DAWiZZ
All about your data

# Pour maîtriser et valoriser son patrimoine data, encore faut-il le connaître !

**Stéphane LE LIONNAIS**
Cofondateur de **DAWIZZ**
stephane.lelionnais@dawizz.fr
06 71 17 73 40

DAWiZZ
All about your data

# Le Contexte

IDC ANALYZE THE FUTURE

•Selon des **études IDC**, les volumes de données devraient atteindre **175 zettaoctets** (un zettaoctet = 1 milliard de téraoctets !) à l'échelle mondiale d'ici 2025 … et en parallèle, **moins de 0,5 % de ces données sont actuellement analysées**

•Les données non structurées représentent 80% de toutes les données des organisations

•60% du temps passé par les data scientists est consacré à chercher, nettoyer et organiser les données

•Coût moyen d'une attaque en France: 773 000 €

•Coût d'une amende CNIL jusqu'à 4% CA.

| Données structurées | Données non structurées |
|---|---|
| • Généralement créées et stockées dans des bases de données ou fichiers structurés (csv, json, …) | • Généralement créées et stockées dans des Fichiers textes, email, images, … |
| • Représentent 20% de toutes les données de l'organisation | • Représentent 80% de toutes les données de l'organisation |
| • Généralement associées à des métadonnées techniques difficilement interprétables (noms de tables, noms de champs, …) | • Généralement impossibles à qualifier car très peu de métadonnées disponibles (taille, dates, acl, …) |
| • Stockage centralisé | • Stockage disparate / éclaté |

DAWiZZ
All about your data

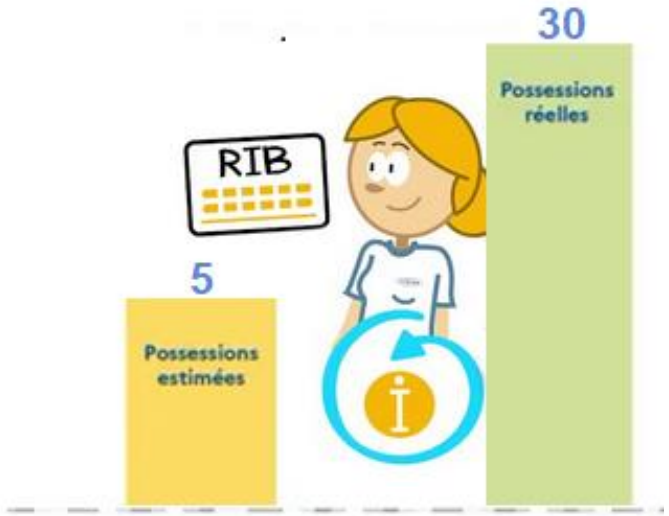
# Une comparaison sur un sujet d'actualité !!

## Le delta entre l'estimation et la réalité des possessions

De mars à octobre 2021, Avec l'aide de l'Ademe, six « home-organisers » – des professionnels de l'organisation et du rangement, métier tout nouveau- sont entrés dans 21 foyers volontaires pour travailler, avec eux, au désencombrement de leurs logements.

x 2

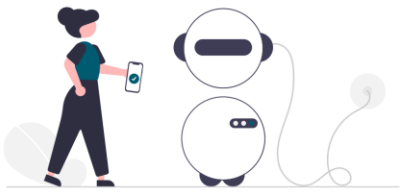
Quid en ce qui concernent les **données personnelles** ?




Moyenne calculée à partir de 20 audits (data-discovery automatisé) réalisés dans des organisations en 2021

x 6

DAWiZZ
All about your data



# Le rôle de l'intelligence artificielle

## L'enjeu de la data driven strategy

Pour guider les décisions à partir des données, l'enjeu, pour les organisations, réside actuellement dans l'exploitation et la valorisation des données non exploitées, avec pour objectifs de les rendre utiles et utilisables.

Mais le volume des données augmente plus vite que la capacité des organisations à les traiter. Le travail pour connaitre son patrimoine de données est effectivement d'une ampleur parfois abyssale car souvent manuel.

