ACTIAN
a division of HCLSoftware

EBOOK

Qu'est-ce qu'un catalogue de données « intelligent » ?

Offrez à chaque utilisateur un catalogue conçu pour les entreprises modernes.

Sommaire

Résumé exécutif

L'idée d'un catalogue de données intelligent existe depuis plusieurs années dans la littérature spécialisée sur la gestion des métadonnées, même si elle ne fait l'objet d'aucune définition officielle. Un consensus général se dégage néanmoins : pour atteindre tout son potentiel, un catalogue de données moderne doit intégrer le machine learning et l'intelligence artificielle (IA).

Cet eBook présente l'approche d'Actian concernant le catalogue de données intelligent, qui ne peut se limiter à de simples capacités de machine learning. Avant d'expliquer ce que signifie réellement le terme « intelligent », il convient de définir ce qu'est un catalogue de données. Il s'agit d'un inventaire détaillé de tous les actifs de données disponibles dans l'entreprise, accompagné des métadonnées nécessaires pour les exploiter.

Un catalogue de données permet aux utilisateurs de localiser efficacement les données dont ils ont besoin, facilitant ainsi l'accès dans de nombreux cas d'usage. Il va donc bien au-delà d'un simple inventaire structuré : c'est un système opérationnel conçu pour accélérer les initiatives autour de la donnée. Ce catalogue doit répondre aux besoins de profils variés, comme les analystes, les data engineers, les professionnels de la conformité et des risques, les data scientists, les chefs de produit ou encore les responsables métiers. En résumé, il doit être conçu pour servir les utilisateurs finaux.

Faire de la donnée un actif stratégique

La reconnaissance de la donnée comme un véritable actif de l'entreprise est relativement récente, mais elle est de plus en plus adoptée par les entreprises innovantes. Dans un monde massivement numérisé, les organisations les plus performantes sont celles qui exploitent pleinement les volumes de données à leur disposition. Ces données peuvent être intégrées et analysées pour améliorer le positionnement de produits ou de services, conquérir de nouveaux marchés, ou encore répondre à de nombreux autres cas d'usage.

Donnée : l'actif invisible qui crée de la valeur

La donnée correspond parfaitement à la définition comptable d'un actif : une ressource de l'entreprise qui peut produire de la valeur ajoutée ou contribuer à son bon fonctionnement. À la différence des autres actifs, elle ne figure tout simplement pas dans le bilan comptable.

ACTIAN™ a division of HCLSoftware

La donnée ne devient un actif qu'à une seule condition : elle doit être exploitable. La recherche de cette exploitabilité a conduit la plupart des organisations à investir massivement, tant sur les aspects culturels, que techniques ou opérationnels. Et l'un des piliers d'un catalogue de données se trouve précisément dans la gestion des métadonnées.

L'importance des métadonnées dans l'exploitation de l'information

Il est important de garder à l'esprit que les données de l'entreprise ne sont, au fond, qu'un magma binaire (des zéros et des uns) totalement incompréhensible pour la majorité des gens. Prenons l'exemple des données utilisées au quotidien sur les ordinateurs ou les appareils mobiles. Sur le plan physique, rien de surprenant : ces données sont une suite de 0 et de 1 enregistrée sur un disque dur ou un autre support de stockage. Le propriétaire de l'appareil n'accède jamais directement à ces données brutes.

Une série de logiciels va leur fournir les informations nécessaires pour qu'eux-mêmes (ou un autre logiciel) puissent exploiter ces données :

- Un contrôleur conserve les informations relatives à l'emplacement physique des bits qui composent un jeu de données élémentaire, généralement sous forme de fichier.
- Un système de fichiers organise ces jeux de données de manière logique et gère des informations essentielles comme les répertoires, les noms, les extensions, pour donner un sens aux fichiers. Il s'occupe aussi de la sécurité : propriétaires, autorisations, dates de création ou de modification, etc.
- Un explorateur de fichiers dédié utilise ces informations pour permettre aux utilisateurs de consulter et comprendre le contenu : exploration des arborescences, prévisualisation, recherche, association des extensions aux applications.

Toutes les informations gérées par ces différents composants sont appelées métadonnées (littéralement, des données sur les données), et elles sont indispensables pour donner du sens au contenu d'un jeu de données au sein d'une organisation. L'intervention humaine est très limitée. Il suffit de donner un nom de fichier et un emplacement, et les métadonnées sont ensuite gérées automatiquement.

Qu'est-ce qu'une métadonnée ?

Une métadonnée est ce qui transforme du contenu binaire en information exploitable. Le but du catalogue de données est de regrouper les métadonnées issues de tous les jeux de données disponibles et de les présenter aux utilisateurs de la manière la plus simple et directe possible.

Le vrai défi apparaît quand il faut changer d'échelle. Passer d'un système de fichiers à un système d'information est loin d'être simple.

L'objectif du catalogue de données intelligent

L'objectif d'un catalogue de données intelligent est de consolider efficacement une quantité massive d'informations.

Un système d'information, quelle que soit sa taille, regroupe généralement plusieurs dizaines de systèmes et d'applications qui stockent des données provenant d'une grande variété de sources. Ces sources peuvent inclure des bases de données relationnelles ou non relationnelles, des systèmes de fichiers distribués, des API, des solutions cloud, etc., selon des protocoles, formats et règles spécifiques.

