Qu’est-ce que le Web et pourquoi les moteurs de recherches n’y couvrent que 10 % ?

écrit par Cachou | 20 février 2024

Internet est un outil extraordinaire dans la communication et l'information. Tout le monde le sait. Mais comment fonctionnent les recherches que l'on peut y faire ? Quels sont les résultats présentés, dans leurs nombres et objectivités ?Là, c'est un peu moins évident. Voici simplement et résumés les processus existants.

Le Web est si grand et développé qu'il est maintenant impossible de savoir relativement précisément le nombre de sites Web et de pages qu'il contient.

Qu'est-ce que le Web ?

Tout d'abord précisons la différence entre Internet et le Web

Internet (contraction de Inter Network) est un réseau

informatique qui relie des ordinateurs entre eux à l'échelle du monde. Ce vaste réseau informatique a plusieurs systèmes comme le FTP (protocole de transfert de fichier), la téléphonie VOIP, service l'envoi de courrier électronique (SMTP, POP, IMAP) et le web.

Le web (abréviation de World Wide Web ou toile mondiale), lui, est un **système** hypertexte fonctionnant sur Internet, parmi d'autres. Ce système utilise les techniques de l'hypertexte, c'est-à-dire des hyperliens ou liens qui te permettent de surfer d'un document à un autre.

Le Web n'est qu'un système d'Internet, distinct d'autres systèmes comme le courrier électronique, la messagerie instantanée et le partage de fichiers en P2P.

Invention du World Wide Web

Au cours des années 1960, l'Advanced Research Projects

Agency (A.R.P.A.), l'agence de recherche du ministère américain de la Défense, avait piloté et financé le développement d'un réseau de communications, Arpanet, reliant les ordinateurs des divers laboratoires universitaires travaillant sous contrats militaires. Dès 1967, quelques liaisons fonctionnaient, transmettant bientôt des courriers électroniques ou e-mails.

En 1974, Robert Kahn et Vinton Cerf conçoivent le protocole T.C.P./I.P. (Transmission Control Protocol/Internet Protocol) – voir plus loin – qui facilite ces échanges. En 1983, le réseau Arpanet adopte officiellement la norme T.C.P./I.P. : c'est le démarrage d'Internet, avec à l'époque environ un millier de postes utilisateurs.

Puis en 1989 Tim Berners-Lee crée le World Wide Web (les célèbres www.), une application conviviale de consultation à distance de pages d'informations multimédias, qui permet à Internet de s'ouvrir à un large public. Grâce au Web (World Wide Web), Internet permet à des milliards d'humains d'accéder à des milliards de documents. Ceux-ci contiennent des images, des textes, des vidéos, des sons… sur à peu près tous les sujets.

Le Web, pour sa totalité, est divisé en deux parties :

- **La surface Web, partie « visible » du Web**
- **Le Deep Web (partie « invisible » du Web) constitué du :**
  - **Deep Web en lui-même**
  - **Darknet (réseau informatique chiffré intégrant des fonctions d'anonymisation)**
  - **Darkweb (contenu du web qui existe sur les**

réseaux darknet)

