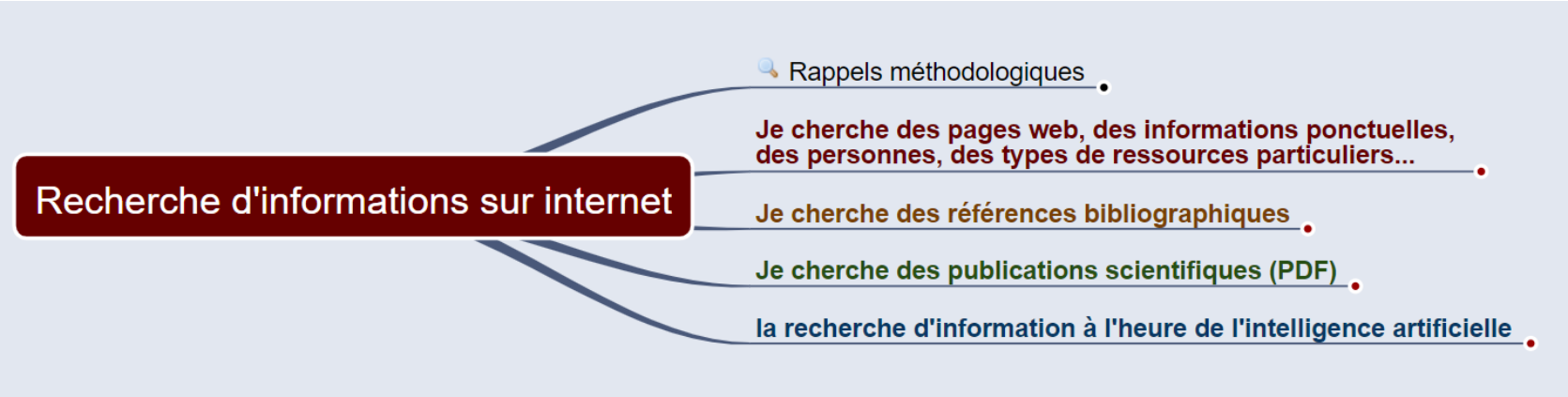
Recherche d'informations sur Internet (perfectionnement)

méthodologie et outils disponibles

TENDANCES 2023



support https://framindmap.org/c/maps/446941/public

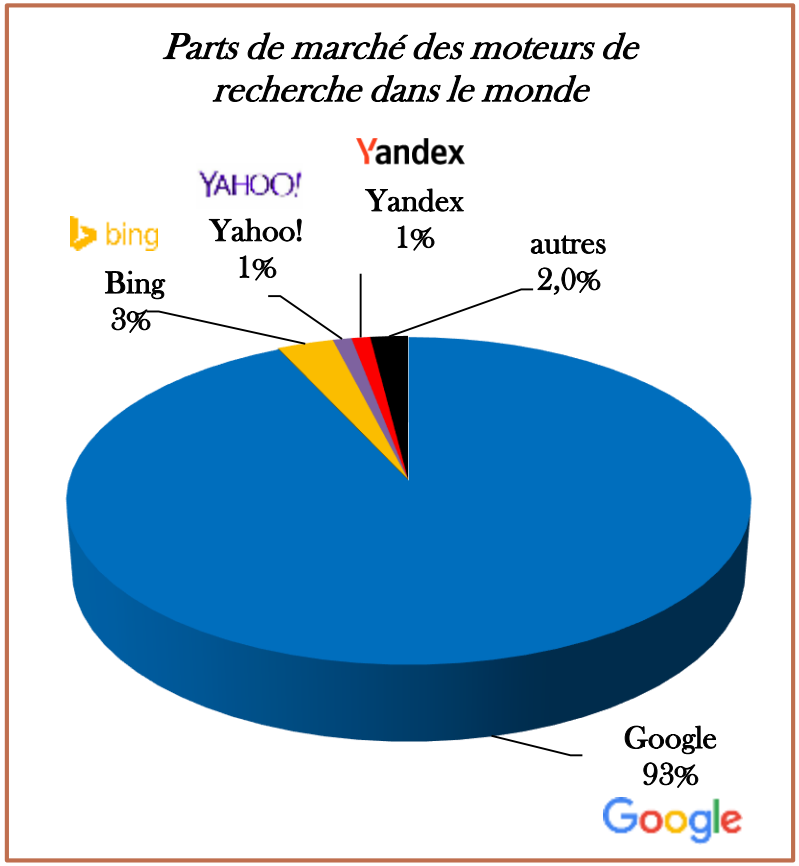
support général

https://urfist.chartes.psl.eu/ressources/recherche-d-informations-sur-internet-perfectionnement