Chaque système gère des centaines, voire des milliers de jeux de données (généralement des tables ou des fichiers), eux-mêmes constitués de dizaines de champs ou de colonnes. Chaque jeu de données et chaque champ alimentent un métamodèle : un ensemble structuré de métadonnées qui permet l'exploration des données.

Un métamodèle d'entreprise est bien plus sophistiqué qu'un simple fichier propre à un système donné. Il peut couvrir un large éventail d'aspects :

- **Métadonnées techniques.** Quels outils utiliser pour accéder aux données, selon les protocoles, les autorisations, les formats, les types, etc.
- **Métadonnées sémantiques.** Quelles informations opérationnelles contient le jeu de données, et quelles règles métier s'y appliquent.
- **Métadonnées organisationnelles.** Qui possède les données, qui les produit et les contrôle, et comment elles sont classifiées.
- **Métadonnées d'usage.** Où et comment sont utilisées les données, quelle est leur qualité, comment sont-elles surveillées.
- **Métadonnées de conformité.** Quelles règles internes et réglementations doivent être respectées pour exploiter ces données.

Au final, un catalogue de données exploite une quantité énorme d'informations très diverses, et ce volume ne cesse de croître de manière exponentielle tout comme le volume des données exploitables. Ce volume d'informations soulève deux grandes questions :

1. Comment alimenter et maintenir ce volume sans tripler (ou plus) le coût de gestion des métadonnées ?
2. Comment trouver les jeux de données les plus pertinents pour un cas d'usage donné ?

Un catalogue de données peut être "smart" dans cinq domaines clés :

1. L'inventaire de données
2. La gestion des métadonnées
3. Le moteur de recherche
4. L'expérience utilisateur
5. Les enseignements à tirer

Pour répondre à ces enjeux, le catalogue de données doit être intelligent. Cela signifie bien plus que l'intégration d'algorithmes d'IA : un catalogue intelligent repose sur un ensemble de fonctionnalités technologiques et conceptuelles intelligentes.

La métamodélisation dans un catalogue de données intelligent

En una empresa, los metadatos necesarios para aprovechar À l'échelle d'une l'entreprise, les métadonnées nécessaires pour exploiter les actifs de données peuvent être considérables. Au-delà d'une fine couche essentiellement technique, les métadonnées sont propres à chaque organisation, et parfois même spécifiques à chaque département. Par exemple, un analyste métier ne recherchera pas les mêmes informations qu'un data engineer ou qu'un chef de produit.

Il est important de souligner qu'un métamodèle universel et figé ne peut pas être intelligent. Chercher à créer un modèle unique n'est donc pas une bonne pratique. En effet, un tel métamodèle devrait pouvoir s'adapter à une infinité de situations, ce qui le condamne inévitablement à l'un de ces trois écueils :

1. Une simplicité excessive, incapable de couvrir l'ensemble des cas d'usage.
2. Un niveau d'abstraction trop élevé, adaptable à différents contextes mais nécessitant une formation complexe et chronophage, ce qui n'est pas souhaitable pour un déploiement à grande échelle.
3. Des abstractions peu structurées, qui finissent par générer de nombreux concepts concrets issus de combinaisons confuses et de contextes variés. Résultat : un métamodèle inutilement complexe, voire incompréhensible.complicado y potencialmente incomprensible.

L'adaptabilité est essentielle

Une métamodélisation intelligente repose sur un métamodèle capable de s'adapter à n'importe quel contexte et de s'enrichir à mesure que les cas d'usage et le niveau de maturité évoluent.

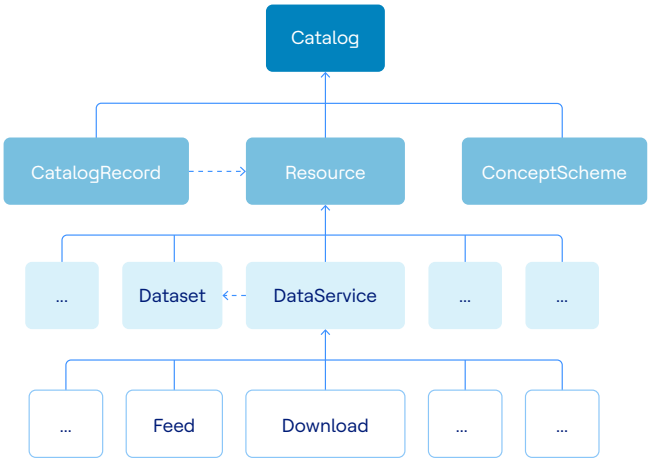


Figure 1 : exemple d'ontologie : structure formelle d'un modèle de connaissances

Une approche organique du métamodèle

Un métamodèle est un champ de connaissances. Sa structure formelle prend la forme d'une ontologie : un modèle qui définit un ensemble de classes d'objets, leurs attributs, ainsi que les relations entre eux. Dans un modèle universel, cette ontologie est figée : les classes, attributs et relations sont prédéfinis, avec des niveaux d'abstraction et de complexité variables (Figure 1).

Le catalogue de données Actian Data Intelligence Platform ne repose pas sur une ontologie statique, mais sur un graphe de connaissance évolutif. Le métamodèle est donc simple au départ. Il se limite à quelques types représentant les différentes classes d'actifs informationnels, comme les sources de données, les jeux de données, les champs ou les tableaux de bord. Chacune de ces classes possède des attributs essentiels comme le nom, la description ou les contacts (Figure 2).