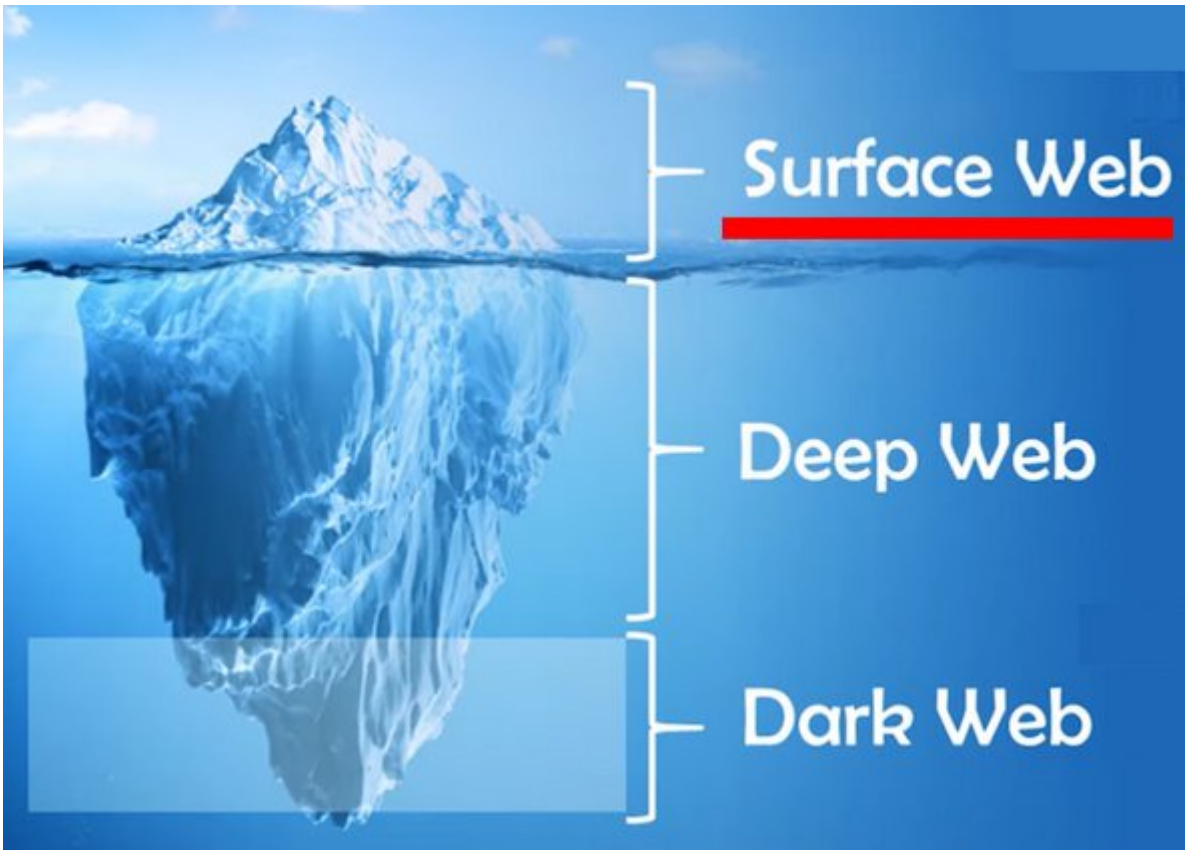
1 - La surface Web

Elle est appelée aussi « *web surfacique* ». Elle constitue le niveau 1 du Web dans sa totalité. **Cette partie regroupe tout ce qui est visible pour tout le monde et très facilement, et représente environ 10 % de tout l'ensemble du Web.**

Imagine qu'Internet soit une ville. Il y a des espaces publics ouverts à tout le monde, comme les rues, les boulevards, les parcs, et tous figurent sur un plan. N'importe qui peut s'y rendre et regarder ce qui s'y passe.

Sur Internet, ces lieux publics sont ce que l'on appelle le web surfacique. Ce sont des pages Web, des applications Web et tout autre élément en ligne que les moteurs de recherche peuvent indexer. On y trouve donc des documents, des fichiers multimédias et bien d'autres choses. N'importe qui peut les trouver grâce à un moteur de recherche et les visualiser sans payer, sans s'inscrire et sans installer de programme spécifique.

Si l'on compare le Web a un iceberg, la surface Web représente la partie visible de l'iceberg. C'est-à-dire une toute petite partie de l'ensemble du Web.



2 – Le Deep Web

En plus des espaces publics, les villes ont des lieux privés auxquels on accède sur invitation, avec un ticket ou encore grâce à un laissez-passer ou par connaissances personnelles. C'est notamment le cas des maisons, des quartiers d'affaires, des clubs privés, des cinémas, etc. En général, aucune carte ouverte au public ne te montre ce qui se passe dans ces lieux.

Le Web regorge aussi de recoins que Google, Bing, Yahoo, DuckDuckGo, Qwant, Startpage.com, et bien d'autres moteurs de recherches ne peuvent pas voir. En effet, ces endroits ne sont pas indexés. Ils appartiennent à la partie du web appelé **Deep Web** appelé aussi le web sombre ou web profond. Il se compose principalement de toutes les pages Internet qui ne peuvent pas

être recherchées et ouvertes par les moyens habituels, et que les bot-cartographes ne peuvent pas indexer.

