

Chapitre 3

Les données juridiques

1. Introduction

Dans notre monde qui aujourd'hui se partage entre physique et virtuel, nous assistons à la **génération d'un volume de données croissant**. Ces données, issues de sources multiples (réseaux sociaux, objets connectés, applications, photos, audio, vidéo, etc.) et aux formats variés, sont devenues essentielles au fonctionnement de notre société, voire vitales.

Il est à noter que cette masse de données n'est aujourd'hui ni complètement exploité, ni complètement exploitable ; en fait, seule une infime partie est utilisée. D'après **Earthweb**, en 2022, **3 % du volume total des données produites serait à ce jour exploité**.

Les applications exploitant les données sont de plus en plus nombreuses, de plus en plus efficaces et de plus en plus innovantes. Elles sont capables aujourd'hui d'utiliser et de croiser de **nombreuses sources de données**, de telle sorte que nos vies deviennent de plus en plus tributaires de ces traitements et, surtout, **de moins en moins privées**.

Des données qui, à leur origine, semblaient anodines, voire anonymes, peuvent devenir, après leur croisement avec d'autres sources, particulièrement sensibles. C'est notamment le cas pour les données du droit.

Les données nous permettent de prendre des décisions, elles nous facilitent la vie mais pas uniquement, elles nous font aussi vivre. Que deviendrions-nous sans elles ; Imaginez en vrac : les comptes bancaires, la fraude, la sécurisation de sites sensibles et des prisons, les réseaux sociaux, les documents dématérialisés, la gestion de l'électricité, les moyens de transport, les moyens de navigation (dont le GPS), les voyages, les réservations de toutes sortes, la gestion des hôpitaux et des dossiers patients, la prise de rendez-vous, les prothèses intelligentes, etc.

Dans ce chapitre, je vais détailler **les concepts et problématiques généraux liés aux données**, en mettant **l'accent sur la gestion, la conservation, l'utilisation et la sécurisation des données** qui sont issues et utilisées par le domaine du droit.

Ce chapitre concerne avant tout à la génération et à l'utilisation des données juridiques **en France** ; je ferai néanmoins référence à d'autres pays lorsque cela sera nécessaire.

2. Le type de données

Les données peuvent être divisées en **trois types** qui seront décrits par la suite : **les données structurées** qui s'organisent dans des formats et selon des modèles prédéfinis, **les données non structurées** qui adoptent des formes différentes qui ne suivent pas nécessairement un format particulier, ce qui rend leur exploitation un peu plus complexe. Elles constituent la majorité des données disponibles aujourd'hui. Et pour finir, **les données semi-structurées**, qui sont essentiellement un mélange de données structurées et non structurées.

Un quatrième type de données est aujourd'hui à considérer : **les données synthétiques**. Néanmoins, nous allons voir que les données synthétiques peuvent être vues comme un sous-ensemble des types précédents.

2.1 Les données structurées

Pour la plus grande part, **les données structurées sont des informations qui sont organisées dans des fichiers texte, des feuilles de calcul et des bases de données relationnelles**. On y trouve : les numéros de téléphone, de sécurité sociale, les divers codes, les champs textuels fixes ou variables dans les enregistrements des bases de données, des ensembles de nombres, etc.

Ces données peuvent être **générées par un humain ou par une machine**, mais la condition première est qu'elles puissent être facilement **classées, retrouvées et extraites**.

Les entreprises et organismes sont familiarisés avec ce type de données, que l'on considère comme des sources de données « **traditionnelles** ».

Ces données sont généralement stockées dans des systèmes de bases de données, le plus souvent relationnelles, et associées au système *back-end* de l'entreprise qui comprend les solutions de :

- **CRM (*Customer Relationship Management*)** : plateforme logicielle conçue pour aider les entreprises à gérer et organiser leurs interactions avec leurs clients et prospects. Elle centralise les données clients, permettant de suivre l'historique des communications, les préférences, les achats et les besoins spécifiques.
- **ERP (*Enterprise Resource Planning*)** : solution informatique intégrée qui permet à une entreprise de gérer l'ensemble de ses ressources, processus et activités en un seul système centralisé. Elle couvre des fonctions clés telles que la finance, la gestion des ressources humaines, la production, la logistique, la gestion des stocks et la comptabilité.
- **SCM (*Supply Chain Management*)** : solution logicielle qui optimise la gestion de la chaîne d'approvisionnement. Le système permet de coordonner et de suivre l'ensemble des activités liées à la planification, à l'approvisionnement, à la production, au stockage, à la distribution et à la livraison des produits.

66 ——— Quand l'IA révolutionne le droit

Opportunités et défis

- **BI (*Business Intelligence and Data Mining*)** : plateforme qui permet d'analyser et de visualiser les données d'entreprise pour faciliter la prise de décision. Le système collecte, compile et interprète des informations provenant de différentes sources afin de produire des rapports, des tableaux de bord et des indicateurs clés de performance (KPI).
- et des **applications spécifiques** au secteur ou à l'entreprise.