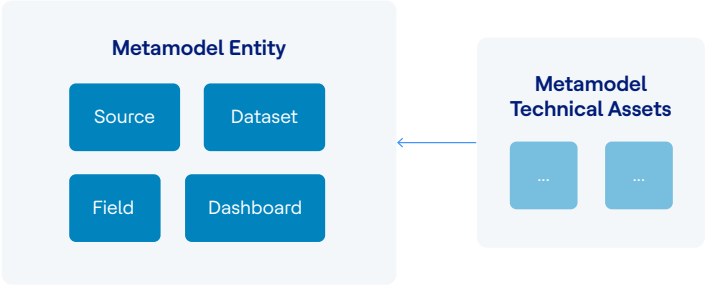


Figure 2 : Classes d'actifs informationnels

Ce métamodèle est alimenté automatiquement par les métadonnées techniques extraites des sources de données, qui varient selon les technologies utilisées. Par exemple, les métadonnées techniques d'une table dans un data warehouse ne sont pas les mêmes que celles d'un fichier stocké dans un data lake.

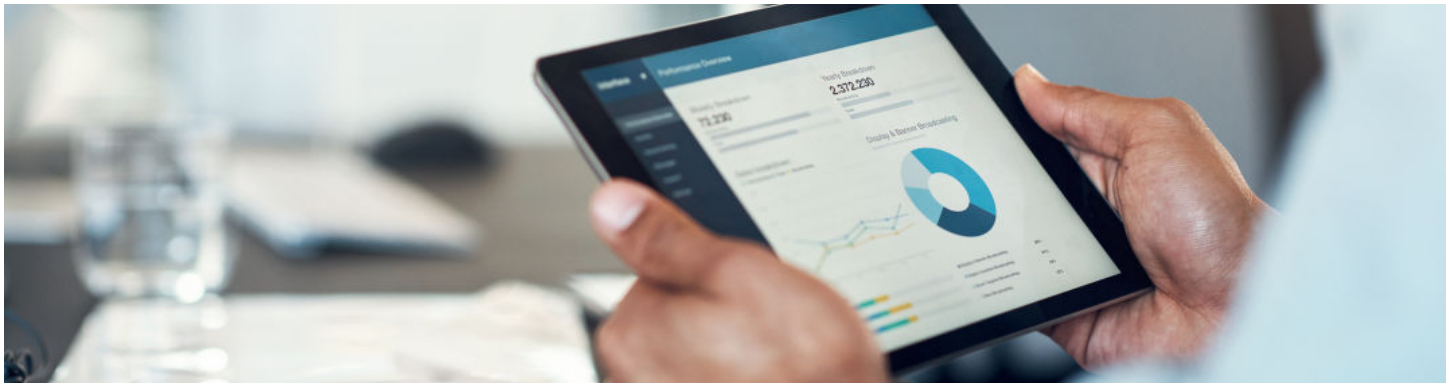
Il est possible de définir de nouvelles classes d'objets, ou d'ajouter des attributs aux classes existantes, ainsi que de préciser les relations entre les objets. L'ontologie n'est donc pas gravée dans le marbre : il est facile d'intégrer des évolutions du métamodèle de manière itérative et expérimentale. Cela rend l'extension progressive du catalogue à de nouveaux cas d'usage beaucoup plus fluide.

Le catalogue de données Actian Data Intelligence Platform repose sur un graphe de connaissance

La métamodélisation organique adoptée dans le catalogue de données Actian Data Intelligence Platform est la manière la plus intelligente de traiter la question de l'ontologie. Cette approche offre plusieurs avantages :

- Le métamodèle peut s'adapter à chaque contexte, souvent en s'appuyant sur un modèle déjà existant, et en intégrant la nomenclature et la terminologie internes de l'entreprise, sans nécessiter de courbe d'apprentissage longue et coûteuse.
- Le métamodèle n'a pas besoin d'être entièrement défini dès le départ. Les organisations peuvent se concentrer sur quelques classes d'objets et les attributs essentiels pour couvrir les premiers cas d'usage. Le modèle pourra ensuite être enrichi au fil de l'adoption du catalogue.
- Les retours des utilisateurs peuvent être intégrés progressivement, ce qui facilite l'adoption du catalogue et garantit, in fine, un meilleur retour sur investissement pour la gestion des métadonnées.

L'ajout d'attributs fonctionnels au métamodèle peut également faciliter la recherche des actifs. Cette approche de la métamodélisation présente donc de nombreux avantages. Elle comporte aussi un inconvénient majeur : comme le métamodèle est entièrement dynamique, le moteur a plus de mal à comprendre sa structure. Il lui est donc plus difficile d'assister les utilisateurs pour alimenter le catalogue et exploiter les données, ce qui constitue pourtant le cœur d'un catalogue de données intelligent.



Une partie de la solution réside dans la définition des attributs du métamodèle et de l'ontologie. En général, ces attributs sont définis par des types techniques : date, nombre, chaîne de caractères, liste de valeurs, etc.

Dans le catalogue de données Actian Data Intelligence Platform, ces bibliothèques de types incluent bien sûr ces types techniques, mais aussi des types fonctionnels comme les niveaux de qualité, les niveaux de confidentialité, etc. Ces types fonctionnels permettent au moteur de mieux comprendre l'ontologie, d'affiner les algorithmes, et d'adapter la manière dont l'information est représentée.

Comment l'inventaire des données soutient le catalogue

Une autre façon de rendre un catalogue de données « intelligent » passe par son inventaire. Un catalogue de données est, en essence, un inventaire complet des actifs informationnels, accompagné de nombreuses métadonnées qui permettent d'exploiter les données de manière aussi efficace que possible. La mise en place d'un catalogue de données repose donc, avant toute chose, sur un inventaire des actifs à l'échelle de l'organisation.