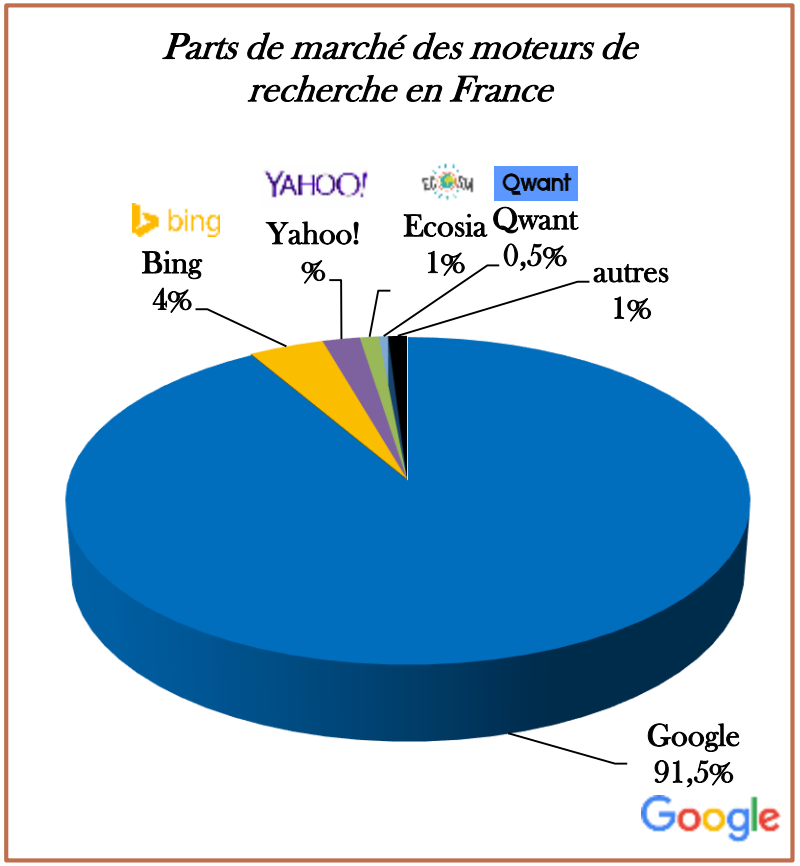
(v. 2021)

Moteurs de recherche

Monde



France



d'après StatCounter, 06/2023,
desktop et mobile

Moteurs de recherche

- les moteurs de recherche en 2023

des alternatives aux *Big tech* ?

- Google (et Bing) dominent toujours le marché
- quelques alternatives à tester/utiliser (Brave, DuckDuckGo, Qwant), avec cependant des limites (taille de l'index, couverture des sources francophones, fonctionnalités de recherche inégales...)

modèle économique et données personnelles ?

- des modèles encore souvent dominées par le recours à la publicité et/ou la récupération de données personnelles
- développement d'offres ou de solutions payantes qui doivent néanmoins trouver leur place dans un paysage avant tout gratuit

des fonctionnalités plus poussées

- développement de nouvelles fonctionnalités (ex. recherche sémantique cf. *Knowledge graph* de Google, recherche inversée d'images)
- Google reste cependant dominant dans la panoplie de fonctionnalités avancées (tris, filtres, équations) possibles

des moteurs de recherche aux assistants personnels

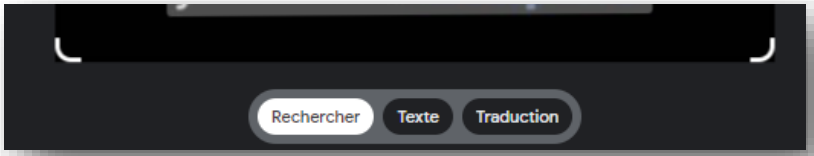
- de moins en moins des moteurs de recherche (de liens) et de plus en plus des moteurs de réponse
- développement d'autres formes d'interactions (ex. : conversation)
- prise en compte de fonctionnalités / contenus issus de l'intelligence artificielle [IA]