L'usage de l'intelligence artificielle, permet :

- non seulement un gain de productivité (une économie du temps de traitement par les collaborateurs) en particulier dans le domaine de l'analyse des données qu'elles soient structurées ou non et quel que soit le support / l'application utilisée ;
- mais aussi et surtout une connaissance exhaustive de ses données, ce qu'aucune solution « manuelle » ne permet d'atteindre.

## L'enjeu de la conformité et de la cybersécurité

L'analyse et la surveillance des données est indispensable pour s'assurer de leur bonne utilisation et veiller ainsi au respect des politiques de conformité et des politiques sécuritaires

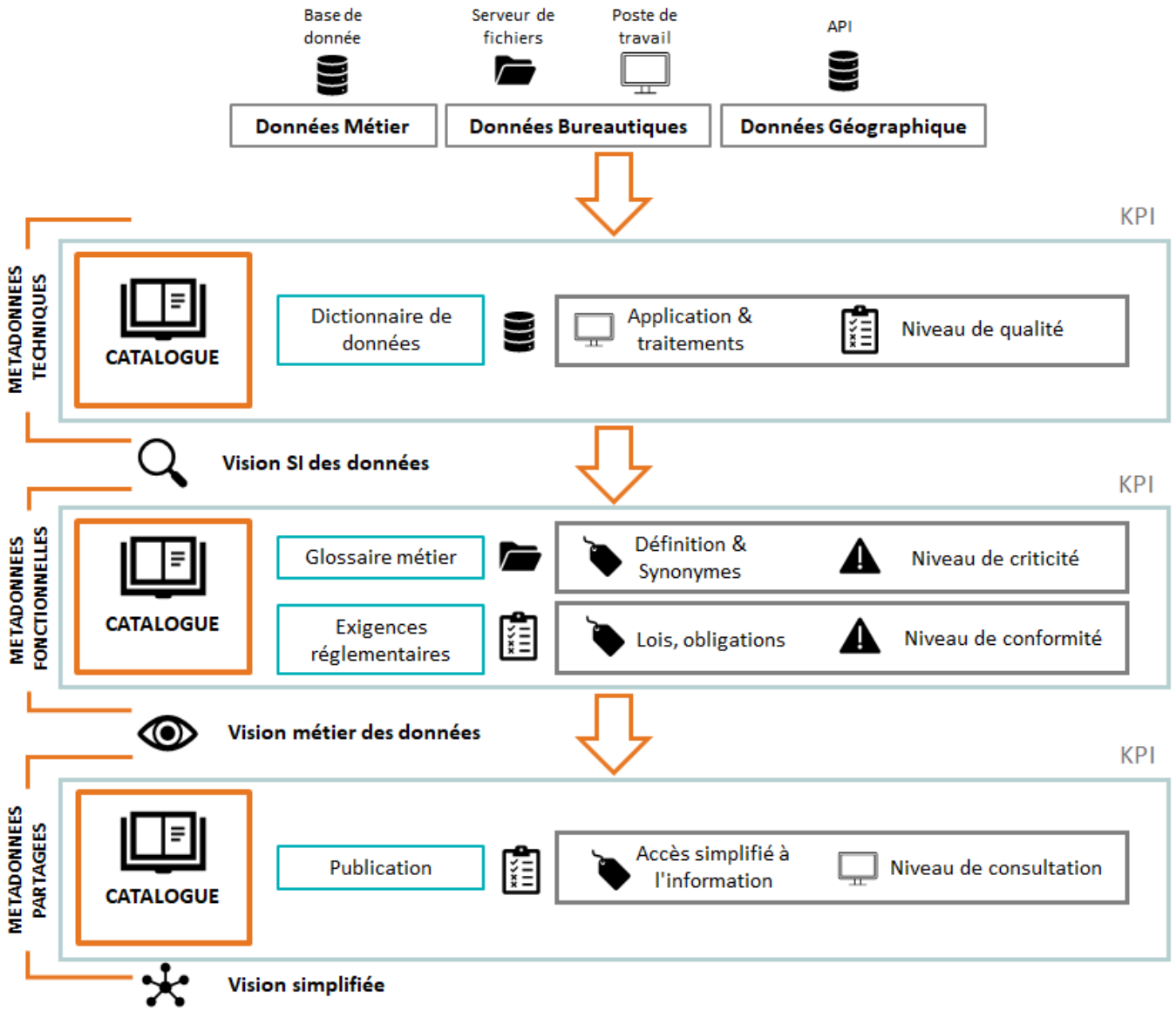
L'IA permet d'accélérer la découverte des données (avec leur niveau de sensibilité / confidentialité) ce qui permet ensuite l'organisation de se mettre en conformité