Si un site te demande de saisir un identifiant, un mot de passe, ou tout autre captcha pour y accéder, alors le bot de recherche est incapable d'en savoir plus sur le contenu du site. C'est le but de ces éléments de se débarrasser des bots.

Si un article n'est disponible que sur abonnement, un bot ne peut pas l'ouvrir et ne peut pas l'indexer puisqu'il n'a pas de compte ni d'argent pour payer l'abonnement. Il faut saisir un mot de passe pour accéder à l'article ? Là encore, le bot n'a pas de chance : il n'en connaît aucun.

Tous les sites web peuvent contenir des pages dans le web profond. On y trouve notamment les Webmails, les banques en ligne, les sites dont une partie est payante.

Par exemple, le service payant de streaming Netflix contient une grande partie des pages dans le web profond. De toute évidence, les moteurs de recherche n'opteront pas pour un abonnement mensuel pour indexer le catalogue de Netflix.

L'URL d'un email que tu envoies ou que tu reçois, c'est le Deep Web, Aucun moteur de recherche ne peut l'indexer et donc personne ne peut le trouver.

Même si quelque chose qui peut être ouvert est introuvable alors il appartient tout de même à la catégorie Deep Web. Par exemple, tu configures Facebook pour cacher ton profil et éviter qu'il n'apparaisse dans les moteurs de recherche, même si un bot de recherche le trouve, il devra l'ignorer.

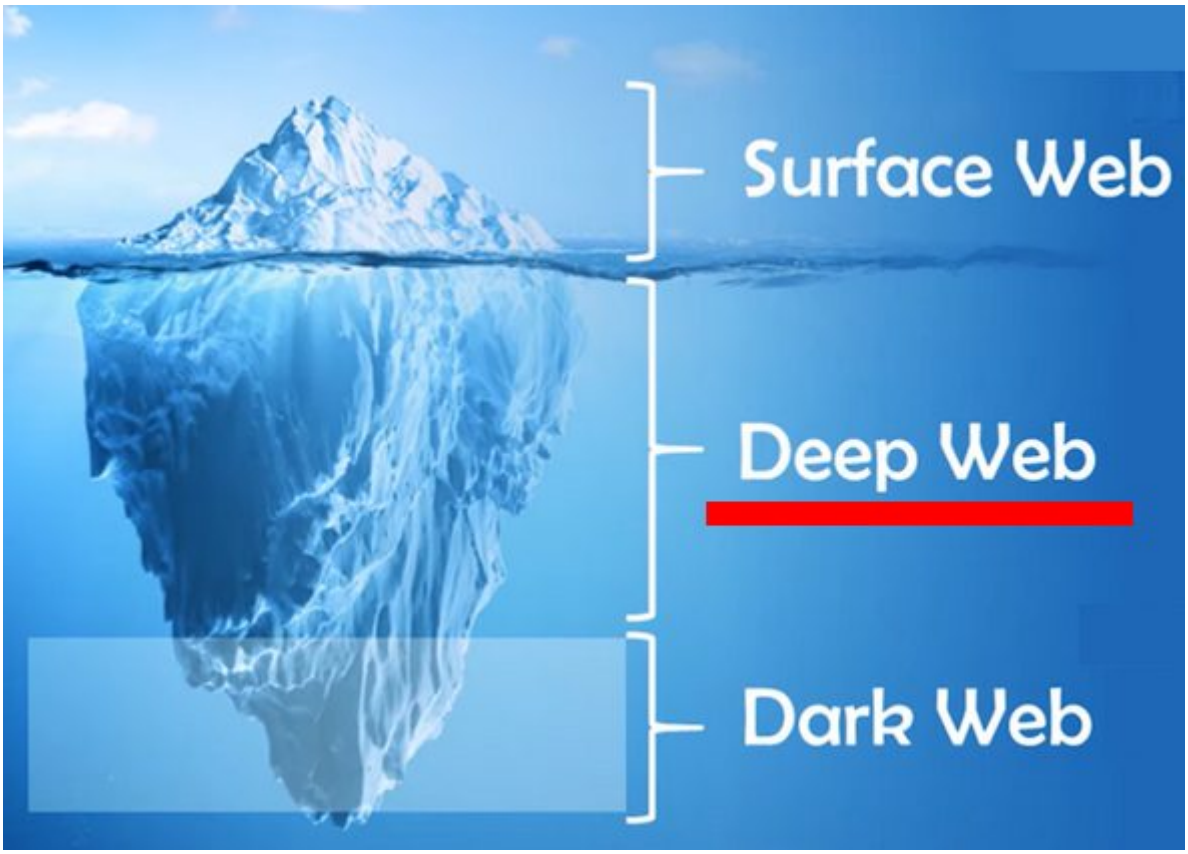
Enfin, le Deep Web fait aussi référence au contenu pour lequel il n'existe aucun lien sur le web surfacique ou visible. Un bot de recherche ne sait pas que ce contenu existe. Il trouve de nouvelles pages en suivant les liens des autres pages déjà indexées.