Les défis de l'automatisation de l'inventaire des données

Une approche déclarative pour construire l'inventaire, même bien pensée, n'est pas une bonne pratique. Elle demande un travail important pour lancer et maintenir le catalogue. Et dans un environnement numérique en constante évolution, cet effort initial devient rapidement obsolète.

La première étape logique pour créer un inventaire intelligent consiste donc à l'automatiser. Dans la grande majorité des cas, les jeux de données d'entreprise sont gérés par des spécialistes des systèmes, et s'appuient sur des solutions variées : systèmes de fichiers distribués, progiciels de gestion intégrés (ERP), bases de données relationnelles, applications métiers, entrepôts de données, etc.

Ces systèmes sont gérés avec toutes les métadonnées nécessaires à leur bon fonctionnement. Il n'est donc pas nécessaire de recréer manuellement ces informations : il suffit de se connecter aux différents registres et de synchroniser le contenu du catalogue avec les systèmes sources. En théorie, c'est simple. En pratique, c'est beaucoup plus complexe. Le fait est qu'il n'existe aucun standard universel auquel les différentes technologies se conforment pour permettre un accès homogène aux métadonnées.

Certains systèmes proposent des protocoles simples et bien documentés pour accéder aux métadonnées, tandis que d'autres nécessitent un travail plus approfondi. Même avec des bases de données relationnelles ou d'autres systèmes courants, les méthodes standard sont souvent décevantes.

Prenons par exemple l'API Java Database Connectivity (JDBC), qui permet de connecter des bases de données relationnelles en Java. Cette API propose des interfaces standards pour accéder aux métadonnées. En théorie, cela devrait permettre de récupérer facilement les informations de tous les systèmes disposant d'un driver, autrement dit, de pratiquement toutes les technologies utilisées aujourd'hui.

Mais en réalité, cette API standard ne fournit que des informations basiques. Les tables système fournies par ces solutions contiennent en revanche des métadonnées bien plus riches et complètes, et il est souvent préférable de les consulter directement.

Le rôle essentiel de la connectivité aux systèmes sources

Une couche de connectivité intelligente est un élément clé d'un catalogue de données intelligent. C'est pourquoi le catalogue de données Actian Data Intelligence Platform repose sur plusieurs caractéristiques majeures :

- **Propriétaire.** Le catalogue ne dépend pas de solutions tierces pour maintenir une extraction de métadonnées hautement spécialisée.
- **Distribué.** Cela permet d'étendre au maximum la portée du catalogue de données.
- **Ouvert.** Toute personne souhaitant enrichir le catalogue peut facilement développer ses propres connecteurs.
- **Universel.** Le catalogue peut synchroniser n'importe quelle source de métadonnées.

Cette couche de connectivité ne se contente pas de lire et synchroniser les métadonnées présentes dans les registres sources : elle est aussi capable de produire des métadonnées. Et produire des métadonnées demande plus qu'un simple accès aux registres, il faut aussi accéder aux données elles-mêmes, qui seront analysées par des scanners pour enrichir automatiquement le catalogue.

Le catalogue de données Actian Data Intelligence Platform produit ainsi deux types de métadonnées :

1. **Analyse statistique :** elle permet d'établir un profil des données (distribution des valeurs, taux de valeurs nulles, valeurs les plus fréquentes). Le type de métadonnées généré dépend du type natif des données analysées.
- **Analyse structurelle :** elle permet d'identifier la nature opérationnelle de certaines données textuelles, comme une adresse e-mail, une adresse postale, un numéro de sécurité sociale ou un code client. Le système est évolutif et personnalisable.

Le mécanisme d'inventaire doit lui aussi être intelligent

Le mécanisme d'inventaire du catalogue de données Actian Data Intelligence Platform est intelligent à plusieurs niveaux, au-delà de sa simple capacité à alimenter automatiquement le catalogue grâce à sa connectivité avec les différents systèmes. Voici quelques aspects clés :

- La **détection des jeux de données** repose sur une connaissance approfondie des structures de stockage, en particulier dans un contexte big data. Par exemple, un jeu de données issu de l'internet des objets (IoT), constitué de milliers de fichiers contenant des mesures en séries temporelles, peut être identifié comme un jeu de données unique, les fichiers eux-mêmes et leur emplacement n'étant alors que des métadonnées.
- **L'inventaire** n'est pas intégré par défaut dans le catalogue de données, afin d'éviter l'importation de jeux de données techniques ou temporaires, souvent peu exploitables ou redondants.
- **Le processus de sélection des actifs** à importer dans le catalogue bénéficie d'un accompagnement. C'est pourquoi une bonne pratique consiste à identifier en amont les objets les plus pertinents à intégrer dans le catalogue de données.

Tirer parti des bénéfices de la gestion des métadonnées

Le concept de catalogue de données intelligent (souvent associé à des algorithmes, au machine learning et à l'intelligence artificielle), relève avant tout du domaine de la gestion des métadonnées.

Comment la gestion des métadonnées est-elle automatisée ?

La gestion des métadonnées est une discipline qui consiste à valoriser les attributs du métamodèle pour les actifs de données inventoriés. La charge de travail nécessaire est généralement proportionnelle au nombre d'attributs dans le métamodèle et au nombre d'actifs dans le catalogue. Autrement dit, le volume de métadonnées à traiter peut vite devenir colossal.