Moteurs de recherche

- 2023 : nouveautés Google

  mise en valeur et développement de la recherche inversée d'images

intégration directement
dans le moteur classique

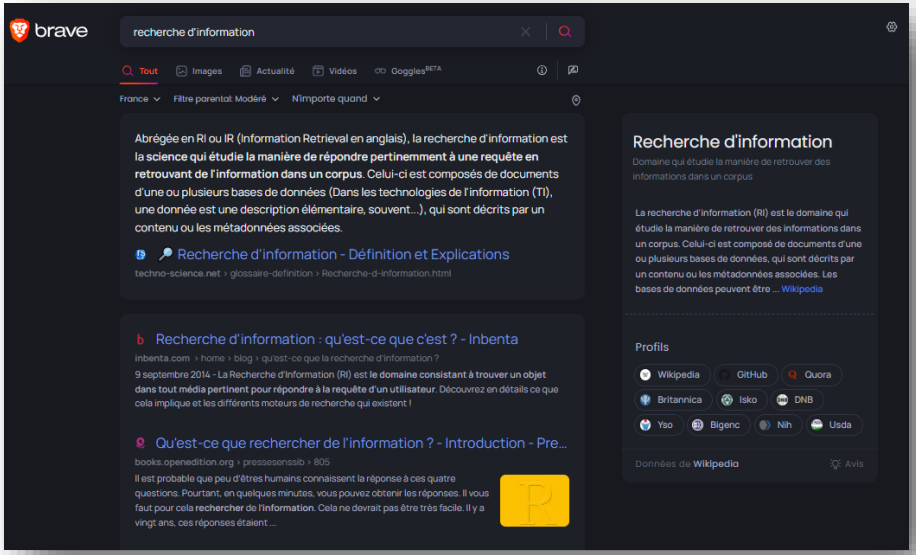


fonctionnalités avancées :
- sélection d'une portion d'image
- sélection de texte sur l'image
- traduction du texte
- remplacement du texte traduit dans l'image d'origine

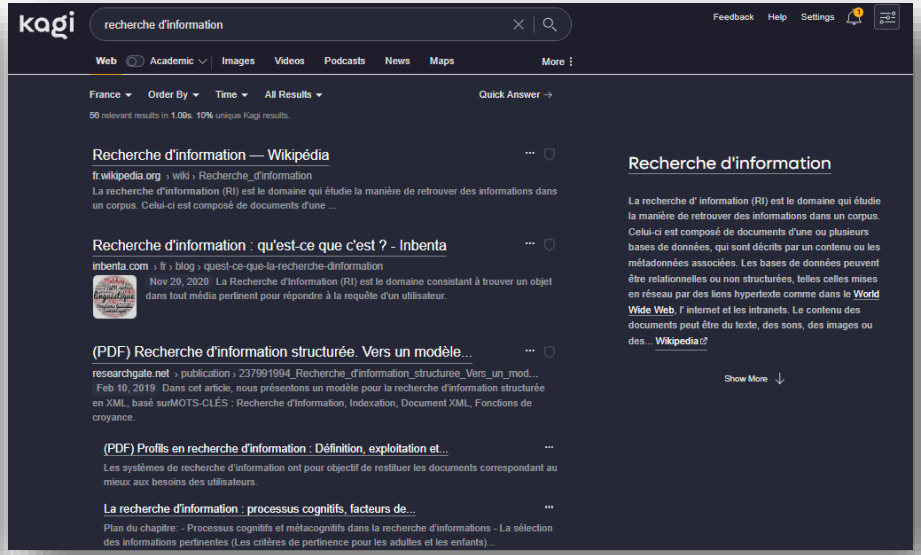
Moteurs de recherche

- 2023 : deux *outsiders* à suivre ?