La voiture de Google Street View ne peut pas entrer dans le jardin privé d'une maison. Il en est de même pour les bots avec les contenus sans lien.

Comme tu peux le constater, l'essentiel du Deep Web est fait de pages Web et de documents inoffensifs, parfois utiles, que la plupart d'entre nous utilise. Tu vas sur ton compte bancaire, tu vas sur ton compte de la Redoute, tu es dans le Deep Web.

Mais il en va autrement avec une autre petite partie du Deep Web, le Darkweb (voir plus loin).

Le Deep Web constitue, et de très loin, la principale partie du Web. Revenons à notre iceberg :



3 – Le Dark Web

Ah, là là ! Le Dark Web, rempli de mystères et de fantasmes… et, il est vrai, ce n'est pas faux.

Le Dark Web forme une petite partie du Deep Web, lequel comme ci-dessus expliqué, forme la partie immergée du web et donc dite « invisible ». Il permet de naviguer de manière totalement anonyme contrairement à la Surface Web et le Deep Web. Même si pour ces deux derniers, il existe les VPN. Mais bon…

Ainsi, un internaute, de façon totalement anonyme peut communiquer, acheter, se renseigner sur l'état de choses avec le Dark Web.

Pour fonctionner et présenter des contenus, le Dark Web a

besoin d’une infrastructure étant un ensemble des réseaux chiffrés permettant d’accéder à ce contenu. Cette infrastructure s’appelle le Darknet.

Ainsi, le Darknet permettant au Dark Web de fonctionner forment, à eux deux, une petite partie du Deepweb.

Alors, me diras-tu, pourquoi commencer ce paragraphe par « Ah, là là ! Le Dark Web, remplie de mystères et de fantasmes… » ?

Parce que le Dark Web est un joujou dangereux, très risqué, et réservé à une toute petite partie d’individus qui baignent dans ce milieu. Explications.

Le Dark Web n’est pas indexée par les moteurs de recherche. Pour y accéder il faut donc faire appel à des réseaux particuliers et surtout connaître l’adresse exacte de ce que l’on cherche. Il existe plusieurs outils permettant d’y accéder comme Freenet, I2P mais le réseau du Darknet le plus connu est Tor. Ce sont des navigateurs spécifiques du Dark Web et en les téléchargeant sur son ordinateur tout le monde peut y aller facilement.

Le Dark Web est surtout connu pour des usages illégaux comme

le trafic de drogues, des armes, des faux papiers, des données bancaires volées ou d'autres marchandises illégales. On y a accès aussi à des recrutements de mercenaires, des contacts avec des prostituées, des réseaux de pédophilies. On y trouve aussi des forums de discussions ou autres médias pour les pédophiles, les terroristes ou amateurs de pornographie extrême.

Le Dark Web est très difficilement contrôlable par les polices du Web du monde entier. Cela dit, aller sur le Dark Web n'est pas un délit tout dépend ce que l'on y fait. Car le Dark Web n'est pas une zone de non-droit – enfin en théorie – malgré tout ce qui s'y passe d'illégales.

