'héritage

Les valeurs des propriétés génériques (« *nom* », « *domaine de définition* », « *description* », « *type* », « *longueur* ») du stéréotype « *numéro de compte* » sont ainsi stockées dans une table « *contrôlée* » de données structurées génériques prétraduites et prévalidées par les juristes et l'IT.

Les valeurs génériques et spécifiques sont ainsi concaténées en un champ semi-structuré. Ces fonctionnalités apportent des avantages en terme de temps de mise à jour (chaque valeur générique ne devant être encodée qu'une seule fois) et en terme de cohérence. Le système permet d'assurer que les données communes reçoivent les mêmes valeurs et d'éviter les erreurs humaines inhérentes à l'encodage manuel.

WOPM (*Write Once Publish Many*), Standards, Graph Database et publication sous forme de REST API

L'application inclut des listes structurées (codes postaux, catégories d'activité, ...) qui, dans la pratique, doivent être diffusées à des fins documentaires (dans l'esprit d'un "système de méta-information") mais aussi en vue de tester les déclarations envoyées par les citoyens et stockées dans les bases de données. Afin de rencontrer les deux fonctions, l'application doit être conçue dans l'optique du concept WOPM (« *Write Once Publish Many* ») de façon à générer automatiquement une même table structurée (liste de codes postaux, par exemple) sous différents formats : formats lisibles par l'humain et par la machine. La même source peut ainsi être utilisée au sein d'applications interdépendantes.

A l'heure actuelle, dans la mouvance du « Web sémantique », les normes en la matière sont devenues nombreuses. Les unes offrent des syntaxes génériques permettant le déploiement de méta-données, tel que DCAT, recommandation de l'UE. A ces normes, sur le plan technique, peuvent s'ajouter XML ou JSON, particulièrement utile pour la fusion de tables (Figure 4) et d'autres formats encore.

Une graph database (Figure 3) permet de visualiser l'état des relations entre différents « Data Catalogs » et pour ceux-ci, la part des méta-données complétées ou pas. En fonction de leur état plus ou moins

complet, on peut décider de la publication d'un "data catalog" sous forme de REST API au sein d'une institution (Figure 3).

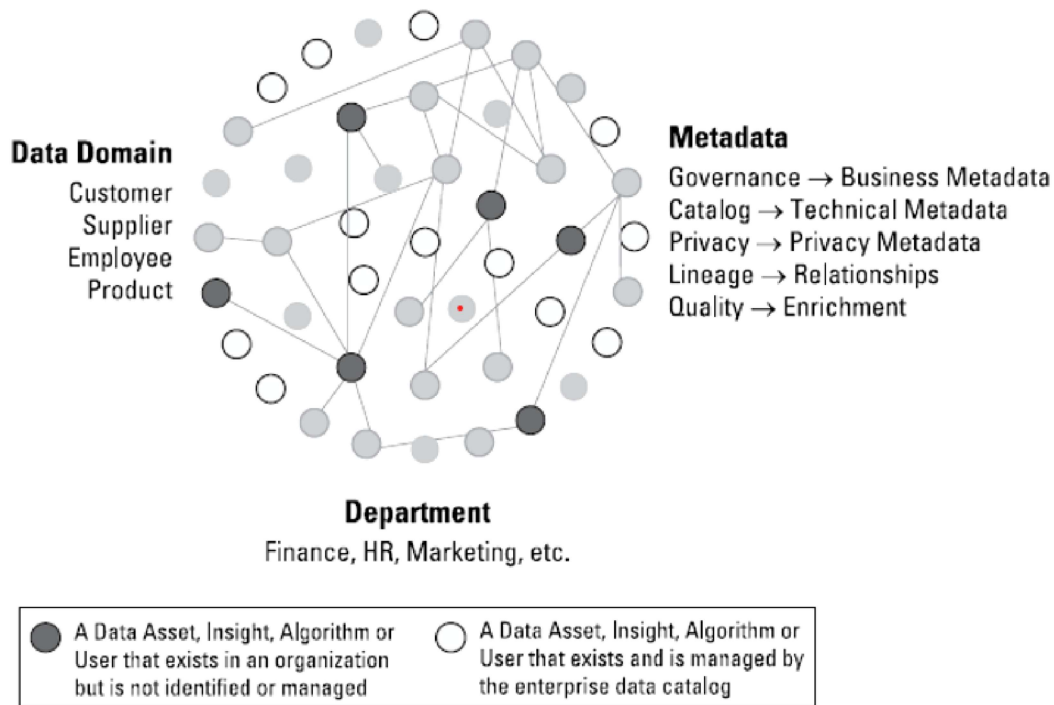


FIGURE 4-1: An active metadata graph.