Brave search : https://search.brave.com/



Kagi : https://kagi.com/



- ouvert en 2021, écosystème Brave (navigateur, etc.), fondé par l'un des cofondateurs de Firefox
- index 100 % autonome
  mais en cours de constitution et ne proposant pour l'instant ni le filtre vidéo ni le filtre images
- fonction « Goggles » pour personnaliser les sources (version bêta)
- possibilité de supprimer la publicité avec l'offre premium

- ouvert en 2022, pour un entrepreneur yougoslave
- nécessite un compte (limité à 100 requêtes en version gratuite)
- index : Google et Bing
- très nombreux tris (ex. : par fraîcheur ou site) et filtres (ex. : *academic*, création de *lenses* personnalisées) et possibilité de booster certains types de résultats
- contenus sémantiques / IA : *summarize page* et *ask questions about document*

Moteurs de recherche

« Chaque outil dispose d'un corpus spécifique
et de fonctionnalités de recherche différentes. [...]

Il est donc crucial d'utiliser **plusieurs moteurs**
pour arriver à un résultat satisfaisant.

D'autre part, il ne faut **pas** utiliser tous ces moteurs
de la même façon »

C. Tisserand-Barthole, *Netsources*, n° 149 (11-12/2020)

Réseaux sociaux

moteurs de recherche « classiques »	réseaux sociaux
indexation du web	interrogation du site en temps réel (fraîcheur et tendances)
filtre algorithmique	filtre humain
popularité (*PageRank*)	sérendipité (signaux faibles)
confiance dans les sources	confiance dans les personnes
	une indexation souvent incomplète par les moteurs de recherche classiques

Réseaux sociaux

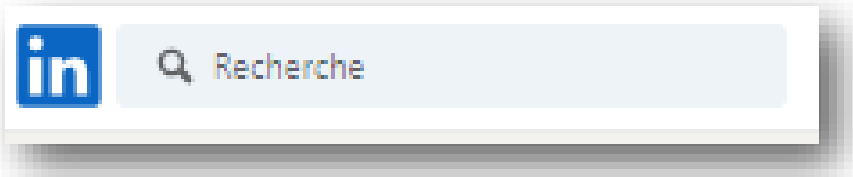
- 2023 : deux réseaux à connaître dans un cadre professionnel et académique

Twitter : https://twitter.com/

- le réseau social avec le plus de contenu professionnel ou académique pertinent
- actualité 2022-203 : rachat par E. Musk
  - disparition de certaines fonctionnalités
  - retour / développement de comptes problématiques
  - mais pas de désertion importante
  - compte nécessaire pour pouvoir interroger le réseau
- des fonctionnalités de recherche (booléens, filtres, recherche avancée) et de tris avancées permettant des recherches fines et maîtrisées

pour aller plus loin sur Twitter : support de formation

LinkedIn : https://www.linkedin.com/



- réseau social avec d'importants contenus professionnels ou académiques, mais variable selon les domaines
- compte nécessaire pour pouvoir interroger le réseau
- des fonctionnalités de recherche et de tris peu développées
- phénomène de « boîte noire » (dont résultats non exhaustifs ou aléatoires)
- les abonnements sont pris en compte dans la présentation des résultats

« Quoi qu'on en dise, [Twitter est] toujours à ce jour le réseau social le plus important pour les professionnels de l'information, que ce soit en termes de **contenus** ou de **fonctionnalités**. »

(C. Tisserand-Barthole, *Bases,* 01/2023)

Réseaux sociaux

- les réseaux sociaux en 2023

des informations complémentaires des moteurs de recherche

- informations en temps réel et actualités, avis et commentaires
- informations personnalisées (abonnements, recherche dans sa communauté)
- fonctionnalités de recherche spécifiques

des services de plus en plus fermés

- nécessité de plus en plus d'avoir un compte pour accéder au contenu et aux outils de recherche
- des algorithmes de plus en plus imposés