Le rôle du catalogue de données intelligent est donc d'automatiser au maximum ces processus, ou, à défaut, d'aider les data stewards en améliorant leur productivité et en renforçant la fiabilité globale. Une couche de connectivité intelligente permet d'automatiser partiellement la gestion des métadonnées, mais cela reste souvent limité à une partie du métamodèle, principalement les métadonnées techniques. Un métamodèle complet, même modeste, peut contenir de nombreux attributs qui ne sont pas présents dans les registres des systèmes sources, tout simplement parce que ces informations n'ont jamais été enregistrées.

Par exemple, les métadonnées telles que les structures de tables, les noms de colonnes ou les types de données peuvent souvent être extraites automatiquement. D'autres types, en revanche, comme les définitions métier, les règles de qualité des données, ou les politiques d'usage, ne sont pas stockés nativement dans les systèmes sources.

Pour résoudre ce problème, les organisations peuvent adopter plusieurs approches. La plus directe consiste à identifier des modèles récurrents dans le catalogue pour suggérer des valeurs de métadonnées à appliquer aux nouveaux actifs.

Reconnaissance de motifs et empreintes numériques (fingerprinting)

Un motif regroupe l'ensemble des métadonnées associées à un actif de données, ainsi que les métadonnées liées à ses relations avec d'autres actifs ou entités du catalogue. La reconnaissance de ces motifs s'appuie généralement sur des algorithmes de machine learning.

La difficulté de cette approche réside dans la capacité à qualifier les actifs informationnels sous une forme numérique, afin d'alimenter les algorithmes et d'identifier les motifs pertinents. Une simple analyse structurelle ne suffit pas : deux jeux de données peuvent contenir les mêmes informations, mais dans des structures différentes.

S'appuyer sur l'identité des données n'est pas non plus efficace. Deux jeux de données peuvent avoir un contenu similaire mais avec des valeurs différentes. Par exemple, « Facturation client 2024 » dans un jeu de données, et « Facturation client 2025 » dans un autre.

Pour résoudre ce problème, le catalogue de données Actian Data Intelligence Platform utilise une technologie appelée *fingerprinting* (Figure 3).

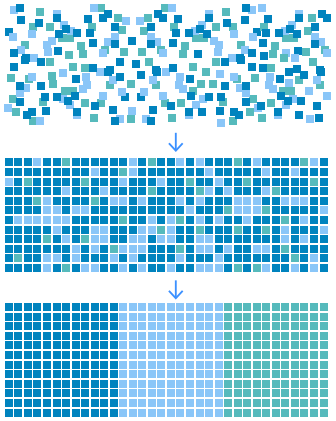


Figure 3 : le catalogue de données Actian Data Intelligence Platform utilise le fingerprinting

Le fingerprinting consiste à réduire un jeu de données (ou plus précisément un champ) à un vecteur numérique qui le décrit, appelé feature ou caractéristique. La création de cette empreinte repose sur deux types de caractéristiques extraites des données :

- Un ensemble de caractéristiques adaptées aux données numériques (essentiellement des indicateurs statistiques)
- Des données issues de modèles de vectorisation sémantique (word embedding) pour les données textuelles

Le fingerprinting est au cœur des algorithmes intelligents du catalogue de données Actian Data Intelligence Platform.

Autres approches intégrées dans un moteur de suggestions

La reconnaissance de motifs est une méthode efficace pour suggérer des métadonnées à associer à un nouvel actif dans un catalogue de données. Mais elle repose sur un prérequis essentiel : pour reconnaître un motif, il faut qu'il y en ait un à reconnaître. Autrement dit, cette méthode ne fonctionne que si le catalogue contient déjà un certain nombre d'actifs, ce qui n'est pas le cas au lancement d'un projet.

Or, c'est justement lors des phases initiales d'un projet de catalogue de données que la charge liée à la gestion des métadonnées est la plus élevée. Il est donc crucial de proposer d'autres approches pour accompagner les data stewards au début, quand le catalogue est encore peu fourni.

Le moteur de suggestions Actian Data Intelligence Platform, qui intègre des algorithmes intelligents pour faciliter la gestion des métadonnées, offre justement plusieurs approches complémentaires :

- **Détection de similarité structurelle.** Elle fonctionne lorsque plusieurs jeux de données ont des structures identiques, ce qui est courant dans les architectures de data lakes *en couches*.
- **Détection de similarité d'empreinte (fingerprint).** Elle ne repose pas sur la reconnaissance de motifs, mais sur un calcul direct de distance euclidienne entre les empreintes numériques de différents jeux de données. Deux ensembles aux contenus similaires auront des empreintes également proches.
- **Approximation de noms.** Il s'agit de construire dynamiquement un dictionnaire de noms techniques associé à des métadonnées spécifiques. Cette méthode fonctionne bien pour certaines associations types. C'est le cas, par exemple, dans la couche sémantique, où l'on pourrait proposer d'associer un champ nommé « *txt_email* » à une définition du glossaire intitulée « *Email* ».

Le moteur de suggestions, qui analyse le contenu du catalogue de données pour déterminer les valeurs probables des métadonnées associées aux actifs intégrés, est mis à jour en continu. De nouvelles approches y sont régulièrement ajoutées. Parfois très simples, parfois bien plus sophistiquées. L'architecture Actian Data Intelligence Platform permet ainsi d'améliorer les performances du moteur à mesure que le catalogue s'enrichit et que les algorithmes gagnent en maturité.