Pour ceux qui fréquentent cette partie du web caché il faut qu'ils respectent des procédures drastiques pour préserver leur anonymat. C'est la base incontournable pour aller ensuite sur le Dark Web.

Est-ce légale d'accéder au Dark Web ?

Oui, il est tout à fait légal de consulter le Dark Web. Aujourd'hui, de nombreux politiciens et journalistes utilisent ce réseau (darknet) pour exprimer leurs opinions en toute liberté, tout en préservant leur vie privée. En plus, certains experts en cybercriminalité s'en servent pour surveiller le mouvement des hackeurs et des criminels.

L'accès au Dark Web devient illégal si tu t'en sers pour accéder à des activités illégales ci-dessus citées. Mais en fait, les lois et les réglementations liées à l'accès du Dark Web sont différents selon les pays.

Quels sont les risques en se rendant sur le Dark Web ?

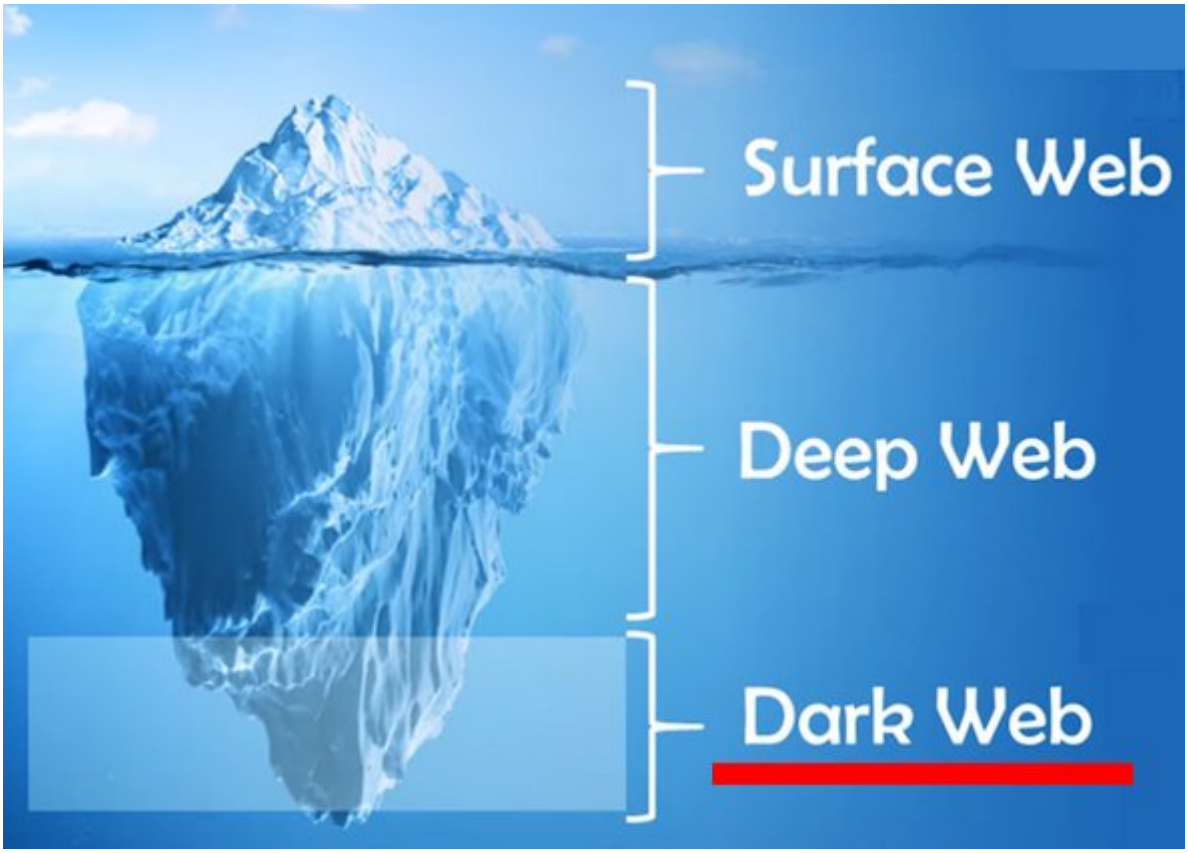
Il y a des risques de sécurité et de confidentialité si tu accèdes sur le Dark Web sans protection efficace. Tu peux faire l’objet de pirates informatiques qui volent ton identité pour l’utiliser sur d’autres personnes ainsi que tout ce que contient ton ordinateur. Cela s’appelle le phishing.