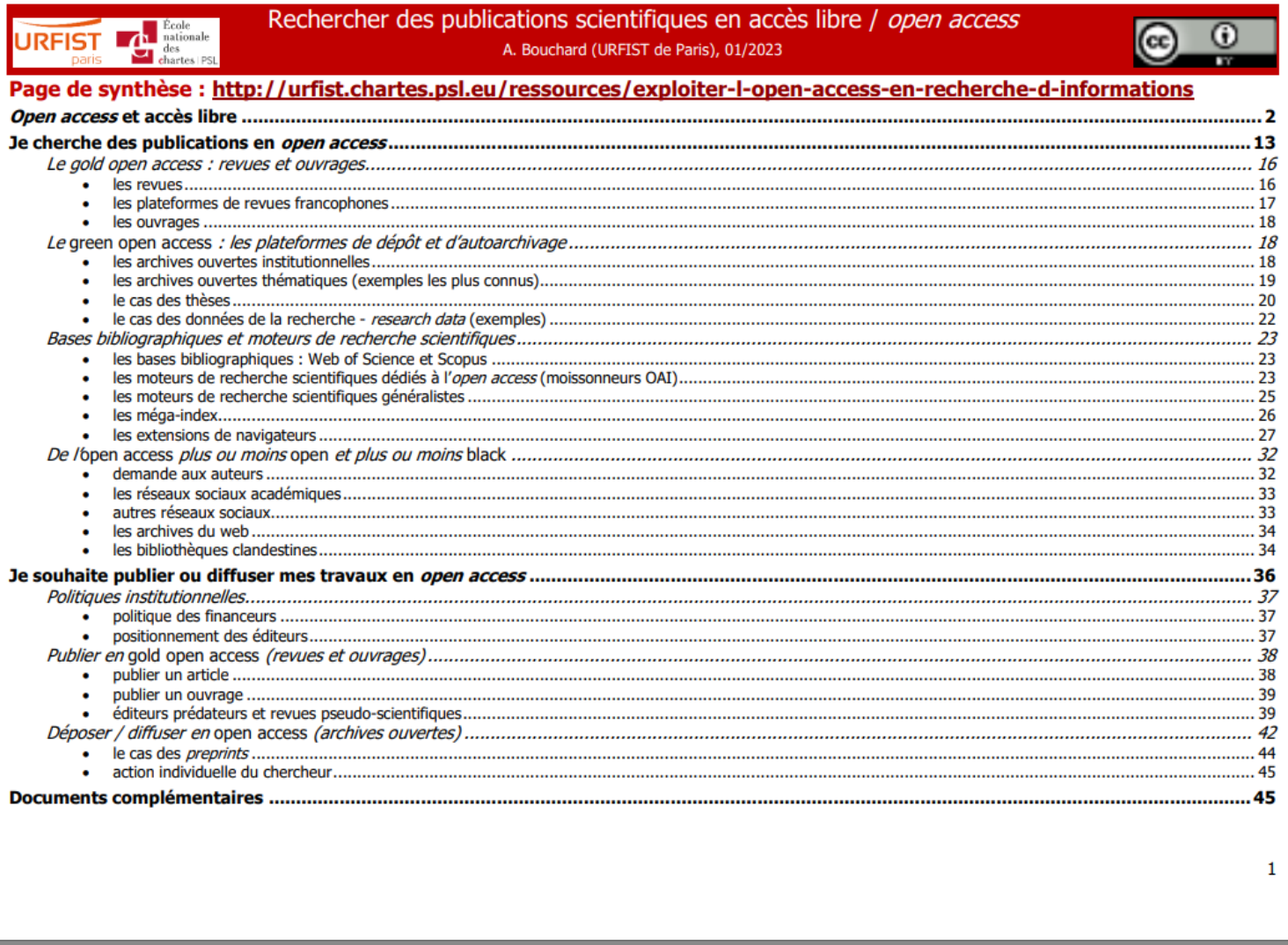
le maintien de Twitter et le développement de LinkedIn ?

- Twitter : maintien de la communauté et de la plupart des fonctionnalités de recherche malgré le rachat
- LinkedIn : de plus en plus de contenus professionnels (y compris par des acteurs ayant quitté Twitter)

toujours l'intérêt de l'équation site: sur Google ?