Le développement du moteur de suggestions est de nature expérimentale. Les utilisateurs identifient une piste prometteuse, la mettent en œuvre, en mesurent les performances, puis recommencent. C'est une démarche classique, mais qui soulève une question majeure : comment mesurer l'efficacité des algorithmes intelligents du catalogue de données ?

Une bonne pratique consiste à utiliser le lead time comme indicateur principal de productivité des data stewards. Le lead time, notion issue du lean management, représente ici le temps écoulé entre l'inventaire d'un actif et la valorisation complète de ses métadonnées.

Optimiser un moteur de recherche pour trouver rapidement les actifs

L'objectif principal d'un projet de catalogue de données est de trouver rapidement les actifs de données pertinents. Pour permettre une exploration efficace, il faut un moteur de recherche puissant.

Étant donné les volumes massifs de données contenus dans un catalogue d'entreprise, le moteur de recherche constitue le principal moyen d'exploration pour les utilisateurs. Il doit être simple d'utilisation, puissant, et surtout efficace : les résultats doivent être conformes aux attentes. Google et Amazon ont placé la barre très haut en la matière, et leurs expériences de recherche font désormais office de référence.

Une expérience de recherche optimale doit permettre :

- À l'utilisateur de taper quelques mots dans la barre de recherche, souvent avec l'aide d'un système de suggestions proposant des termes fréquemment associés pour affiner la requête.
- D'obtenir des résultats quasi instantanés, triés selon un ordre spécifique, avec le plus pertinent en tête de liste.
- De pouvoir ajouter facilement des mots pour affiner la recherche, ou d'utiliser des filtres disponibles pour écarter les résultats non pertinents.

De nombreuses solutions de catalogage de données sur le marché se limitent à l'indexation système, au scoring et au filtrage. Cette approche est satisfaisante lorsque l'utilisateur sait précisément ce qu'il cherche — ce qu'on appelle une recherche à forte intention. Mais elle se révèle souvent décevante dans les recherches plus exploratoires — à faible intention, voire sans intention du tout, lorsque l'objectif est simplement de suggérer spontanément des résultats pertinents à l'utilisateur.

En résumé, une indexation simple fonctionne bien pour retrouver des informations dont on connaît déjà les caractéristiques… mais elle atteint vite ses limites dans une recherche exploratoire. Les résultats comprennent alors souvent des faux positifs, et l'ordre d'affichage est surchargé de correspondances exactes, ce qui nuit à la pertinence globale.

Adopter une approche de recherche multidimensionnelle

Un système d'indexation simple présente des limites et ne permet pas d'obtenir des résultats véritablement pertinents pour les utilisateurs. C'est pourquoi le catalogue de données Actian Data Intelligence Platform isole le moteur de recherche dans un module dédié. Cela marque une rupture avec les moteurs classiques reposant sur une indexation plate de l'information.

L'approche adoptée par Actian Data Intelligence Platform s'inspire du célèbre algorithme PageRank de Google PageRank prend en compte plusieurs dizaines de facteurs, appelés features. Ces caractéristiques incluent la densité des relations entre différents objets du graphe, comme les liens hypertextes entre pages web, le traitement linguistique des termes recherchés, ou encore l'analyse sémantique du graphe de connaissance.

Plusieurs features sont intégrées au moteur de recherche d'Actian Data Intelligence Platform afin de garantir un haut niveau de pertinence des résultats, et ces fonctionnalités évoluent en permanence. Les principales sont les suivantes :

- Une indexation standard et plate de tous les attributs d'un objet (nom, description, propriétés), avec une pondération en fonction du type de propriété.
- Une couche de traitement du langage naturel (NLP), qui prend en compte les erreurs de frappe ou d'orthographe.
- Une couche d'analyse sémantique, basée sur le traitement du graphe de connaissance.
- Une couche de personnalisation, reposant sur une classification simple des utilisateurs selon leurs usages.

Un filtrage intelligent pour contextualiser et affiner les résultats de recherche

Pour compléter le moteur de recherche, un *système de filtrage intelligent* est intégré. Ce type de filtrage, bien connu sur les sites d'e-commerce, permet de proposer des filtres contextuels pour restreindre les résultats de recherche.

Ces filtres fonctionnent selon les principes suivants :

- Seules les propriétés réellement discriminantes (c'est-à-dire utiles pour réduire le nombre de résultats) sont proposées. Les propriétés non pertinentes ne s'affichent pas.
- Chaque filtre indique son impact, c'est-à-dire le nombre de résultats restants une fois appliqué.
- L'application d'un filtre actualise instantanément la liste des résultats.

Cette combinaison de recherche multidimensionnelle et de filtrage intelligent permet d'offrir une expérience de recherche nettement supérieure. L'architecture découplée d'Actian Data Intelligence Platform (où le moteur de recherche fonctionne comme un composant autonome) facilite l'exploration continue de nouvelles approches, rapidement intégrées dès qu'elles s'avèrent efficaces.

L'expérience utilisateur dans le catalogue de données

Elle repose sur l'identification des personas clés, la compréhension de leurs comportements et de leurs objectifs, puis la conception d'une interface graphique fluide et efficace répondant à leurs besoins. C'est précisément cette expérience qu'offre le catalogue de données Actian Data Intelligence Platform.