La particularité des données structurées est que leur volume, durée de vie et croissance sont connus et peuvent être anticipés.

2.2 Les données non structurées

Contrairement à ce que leur nom pourrait laisser croire, les données non structurées contiennent en principe une structure, mais celle-ci ne suit pas les modèles de données ou schémas prédéfinis. Ces données sont très variées, elles peuvent être textuelles ou non textuelles, générées par un humain, une machine ou un objet et peuvent être stockées dans des bases de données spécifiques aux données non structurées.

Exemples de **données non structurées générées par des humains** :

- **Les applications métiers** : documents, notes, etc.
- **Les communications instantanées ou pas** : chats, IM, enregistrements téléphoniques, outils de collaboration.
- **Les courriers électroniques** : bien qu'ils soient parfois considérés comme semi-structurés, puisqu'on peut leur associer des métadonnées, ces contenus sont non structurés, car ils ne peuvent pas être traités par des outils d'analyse classiques.
- **Les divers médias** : MP3, photos digitales, fichiers audios et vidéos.
- **Les données en provenance des téléphones portables et tablettes** : messages, positionnements, comportements, habitudes d'achat, sites visités, contacts, etc.
- **Les fichiers texte** : documents créés à l'aide d'un traitement de texte, feuilles de calcul, présentations, logs (fichiers de trace).

- **Les données fournies aux réseaux sociaux** : ce sont les données en provenance des réseaux privés ou des réseaux professionnels, tels que Facebook, X, LinkedIn, Tik-Tok, etc.
- **Les données fournies et stockées sur les sites web spécialisés** : YouTube, Instagram, Flickr et les sites de partage de contenus (photos, vidéos, jeux, sons, musiques, etc.).

Exemples de **données non structurées générées par des machines ou des objets** :

- **Les données en provenance des capteurs et des sondes** : trafic, climat, géographie, centre de données, véhicules, etc.
- **Les données scientifiques** : exploration pétrolière, exploration spatiale, imagerie sismique, données atmosphériques.
- **Les images satellites** : données météo, paysages, mouvements militaires ou autres.
- **La surveillance digitale** : photos ou vidéos de surveillance.

Un nombre important de données non structurées est produit par des machines ou des objets. L'humain n'ayant pas obligatoirement connaissance de celles-ci, on parlera dans ce cas de données cachées (*Hidden Data*). Les données cachées représentent **60 % des données non structurées**.

2.3 Les données semi-structurées

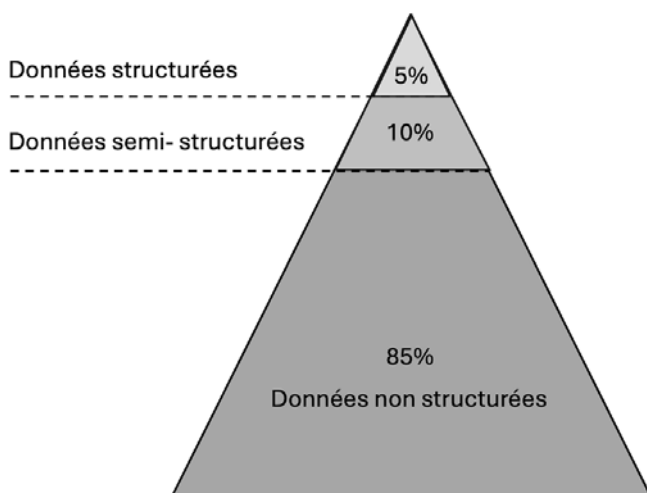
Les données semi-structurées sont une forme de données structurées qui n'obéissent pas à la structure des modèles de données associés aux bases de données relationnelles ou à d'autres formes de tables de données. Néanmoins, elles contiennent une certaine structure, notamment grâce aux **balises** ou autres **marqueurs** qui permettent de séparer les différents éléments sémantiques. On parle aussi de **structure auto-descriptive**.

Même si les données ne sont pas totalement structurées, **la structure auto-descriptive permet d'analyser plus facilement les données**. Elle peut aussi, au travers de processus et développements assez simples, permettre le stockage de ces données dans des bases de données relationnelles.

68 ——— Quand l'IA révolutionne le droit

Opportunités et défis

Le stockage et l'utilisation des données semi-structurées sont moins faciles qu'avec des données structurées, mais plus simples qu'avec des données non structurées. Il faut noter que l'interprétation de relations entre les données semi-structurées est rendue difficile par leur structure irrégulière et partielle et par le fait que certaines sources ont une structure de données implicite.



Données structurées, semi-structurées et non structurées (Source IBM, 2025)

D'après une enquête réalisée en 2023 par Statistica (www.statista.com), **l'accroissement exponentiel du volume de données produit** permet d'avancer le nombre de **2142 Zo (2142, suivi de 21 zéros) comme étant le volume global** des données qui sera stocké un peu partout dans le monde et sur tout type d'objet en 2035.