Moteurs de recherche scientifiques

moteurs de recherche « classiques »	moteurs de recherche scientifiques
indexation du web	indexation de corpus spécifiquement académiques
fonctionnalités de recherche générales	fonctionnalités et filtres spécifiquement liés aux publications (auteur, titre...), aux types de documents scientifiques (*preprints*, brevets...) ou aux acteurs (auteurs, institutions...)
des acteurs généralement commerciaux (Google, Microsoft...)	des acteurs plus divers (Google, fournisseurs de service, services de documentation)



Rechercher des publications scientifiques en accès libre / *open access*

A. Bouchard (URFIST de Paris), 01/2023

Page de synthèse : http://urfist.chartes.psl.eu/ressources/exploiter-l-open-access-en-recherche-d-informations

support complémentaire

https://urfist.chartes.psl.eu/ressources/exploiter-l-open-access-en-recherche-d-informations

(v. 2023)

Moteurs de recherche scientifiques

- les moteurs de recherche scientifiques en 2023

des *outsiders* qui cherchent à se démarquer

- un outil gratuit toujours central : Google scholar
- des acteurs plus spécialisés (ex. : *open access* [Brave], différents types de productions scientifiques, ajouts de données bibliométriques [Dimensions, The Lens])

l'importance de diversifier les outils gratuits et payants

- complémentarité des couvertures et des fonctionnalités de recherche
- accès au texte intégral payant

→ **voir avec son institution**

l'intérêt des extensions de navigateurs

l'importance de rester vigilant face à la notion d'*open access*

- quels accès au texte intégral (texte, images)
- quelles versions disponibles (*preprints*, *postprints*...) pour quelle validation scientifique
- interrogation du texte intégral ou des seules métadonnées

le développement d'outils utilisant de l'intelligence artificielle

- sémantique et clustering
- suggestions
- tâches d'assistance

Intelligence artificielle

- évolutions de la recherche internet

recherche multimodale et universelle : élargir les bases de données
- des données de plus en plus nombreuses (*big data*)
- une diversification des contenus

recherche sémantique : se faire comprendre par la machine comme par un humain
- une compréhension des concepts (vocabulaires, entités, relations…, ex. *knowledge graphs*)
- une compréhension du langage naturel (ex. : NLP – *natural language processing*)
- des modèles de langages (ex. : LLM - *large language models*)

recherche contextuelle et conversationnelle : communiquer avec l'humain
- une compréhension du contexte de la requête et de l'intention de l'internaute
- un historique des échanges précédents

+ un entraînement (*deep learning,* réseaux de neurones et travailleurs du clic)

pour comprendre les évolutions en cours

https://urfist.chartes.psl.eu/ressources/evolutions-des-moteurs-de-recherche-sur-internet
(2016)

https://urfist.chartes.psl.eu/ressources/tendances-de-la-recherche-web-2020
(2020)

Intelligence artificielle

- évolutions des outils de recherche

IA conversationnelle : répondre à partir de bases de données
- chatbots
- agents conversationnels (ex. : Siri, Alexa, Cortana)

IA générative (GenIA) : créer du contenu original à partir d'entraînements pour fournir des contenus similaires aux données d'apprentissage

ChatGPT

Chat <-> chatbot
GPT : *Generative Pre-trained Transformers* (modèles de langage pré-entraînés)

Intelligence artificielle

« ChatGPT est capable de générer des réponses à des questions, de compléter des phrases, de traduire des textes, d'écrire des articles et de tenir des conversations avec des humains.
Il peut également synthétiser des textes suivant un ensemble de contraintes, telles que le ton, le style et le sujet. Il a également été utilisé pour différents usages, tels que la génération de sous-titres pour des vidéos et la création de chatbots. »