Figure 3. intérêt d'une graph database pour suivre la complétude d'un data Catalog Source : [Collibra website](#)

Le Data Catalog peut être publié sous forme de REST API et accueillir lui-même d'autres REST API ou « pluggier » des logiciels commerciaux préexistants, certains standards comme JSON cité plus haut (Figure 4) favorisent ces liens (1).

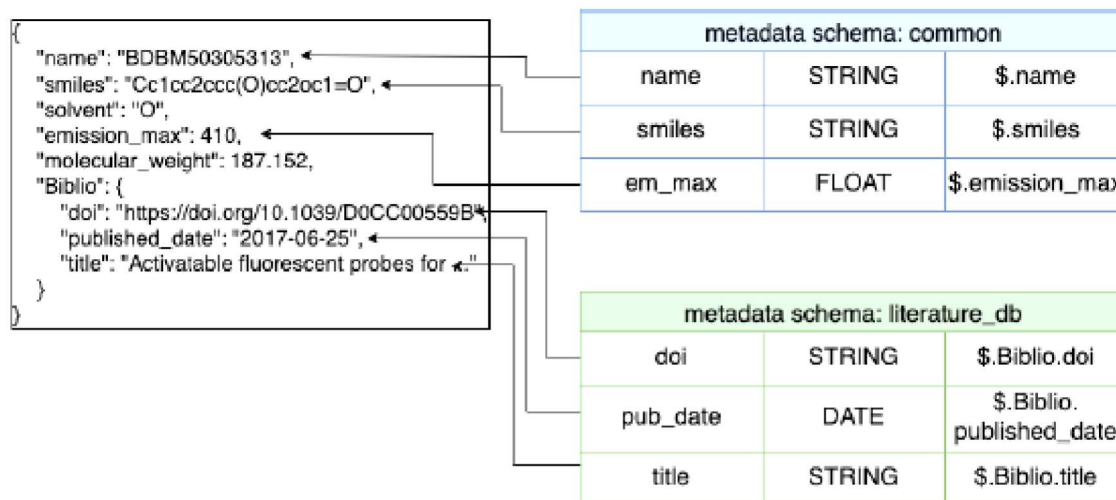


Figure 2: Example of mapping two metadata schemas to JSON properties in a data product's metadata. On the right are two metadata schemas, each with three fields. For each field, the first column is the field's name, the second is the data type, and the third is the JSON path to the field's value within the JSON metadata of a data product.

Figure 4. Exemple de mapping de 2 systèmes de méta-données via JSON (Source voir note 3)

Système de recherche multibases

Un outil de recherche «multibase » (Figure 5) doit être mis en place, permettant une recherche « full text » à travers le système documentaire intégré sur base de paramètres spécifiques avec recours à la logique booléenne de même que des systèmes de tri et de filtrage. L'output de l'outil de recherche peut se présenter sous différents formats en fonction des usages poursuivis (lisible par l'homme ou par la machine).

Search

GLOSSARIES
Select glossaries : (Empty means Any Glossary)

OPTIONS
Document type :

Search in version:

SEARCH
Search for the following word :

In fields : (Empty means Any Field)

RESULTS
Display for each doc the fields :

Figure 5. Exemple de recherches multibases, multilingues et multichamps avec options (source : glossaires de la sécurité sociale)

Evaluation continue et maintien de la qualité des données et des méta-données

Un maintien de la qualité des données et des méta-données est fondamental. Deux approches complémentaires existent. On peut travailler via un data quality tool complet afin de traiter les problèmes déjà présents dans les bases de données incluant les fonctions de profiling, standardization et matching (approche curative). On peut pour que les mêmes erreurs ne reviennent pas "ad infinitum" à la source, procéder via back tracking et ATMS (approche préventive), inventée au sein de Smals Research pour résoudre à la source les causes des problèmes de qualité (voir catalogue Reuse). Améliorer de manière continue la qualité des données et des méta-données correspondantes est crucial (voir le centre de compétence "data quality" sur le web site de Smals incluant des REST API sur le catalogue Software ReUse de Smals) (5).

Softwares

Au niveau **software**, outre des solutions « *home made* », comme les glossaires de la sécurité sociale auxquels plusieurs images de cet article de blog réfèrent, il existe des environnements de développement « open

source » comme Egeria demandant du développement, ou bien des outils commerciaux, comme Collibra, Altan, Infosphere, ...

Les systèmes de méta-information : recommandations méthodologiques

Les systèmes de méta-information comportent potentiellement trois écueils. Le premier est lié à ce que ces systèmes sont extensibles à l'infini., surtout lorsque les champs à compléter sont « libres », le langage naturel étant son propre méta-langage. Ceci implique des coûts importants en termes de gestion, lorsque les mises à jour manuelles sont nombreuses. Le second écueil tient à ce que les méta-données peuvent être elles-mêmes erronées et incertaines : lorsqu'elles sont d'ordre contextuel, leur validation ne peut faire l'objet de contraintes d'intégrité rigoureuses. Le troisième écueil tient au décalage temporel entre la mise à jour d'une donnée et de la méta-donnée correspondante, cette dernière, surtout lorsqu'elle se présente sous une forme textuelle, n'étant généralement créée qu'au terme d'une phase d'analyse.