Offrir une expérience utilisateur conviviale et intuitive

Un catalogue de données doit être intelligent dans l'expérience qu'il propose aux différents types d'utilisateurs. L'un des principaux défis liés à son déploiement réside dans son adoption par les utilisateurs finaux, c'est-à-dire les consommateurs de données. L'expérience utilisateur joue un rôle central dans cette adoption.

Il est difficile de définir clairement les personas dans un catalogue de données. C'est un outil universel, qui apporte de la valeur à toute organisation, quelle que soit sa taille, son secteur ou sa localisation. Plutôt que d'essayer de modéliser des personas parfois flous ou mouvants, il est préférable de se concentrer sur l'adoption du catalogue de données.

Deux populations d'utilisateurs se démarquent :

- **Les producteurs de métadonnées.** Ils alimentent le catalogue et surveillent la qualité de son contenu. On les appelle généralement les data stewards.
- **Les consommateurs de métadonnées.** Ils utilisent le catalogue pour répondre à leurs besoins métiers. On les désigne souvent simplement comme des utilisateurs.

Ces deux groupes peuvent bien sûr se recouper : certains data stewards sont aussi des utilisateurs du catalogue.

Les défis de l'adoption à l'échelle de l'entreprise

La véritable valeur d'un catalogue de données se révèle lorsqu'il est largement adopté par un nombre important de consommateurs de données (et de métadonnées), et pas seulement par les data stewards ou les spécialistes de la gestion des données. Or, ce groupe d'utilisateurs est très hétérogène :

il inclut des experts data (data engineers, architectes, data analysts, data scientists), mais aussi des profils métier comme les chefs de projet, responsables d'unité ou chefs de produit. On y trouve également des responsables conformité et gestion des risques, et plus largement tous les managers opérationnels susceptibles d'exploiter la donnée pour améliorer leurs performances.

Les data stewards utilisent le catalogue de manière régulière, donc l'adoption n'est généralement pas un problème pour eux. Ils acceptent une courbe d'apprentissage, à condition que la solution soit conviviale et utile dans leur quotidien. Le véritable enjeu, c'est l'adoption par les utilisateurs.

L'adoption du catalogue de données par les utilisateurs est souvent lente, pour plusieurs raisons :

- **L'usage du catalogue est ponctuel.** Les utilisateurs ne s'y connectent qu'occasionnellement pour obtenir une réponse très précise à une question spécifique. Ils ont rarement le temps (ou l'envie) d'affronter une courbe d'apprentissage pour un outil qu'ils n'utiliseront que toutes les quelques semaines.
- **La perception des métadonnées varie selon les profils.** Certains utilisateurs se concentrent sur les aspects techniques, d'autres sur les enjeux sémantiques, d'autres encore sur les dimensions organisationnelles ou de gouvernance.
- **Tout le monde ne comprend pas le métamodèle,** ni l'organisation interne de l'information dans le catalogue. Résultat : certains utilisateurs se sentent vite perdus, noyés dans un flot de concepts perçus comme éloignés de leurs préoccupations quotidiennes.

Le catalogue de données intelligent lève ces obstacles pour accélérer l'adoption. Par exemple, Actian Data Intelligence Platform facilite l'adoption du catalogue de données en proposant une interface graphique familière, directement inspirée des plateformes d'e-commerce.

L'objectif est de réduire au maximum la courbe d'apprentissage. Concrètement, l'utilisateur doit pouvoir s'y retrouver sans aucune formation. Pour cela, Actian Data Intelligence Platform propose deux interfaces distinctes, adaptées aux différents profils (Figure 4) :

Studio. L'interface dédiée aux data stewards, conçue pour la gestion et le suivi du contenu du catalogue, un outil expert.

Explorer. Pensée pour les utilisateurs, elle offre une expérience de recherche et d'exploration aussi simple que possible.

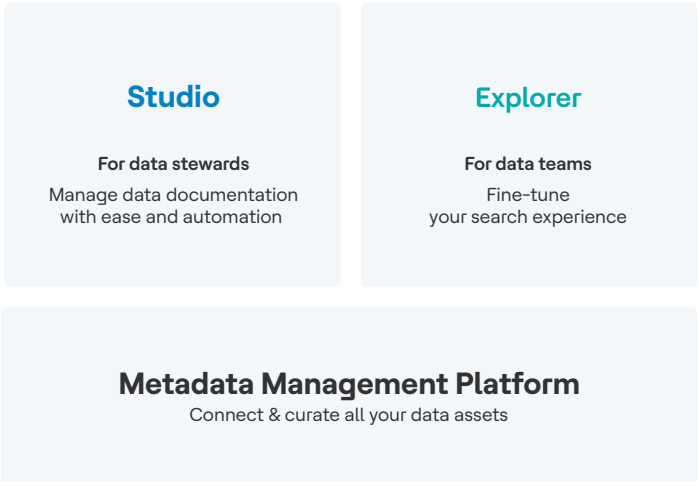


Figure 4 : Interfaces dédiées aux data stewards et aux utilisateurs de données

Cette approche s'inscrit dans la logique user-friendly des plateformes de type marketplace. Ces solutions reposent généralement sur deux applications distinctes. La première, une solution de « back office », permet au équipes ou partenaires de la marketplace d'alimenter le catalogue de façon aussi automatisée que possible, tout en contrôlant la qualité des contenus.

La seconde s'adresse aux consommateurs de données et prend souvent la forme d'un site e-commerce permettant aux utilisateurs finaux de rechercher des articles de données ou d'explorer le catalogue. Explorer et Studio reflètent clairement ces deux usages.