Wikipédia

Intelligence artificielle

moteurs de recherche généralistes	chatbots IA de type ChatGPT
bases de données mises à jour régulièrement	données préformatées à partir de bases de données parfois anciennes
approche par mots-clés (et de plus en plus langage naturel)	approche par énoncé (langage naturel)
requête (avec booléns, filtres et tris)	« prompt » plus ou moins précis
liste de résultats (et de plus en plus réponses)	réponse (parfois sans source)
pas de conservation de l'historique d'une requête à l'autre (itérations)	recherche conversationnelle (échanges)
→ moteurs de recherche	→ moteurs de réponses

Intelligence artificielle

Prompt : "photographic close up portrait of a smiling male robot, hyperrealistic, cinematic lighting" - Negative Prompt : "hair, skin" - Leonardo.ai / DreamShaper v5

l'importance du « prompt »

Prompt : ""Portrait of an asiatic man with face tattoos wearing a beanie and earrings studio lighting cinematic angle 85mm 1.4 lit from above ultra detailed" Dall-E

source

Intelligence artificielle

les « hallucinations » de l'IA

source

Intelligence artificielle

- outils de recherche IA en 2023

de nouvelles manières de chercher de l'information

- recours au langage naturel et à l'incrémentation des résultats précédents
- une perte de maîtrise (absence de tris / filtres explicites)

des tentatives généralement prudentes

- des fonctionnalités non stabilisées
- de l'IA sur certains points seulement

beaucoup trop d'informations incorrectes voire fausses

- phénomène de boîte noire (sources utilisées ?, algorithmes de classement ?, exhaustivité ?)
- des réponses non sourcées
- des réponses et des citations biaisées voire inventées

des réponses généralement superficielles

- de simples copier-coller ou paraphrases de contenus internet
- une prépondérance des premiers résultats des moteurs de recherche/bases de données bibliographiques

→ ne peuvent suffire seuls pour un état de l'art ou une revue systématique
→ nécessitent de savoir se poser des limites

Intelligence artificielle

- point de vue du documentaliste

  « L'IA générative oui, mais pas pour trouver l'information dans les moteurs. [...] L'IA générative a un rôle à jouer dans la recherche d'information, mais cela ne pourra être qu'en amont en phase de recherche de mots-clés, par exemple ou après avoir identifié les informations et les sources pertinentes. » (C. Tisserand-Barthole, *Bases*, 04/2023)

- point de vue du chercheur

  « *Although ChatGPT provides succinct answers to most queries, its lack of reliability, transparency, and up-to-date knowledge compared with conventional search engines is a major drawback, particularly for health-related queries.*» (F. Eggmann et M. B. Blatz, *Compendium...*, 04/2023)

Résumé

- la recherche d'information en 2023

  - un outil toujours central : Google
  - une diversification des autres acteurs et des fonctionnalités qui évoluent (moteurs de recherche « classiques » + réseaux sociaux + moteurs de recherche scientifiques + outils de recherche IA)
  - des questions sur les modèles économiques et la pérennité des services
  - des questions sur les résultats (algorithmiques, chronologiques, personnalisés)
  - de nouvelles manières de rechercher (langage naturel)
  - une interrogation sur la plus-value de l'intelligence artificielle

→ diversifier les outils

→ être de plus en plus vigilant sur les sources couvertes et les résultats (crédibilité, classement, fraîcheur, pertinence, exhaustivité)

Pour aller plus loin

- méthodologie et outils
  - support de la formation : Aline Bouchard. Recherche d'informations sur internet (perfectionnement). MAJ 2021, à compléter par la carte : https://framindmap.org/c/maps/446941/public. https://urfist.chartes.psl.eu/ressources/recherche-d-informations-sur-internet-perfectionnement.
  - Véronique Mesguich. *Rechercher l'information stratégique sur le web.* Sourcing*, veille et analyse à l'heure de la révolution numérique.* 2e éd. ADBS-éd. De Boeck, 2021. 236 p. (« Information et stratégie »).

- pour suivre l'actualité du domaine, les incontournables :
  - ressources de FLA Consultants, plus ou moins gratuites : revues *Bases* et *Netsources*, brèves : https://www.bases-netsources.com/ et compte Twitter : https://twitter.com/BasesNetsources
  - blog d'Aaron Tay, https://musingsaboutlibrarianship.blogspot.com