Ainsi, dans une communication retentissante, "*The Metadata Myth...*"^[4], plusieurs auteurs évoquent les inextricables difficultés pratiques que soulève "l'usage abusif" des méta-données. Dans le domaine des bases de données géospatiales exploitées par le *Bureau of census* et la *National Aeronautics and Space Administration* (NASA), la mise en place d'un système de méta-information fédéral pour lequel chaque nouvel enregistrement nécessitait l'intégration d'environ 300 méta-données a entraîné les avatars suivants : coûts exorbitants en personnel et en ressources, lourdeur des mises à jour, ésotérisme de la documentation et finalement, réduction considérable de l'échange des données. Cependant, la NASA n'a pas abandonné ce système qui a toutefois fait l'objet d'une simplification et d'une restructuration.

Sur base des expériences en la matière, nous proposons les cinq recommandations suivantes :

- l'identification d'un ensemble minimal de méta-données obligatoires.
- une préférence pour les méta-informations générées automatiquement (ou sur base de listes de valeurs contrôlées par exemple) car ces informations sont moins "coûteuses" en termes de

mise à jour et plus fiables (cfr ML supervisé sous les conditions indiquées plus haut).

- la création de plusieurs niveaux de méta-données adaptés en fonction des usages (méta-données génériques et spécifiques, par exemple).
- La mise en place de liens directs entre les applicatifs documentés et les méta-données correspondantes (principe d'intégrité et de cohérence).
- Appliquer tout au long du cycle de vie du Data Catalog des KPI pour monitorer différentes métriques importantes, comme le taux de consultation des différentes parties du Data Catalog (6).

Au delà de l'application présentée dans cet article, ces recommandations s'appliquent à toute base de données empiriques dont l'interprétation est stratégique, en tant qu'instrument d'action sur le réel et donc, à tout « Data Catalog » .

Ce post est une contribution 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.

[1] O. Olesen-Bagneux, *The Enterprise Data Catalog :Improve Data Discovery, Ensure Data Governance, and Enable Innovation*. Boston, O'Reilly, 2023.

[2] « En mai 1999, pendant son intervention au Kosovo, l'Otan a bombardé par erreur l'ambassade de Chine à Belgrade : les bases de données cartographiques alors utilisées pour guider les missiles répertoriaient un plan de la ville obsolète et, donc, inadéquat" BOYDENS I., L'océan des données et le canal des normes. In CARRIEU-COSTA M.-J., BRYDEN A. et COUVEINHES P. édts, *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 ([lien vers l'article](#) – [sommaire du numéro 67 des Annales des Mines](#)).

[3] Marcus Christie, Suresh Marru, Sudhakar Pamidighantam, Isuru nawaka, and Dimuthu Wannipurage. 2023. *Airavata Data Catalog: A Multi-tenant Metadata Service for Efficient Data Discovery and Access*

Control. In Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA. ACM, New York, NY, USA [https://doi.org/ 10.1145/3569951.3597572](https://doi.org/10.1145/3569951.3597572)

[4] Foreman T. W., Wiggins H. V., Porter D.L., *Metadata Myth : Misunderstanding the Implications of Federal Metadata Standards. Proceedings of the First IEEE Metadata Conference*. Maryland : IEEE, 1996 (http://www.llnl.gov/liv_comp/metadata/ieee-md.4-96.html).

[5] 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](#)). BOYDENS I., HAMITI G. et VAN EECKHOUT R., A service at the heart of database quality. Presentation of an ATMS prototype. In *Le Courrier des statistiques*, Paris, INSEE, 2023, n°6, 11 p. (publié le 2/10/2023). [Lien vers l'article](#).

[6] [Asmae Boufassil](#); [Fadwa Bouhafer](#); [Mohamed Cherradi](#); [Anass El Haddadi](#), *Data Catalog: Approaches, Trends, and Future Directions*. In [17th International Conference on Signal-Image Technology & Internet-Based Systems \(SITIS\)](#), **IEEE** : 21 March 2024, **DOI: [10.1109/SITIS61268.2023.00067](https://doi.org/10.1109/SITIS61268.2023.00067)**

[data catalog](#)[data governance](#)[data quality](#)[egov](#)[information management](#)[metainformation system](#)

MORE POSTS

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

2026-03

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

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](#)

Smals Research

© Smals Research – License/Disclaimer: [FR](#) / [NL](#)