Et puis il y a les risques juridiques. Comme dit ci-dessus si tu vas sur le Dark Web sans toucher à une activité illégale, tu ne risques rien. Mais comme l’essentiel des activités sont totalement illégales, mieux vaut s’abstenir.

En théorie, pour aller sur le Dark Web et se protéger il faut impérativement avoir un VPN, plus un navigateur respectant l’anonymat (comme Tor par exemple) lequel navigateur garantit à peu près la sécurité en rendant très difficile la

traçabilité de ton activité sur le Dark Web, ainsi qu'un antivirus performant contre le phishing.

Revenons à notre iceberg :



Quels sont les procédés de demande et réception des informations du Web ?

Quand tu demandes une information sur Internet, celle-ci est reliée automatiquement à ton ordinateur par ton adresse IP. C'est l'équivalent de ton adresse où tu habites quand tu envoies une enveloppe par la Poste et que tu mets ton adresse d'expéditeur.

Une fois ton site Internet choisi, le DNS de ton FAI

(Fournisseur d'Accès Internet) transforme l'adresse de ton site en adresse IP, comme ton ordinateur. Il sait donc où acheminer tes demandes d'informations. Encore faut-il que ces dernières arrivent telles que tu les as demandés et dans le bon ordre. C'est le rôle du TCP. Et comme dit ci-dessus une fois que le TCP a tout mise en ordre, il envoie tout à ton site vers son adresse IP. Cela constitue le protocole de communication appelée TCP/IP.

Une fois que le TCP/IP a tout mis en ordre, il faut transférer les données par les deux protocoles de transfert de données : HTTP (version standard) et HTTPS (version sécurisée).

Pour te faciliter la tâche il existe les navigateurs Web. Il y en a un certain nombre, par exemple : Chrome, Edge, Safari, Mozilla Firefox, Opera, Vivaldi et d'autres… Leurs rôle est de prendre en considération ta demande et en te présentant toutes les réponses.

Quels sont les moyens physiques de communication du Web ?

Ils sont nombreux et variés. On y trouve :