## L'enjeu de la conformité

Une surveillance régulière des données par une IA peut permettre la découverte d'anomalies et mettre en exergue des défauts de qualité (erreurs de saisie, …)

# Exemple d'IA intégrée dans un Data-Catalog

DAWiZZ
All about your data




## CHOIX DU MODE D'ANALYSE

## ALGORITHME D'ANALYSE

Des **algorithmes** peuvent permettre de **normaliser** et rendre compréhensible les attributs audités. Par exemple, l'attribut « AdrCliFact » sera automatiquement normaliser en « Adresse Client facturation ». Autre exemple, CA_HT sera normé en « Chiffre d'Affaires Hors Taxes » et non en « Comice Agricole de Haucourt ».

## ALGORITHME DE CLASSIFICATION

Les attributs normalisés peuvent être, ensuite, **classés automatiquement** en fonction de glossaire métier défini sous forme de thésaurus publics (RGPD, cybersécurité) ou de thésaurus privés (urbanisation, Analyse des données sensibles, Référentiels de données, ..).
Exemple concernant l'analyse des données non structurés, « M. Andouille » sera prioritairement taggué comme un nom de famille et non une injure.

DAWiZZ
All about your data

Thésaurus | Attributs 18 | Contacts 1 | Liens

Analyse des identifiants 1 17

| Index | Identifiant | Enregistrements % | Analyse des titres et concepts (Classification pour thesaurus RGPD ) | Analyse des données % (Exécution des matchers de la sonde) |
|---|---|---|---|---|
| 0 | **sootherp.annu_client** | 0 | client<br>Identification personnelle | |
| 1 | ref_contact | 0 | référence contact<br>Identification personnelle | |
| 2 | id_client_categ | 0 | identifiant client catégorie<br>Identification personnelle | |
| 3 | type_client | 0 | type client | |
| 4 | id_tarif | | | |
| 5 | ref_commercial | | | |
| 6 | ref_adr_livraison | 0 | référence adresse livraison | |

Fourni par le processus de normalisation.
Découpage selon séparateurs : ref, adr, livraison.
Le mot suivant est reconnu comme acronyme et remplacé : 'ref' ⇒ 'référence' (fr).
Le mot suivant est reconnu comme acronyme et remplacé : 'adr' ⇒ 'adresse' (fr).
Le mot suivant est reconnu dans le dictionnaire : 'livraison' (fr).

*Pourquoi le choix d'une* ***analyse par une Approche algorithmique ...***
*... et non par une* ***Approche de centralisation****/d'indexation* (Analyse type Big Data) ?

=> Pour des raisons de sécurité, de coût, l'ensemble des données du SI ne se retrouvent pas dans cette centralisation
=> Pour participer à une démarche RSE en évitant la recopie de données

Thésaurus | Attributs 14 | Contacts 1 | Liens

Analyse des identifiants 1 12 1

| Index | Identifiant | Enregistrements % | Analyse des titres et concepts (Classification pour thesaurus RGPD ) |
|---|---|---|---|
| 0 | **sootherp.comptes_bancaires** | 0 | comptes bancaires |
| 1 | id_compte_bancaire | 0 | identifiant compte bancaire |
| 2 | ref_contact | 0 | référence contact<br>Identification personnelle |
| 3 | lib_compte | 0 | libellé compte |
| 4 | ref_banque | 0 | référence banque |
| 5 | code_banque | 0 | code banque<br>Information bancaire |
| 6 | code_guichet | 0 | code guichet<br>Information bancaire |
| 7 | numero_compte | 0 | numero compte<br>Information bancaire |
| 8 | cle_rib | 0 | cle relevé identité bancaire<br>Information bancaire |

DAWiZZ
All about your data

# Merci de votre attention

**Stéphane LE LIONNAIS**
Cofondateur de **DAWIZZ**
stephane.lelionnais@dawizz.fr
06 71 17 73 40