Une courbe d'apprentissage raccourcie

Avec le catalogue de données Actian Data Intelligence Platform, la courbe d'apprentissage est considérablement réduite, grâce à une interface familière pour les utilisateurs. Une fonctionnalité de guide interactif pour les nouveaux arrivants accélère encore la prise en main : aucune formation n'est nécessaire.

Explorer est la clé de l'adoption à grande échelle du catalogue de données, et s'inspire directement des sites e-commerce dans sa conception :

- Son design intuitif reprend les codes des grandes marketplaces, en intégrant un moteur de recherche efficace, des fonctionnalités d'exploration avancées, ainsi qu'un système de recommandation capable de proposer une sélection d'objets adaptée au profil de chaque utilisateur.
- Comme dans les marketplaces, la recherche est le principal moyen d'accéder à l'information. L'interface s'inspire fortement des moteurs de recherche classiques et des sites e-commerce grand public.

Le catalogue de données Actian Data Intelligence Platform classe l'information en fonction du rôle de l'utilisateur dans l'organisation. Il adapte dynamiquement la hiérarchie des informations selon le profil utilisateur.

C'est cette hiérarchie d'information qui différencie un catalogue de données d'un catalogue de type marketplace. Dans une marketplace, la hiérarchie est identique pour tous : photos, descriptions, prix et conditions de livraison passent en priorité, suivis par les détails techniques, les avis et les délais.

Dans un catalogue de données, c'est différent : la hiérarchie dépend du rôle opérationnel de chaque utilisateur. Pour certains, les données techniques comme l'emplacement, la sécurité, les formats ou les types seront les plus importantes. D'autres s'attacheront à la sémantique des données et à leur lignée métier. D'autres encore chercheront à comprendre les processus et contrôles qui encadrent la production de la donnée, notamment pour des raisons réglementaires ou opérationnelles. Un catalogue de données intelligent doit être capable d'ajuster dynamiquement la structure de l'information pour s'adapter à chaque type d'utilisateur.

Le dernier défi concerne la manière dont l'information est organisée dans le catalogue de données. Elle l'est sous forme de parcours d'exploration par thématique, un peu comme des rayons dans une marketplace. Mais il est difficile de trouver une structure qui convienne à tout le monde.

Certains vont explorer le catalogue selon une logique technique : systèmes, applications, technologies. D'autres préféreront une approche fonctionnelle, par domaines métiers. D'autres encore adopteront un point de vue sémantique, via des glossaires métier, par exemple.

Tenter de faire adopter une classification unique et universelle est souvent une mission impossible pour les entreprises. C'est pourquoi un catalogue de données intelligent doit rester adaptable, et ne pas imposer aux utilisateurs une structure qui n'a pas de sens pour eux.

En fin de compte, l'expérience utilisateur reste l'un des facteurs clés de succès d'un catalogue de données. Cette expérience dépend à la fois du design des applications proposées par le catalogue, mais aussi de l'efficacité et de la simplicité du métamodèle.

Points clés pour optimiser un catalogue de données intelligent

La gestion efficace des métadonnées dans l'ensemble de l'organisation est la pierre angulaire d'une stratégie data réussie. Et le catalogue de données intelligent est l'outil le plus efficace pour y parvenir.

Mais l'aspect *intelligent* ne se limite pas à l'intégration d'algorithmes. Il doit se refléter dans tous les aspects du catalogue, notamment :

- La manière dont le métamodèle peut être conçu, enrichi, modifié ou amélioré à mesure que le catalogue gagne en adoption.
- L'automatisation avancée de l'inventaire des actifs de données, ainsi que la collecte des métadonnées dans les systèmes qui les hébergent.
- La capacité à aider les data stewards à alimenter et contrôler le contenu du catalogue.
- Le moteur de recherche, qui constitue le point d'accès le plus simple et le plus direct pour les consommateurs de données.
- Et enfin, l'expérience utilisateur, qui doit tenir compte de la diversité des profils qui utiliseront le catalogue.

Un véritable catalogue de données intelligent ne se limite pas à un simple référentiel de métadonnées : c'est un outil dynamique et évolutif, capable de s'adapter aux besoins d'utilisateurs très divers.

En plaçant l'expérience utilisateur au cœur du dispositif, en s'appuyant sur des hiérarchies d'information flexibles et une automatisation intelligente, le catalogue de données permet à chaque utilisateur (des équipes techniques aux dirigeants), d'accéder aux informations qui leur sont utiles, de manière claire et pertinente.

À l'heure où la donnée est un actif stratégique, les organisations qui adoptent un catalogue de données intelligent, centré sur les utilisateurs, seront les mieux armées pour gagner en efficacité, renforcer la collaboration et stimuler l'innovation.

À propos d'Actian

Actian permet aux entreprises de gérer et de maîtriser leurs données en toute confiance et à grande échelle. Les organisations font confiance aux solutions de gestion et d'intelligence des données d'Actian pour rationaliser les environnements de données complexes et accélérer la fourniture de données prêtes pour l'IA. Conçues pour être flexibles, les solutions Actian s'intègrent de manière transparente et fonctionnent de manière fiable dans les environnements sur site, cloud et hybrides.

Pour en savoir plus sur Actian, la division Données de **HCLSoftware**, rendez-vous sur le site **actian.com**.

Actian France 27 Avenue de l'Opéra, 75001 Paris, France Tel: +33 1 73 19 12 87 **Actian.com**

ACTIAN™
a division of HCLSoftware