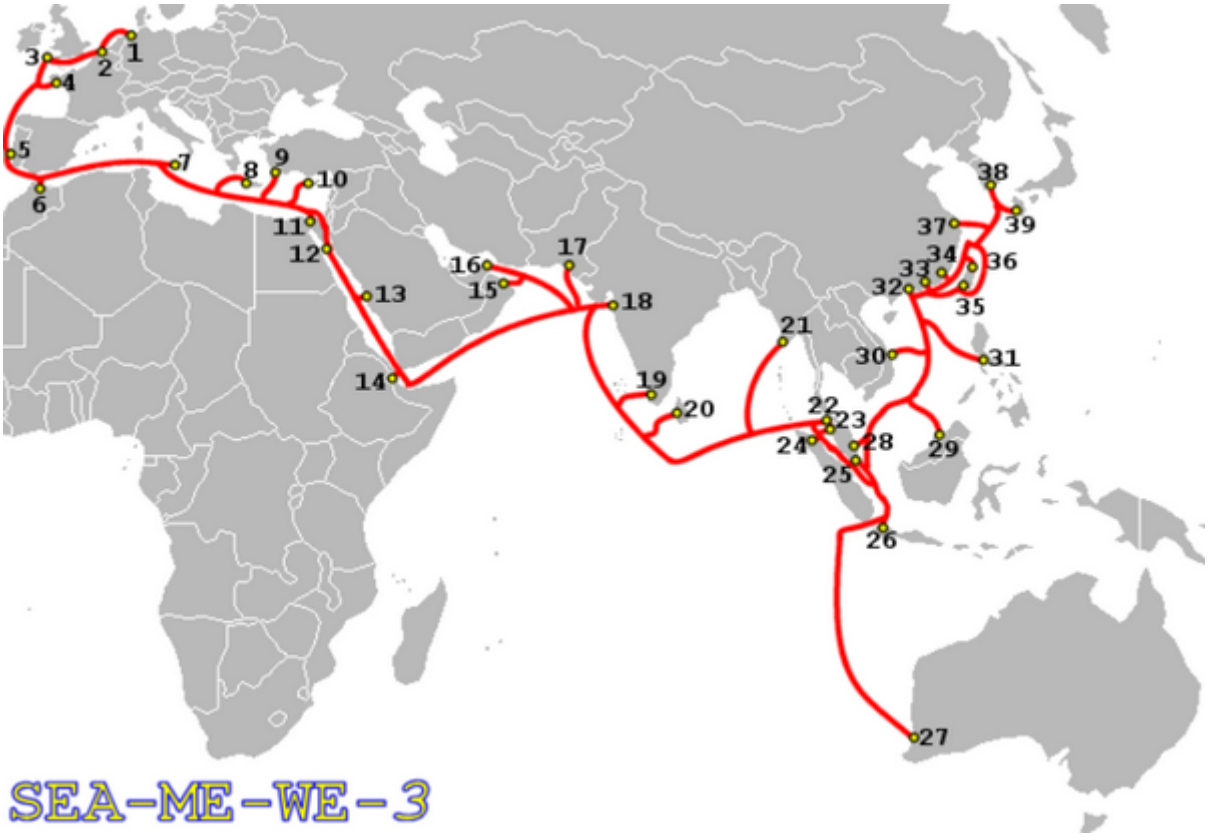
1) Les câbles sous-marins

Une grande partie du trafic Internet mondial transite par des câbles sous-marins qui relient les continents. Ces câbles sont posés au fond des océans et permettent des débits élevés entre les différents pays.

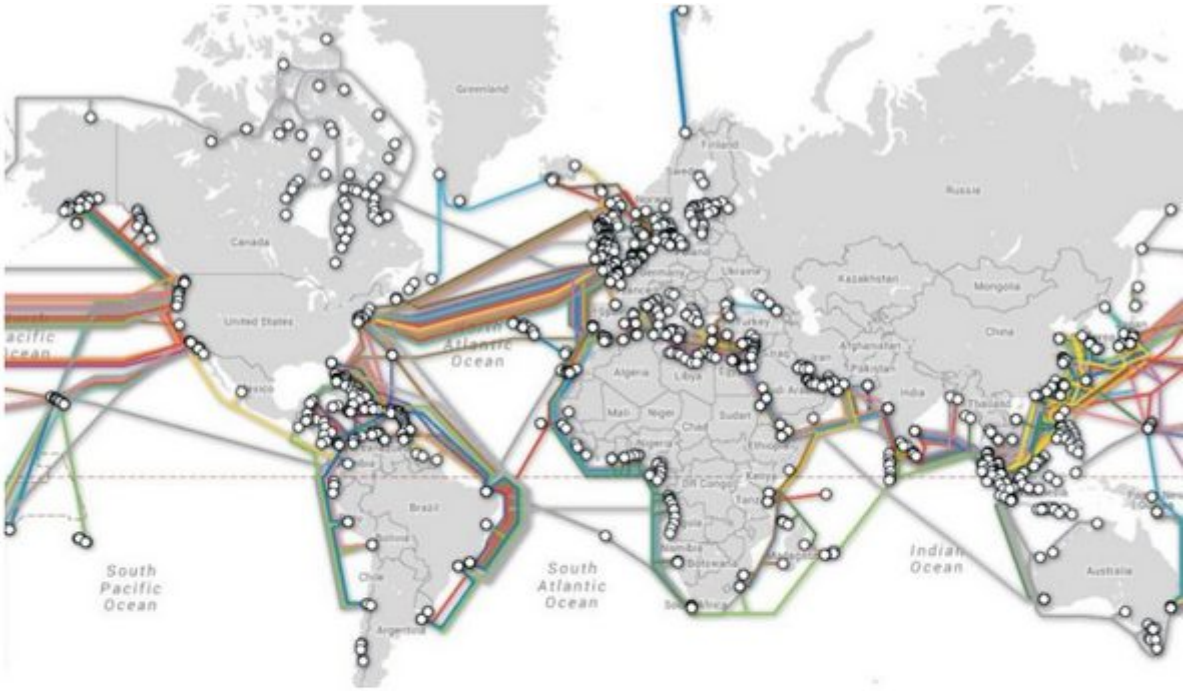
Des centaines de câbles sous-marins, longs de milliers de kilomètres, relient les continents et assurent environ 90 % des communications mondiales. Sans ces câbles, aucune donnée ne pourrait transiter entre les serveurs et nos appareils.

