Extraction d’information :

Table des matières

[Sources : 1](#_Toc63160455)

[Principes : 1](#_Toc63160456)

# Sources :

[1] « Que signifie Données semi-structurées? - Definition IT de Whatis.fr », *Whatis.com/fr*. https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees (consulté le févr. 02, 2021).

[2] « Information extraction », *Wikipedia*. déc. 05, 2020, Consulté le: févr. 02, 2021. [En ligne]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Information\_extraction&oldid=992409176.

# 

# Principes :

## Définition :

L’extraction d’informations désigne la recherche, dans un texte ou un ensemble de textes (corpus) automatisées de données informatives en liens avec un sujet donné et plus ou moins précis.

Ces outils (d’extraction d’information) donnent la possibilité d’extraire des informations à partir de documents textuels, de bases de données, de sites web… Ces sources d’informations peuvent être non structurées[[1]](#footnote-1), semi structurées[[2]](#footnote-2) ou structurées[[3]](#footnote-3).

Pour donner un exemple, un document de traitement de texte écrit avec Word est de manière générale considéré comme un ensemble de données non structurées : pragmatiquement, il s’agit d’une suite de caractères plus ou moins formatée et plus ou moins longue. Ces données ne sont pas à proprement parler « structurées ». On peut néanmoins leur associer une certaine somme d’informations qu’on appelle « métadonnées [[4]](#footnote-4)».

Dans les faits, les délimitations de ces trois catégories ( structurées, semi structurées ou non structurées) ne sont pas clairement définies. On parlera, pour les considérer dans leur ensemble, de « *continuum des données* »[1].

Les techniques de l’extraction d’information sont principalement utilisées dans le cadre du Traitement Automatique du Langage Naturel, dont leur but est d’extraire du texte structuré à partir d’un texte qui ne l’est pas.

## Historique :

L’extraction d’information est une discipline qui a vu le jour dans les années 1970 avec l’apparition du Traitement Automatique du Langage Naturel. L’un des premiers systèmes commercialisés était « JASPER »[2], développé dans le but de fournir des informations financières en temps réel aux trader financiers.

Quelques-unes des problématiques actuelles de ce domaine dont l’annotation automatique ou la reconnaissance et l’extraction de contenu depuis des fichiers d’images ou de vidéos.

Le très grand nombre de règles linguistiques issues de la complexité du langage est la raison principale pour laquelle le développement de systèmes d’extraction d’information de haute qualité se révèle être un véritable enjeu, notamment pour les systèmes d’Intelligence Artificielle.

1. Désignation générique qui vient décrire toute donnée qui n’appartient pas à une structure particulière. Les données textuelles sont majoritairement générées par les documents de traitement de texte (.doc, .rtf …), les logiciels de collaboration ou de messagerie instantanée (Teams, Azure DevOps … ) tandis que les données non textuelles correspondent plutôt à des supports tels que les images .jpeg, les fichiers audios .mp3 ou encore les fichiers vidéo. Il faut retenir que le format ne permet pas d’y accéder ni de les traiter « facilement ».

   Les données totalement « non structurées » sont en réalité assez rares, puisque la présence d’une structure peut être parfois observable à une autre échelle.

   En l’absence d’une gestion ou d’une administration définie, le volume généré annuel de données dites « non-structurées » peut se révéler être un véritable coût en termes de stockage.

   A cela s’ajoute que la présence de données non administrées dans des quantités plus ou moins importantes peut parallèlement poser des problèmes de responsabilité (exemple : cas d’une action en justice durant laquelle des informations ne pourraient pas être localisées)

   Extraire des informations de données non structurées se révèle souvent être une tâche difficile. Cela implique que les documents soient numérisés, afin de permettre à un système numérique d’en extraire des concepts en fonction de contextes spécifiques (recherche sémantique).

   <https://www.lemagit.fr/definition/Donnees-non-structurees> [↑](#footnote-ref-1)
2. Désignation générique qui désigne des ensembles de données qui ne sont pas organisés en référentiel spécialisé (exemple : en dans une base de données), mais qui comportent cependant des informations associées (exemple : des métadonnées) qui les différencient des données « brutes » et les rendent plus facile à traiter. L’organisation n’est pas « complexe », ce qui rend possible une analyse « poussée ».

   <https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees> [↑](#footnote-ref-2)
3. Ce terme désigne le contraire des données non structurées. Elles ont été formatées, et leurs éléments ont été réorganisés en fonction d’une structure leur permettant d’être traitées, organisées et manipulées selon diverses approches, avec diverses technologies.

   <https://whatis.techtarget.com/fr/definition/Donnees-semi-structurees> [↑](#footnote-ref-3)
4. Ce terme signifie « donnée de/à propos de données ». Il désigne une donnée servant à définir ou décrire une autre donnée.

   <https://fr.wikipedia.org/wiki/M%C3%A9tadonn%C3%A9e> [↑](#footnote-ref-4)