Il sera impossible, dès lors, de faire transiter ces données sur les réseaux existants, quelles que soient les technologies employées. Cela signifie que ces données devront être stockées différemment et que pour la majorité d'entre elles, elles ne pourront pas bouger. Elles devront rester et être traitées là où elles sont produites, aussi la sécurité de ces données sera un enjeu essentiel.

2.4 Les données synthétiques

Les données synthétiques (*synthetic data*) sont des données générées artificiellement, via des algorithmes informatiques.

Les jeux de données synthétiques peuvent être de **deux formes** : soit **totale-ment synthétiques**, soit **partiellement synthétiques** et ainsi créées en complément d'un jeu de données réelles (*Data Augmentation*) ou en remplacement de données réelles sensibles, par exemple l'**anonymisation** (*Data Alteration*).

Les données synthétiques offrent plusieurs avantages, notamment la préservation de la vie privée, en générant des données qui imitent les informations réelles des individus mais ne font référence à aucun humain.

3. Les données juridiques

3.1 Définition

Les données juridiques se réfèrent à **l'ensemble des informations et documents relatifs aux lois, règlements, décisions judiciaires, doctrines et autres matériaux pertinents dans le domaine du droit**. Elles comprennent les textes législatifs (comme les lois et règlements), la jurisprudence (comme les décisions de justice), les contrats, les avis juridiques, les analyses doctrinales, et d'autres sources qui aident à interpréter et appliquer le droit.

Ces données servent de base pour la recherche juridique, la pratique professionnelle et l'éducation dans le domaine légal. Elles sont essentielles pour les divers acteurs du domaine du droit. **Voici pour ces données, quelques aspects à considérer :**

- **L'accessibilité et la numérisation** : de plus en plus de données juridiques sont disponibles sous forme numérique, rendant l'accès et la recherche plus faciles et plus rapides, notamment à travers des bases de données en ligne (voir sous-section La variété des bases de données juridiques de ce chapitre).

70 ——— Quand l'IA révolutionne le droit

Opportunités et défis

- **L'analyse prédictive** : les technologies d'intelligence artificielle et d'analyse de données sont utilisées pour tirer des tendances et des prédictions à partir de vastes ensembles de données juridiques, ce qui peut aider à anticiper des résultats potentiels dans des litiges.
- **La conformité et le conseil** : les entreprises et les individus utilisent les données juridiques pour assurer leur conformité aux lois en vigueur et pour obtenir des conseils légaux avisés.
- **L'interprétation** : les données juridiques nécessitent souvent une interprétation experte en raison des complexités du langage juridique et des particularités des systèmes juridiques locaux, nationaux et internationaux.
- **La mise à jour constante** : le droit est un domaine en constante évolution, avec des lois modifiées, abrogées ou introduites régulièrement. Les données juridiques doivent donc être mises à jour fréquemment pour rester pertinentes.

L'utilisation efficace des données juridiques peut **améliorer la prise de décision**, la **stratégie juridique**, et aider à **naviguer dans le système judiciaire** de manière plus efficiente.

Il faut être conscient que le croisement des données juridiques avec d'autres données peut permettre d'accéder à des données plus personnelles et ainsi générer des données juridiques dites sensibles.

Les données sensibles sont des données plus personnelles que les données personnelles classiques (on pourrait parler de données très personnelles). Pour ces raisons, ces données sont soumises à des règles juridiques particulières.

3.2 Catégories des données juridiques

Les données juridiques **peuvent se décliner en plusieurs catégories**, chacune ayant ses propres caractéristiques et utilisations. Ces diverses catégories, sources et types de données juridiques permettent d'obtenir une vue d'ensemble de l'application et de l'évolution des lois. Elles sont essentielles non seulement pour la pratique du droit, mais aussi pour la recherche, l'éducation et la sensibilisation du public sur les questions juridiques. Je mentionne ci-dessous **les principales catégories et des exemples** des données qu'elles contiennent :

3.2.1 Sources écrites

- **Les contrats et les conventions** : documents qui établissent des accords juridiques entre parties. Ils peuvent inclure des contrats de travail, des baux, des accords commerciaux, etc.
- **Les documents de procédure** : ils incluent des formats et formulaires juridiques tels que des requêtes, des plaintes, des jugements, et d'autres documents utilisés dans les procédures judiciaires.
- **La doctrine** : elle comprend les travaux des juristes, chercheurs et universitaires qui analysent, commentent et critiquent le droit et son application. Cela peut inclure des ouvrages, articles de revues et thèses.
- **La jurisprudence** : elle regroupe les décisions rendues par les tribunaux. Cela inclut les arrêts des cours supérieures et les décisions judiciaires qui interprètent et appliquent la loi à des cas spécifiques. Dans le cadre de la jurisprudence, l'application des lois tient compte des cas spécifiques (jurisprudence) qui ont pu avoir lieu. Dans les cas plus classiques (les plus nombreux), la justice applique la loi telle que définie.
- **Les textes législatifs (lois, décrets, règlements)** : les constitutions, les lois, les décrets et les règlements émis par les autorités gouvernementales. Ces documents définissent les normes juridiques et sont soumis au respect des traités internationaux et à la Constitution.