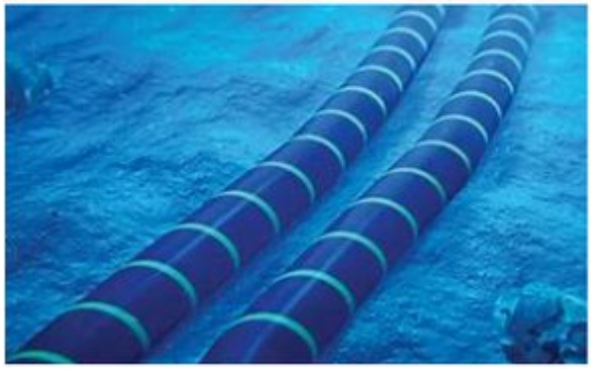
Selon le dernier relevé effectué par la société de conseils en télécommunications TeleGeography, il y aurait actuellement 448 réseaux qui s'étendraient sur 1,2 million de kilomètres dont les célèbres câbles transatlantiques.

Le câble sous-marin en fibre optique le plus long est le SEA-ME-WE 3 (pour South-East Asia – Middle East – Western Europe). Il s’étend sur 39 000 km et interconnecte l’Asie du Sud-Est, le Moyen-Orient et l’Europe de l’Ouest.



Les autres SEA-ME-WE (4 et 5) construits postérieurement au SEA-ME-WE 3 sont plus courts. Le SEA-ME-WE 4 court sur 18 800 km tandis que le SEA-ME-WE 4 court sur 20 000 km.





1) Les réseaux téléphoniques

Pendant très longtemps ils ont assuré les liaisons terrestres au niveau des départs et des arrivées des communications. Leurs transports sur les longues distances passent ensuite par les câbles sous-marins essentiellement. Ces réseaux téléphoniques ont constitué, et constituent encore aujourd'hui, l'ADSL. Ils sont progressivement en voie de disparition pour être remplacés par les fibres optiques terrestres.

3) Fibres optiques terrestres

Les fibres optiques sont utilisées pour les liaisons terrestres. Elles offrent une bande passante élevée et sont largement utilisées pour les connexions haut débit. Elles se

développent rapidement et remplacent les réseaux téléphoniques.

4) Réseaux sans fil

- Wi-Fi : les réseaux Wi-Fi utilisent des ondes radio pour connecter des appareils sans fil à Internet. Ils sont couramment utilisés dans les maisons, les bureaux et les lieux publics.
- Bluetooth : le Bluetooth est une technologie sans fil à courte portée utilisée pour connecter des périphériques tels que des écouteurs, des claviers et des souris.
- Réseaux cellulaires : les réseaux 3G, 4G et 5G utilisent des antennes-relais pour fournir une connectivité mobile à grande échelle.

5) Satellites

Certains services Internet, notamment dans les zones rurales ou éloignées, utilisent des satellites pour transmettre des données. Les signaux sont envoyés depuis la Terre vers les satellites en orbite géostationnaire, puis renvoyés vers les utilisateurs. L'utilisation des satellites est actuellement très peu utilisée, mais n'est-ce pas la voie de l'avenir ?

Bye bye et à la prochaine

Professeur Têtenlair