CSET

Dans la jungle : Pratiques exemplaires pour la recherche à partir de sources ouvertes

Ryan Fedasiuk

Avril 2022

Le présent guide vise à permettre aux chercheurs et aux analystes de mieux connaître les outils, les ressources et les pratiques exemplaires en matière de sécurité lorsqu'ils accèdent à des sources ouvertes à des fins de collecte d'information.

Les sources ouvertes que l'on trouve sur Internet peuvent poser de nombreux risques pour les utilisateurs et le matériel qu'ils utilisent. Face à de tels dangers, une vigilance constante s'impose.

Trois principaux éléments doivent être pris en compte lors de la collecte d'information de sources ouvertes. Ces considérations sont classées par ordre de priorité ci-dessous :

1. **Protéger** ses appareils, son réseau et ses fichiers contre les logiciels malveillants ;
2. **Archiver** ses sources pour la postérité ;
3. **Masquer** ses activités des regards indiscrets.

Le présent guide vise à permettre aux chercheurs et aux analystes de mieux connaître les outils, les ressources et les pratiques exemplaires en matière de sécurité lorsqu'ils accèdent à des sources ouvertes à des fins de collecte d'information.

Les règles essentielles de la recherche à partir de sources ouvertes

1. Il faut toujours supposer que la source a été compromise et qu'elle peut présenter un risque d'atteinte à la vie privée.
2. Il faut toujours rester connecté à un réseau privé virtuel (RPV).
3. Il ne faut jamais télécharger de fichier localement.
4. Dans la mesure du possible, on accédera uniquement aux versions mises en cache ou archivées des pages Web.
5. S'il y a lieu, les sources doivent être archivées sans délai.
6. En cas de doute, une analyse doit être effectuée avant de cliquer.

Ressources, outils et pratiques exemplaires

1. RPV

Un réseau privé virtuel (RPV) sécurise les réseaux en masquant l'adresse protocole Internet (IP) de l'utilisateur et en chiffrant l'information transmise par

les appareils. La plupart des services de RPV permettent de sélectionner un serveur par lequel acheminer le trafic Internet. Cela présente l'avantage supplémentaire de camoufler une adresse IP. L'utilisateur qui souhaite une connexion plus rapide choisira un serveur plus près de lui. En revanche, pour bénéficier d'une connexion plus lente susceptible de passer davantage inaperçue, l'utilisateur optera pour un serveur situé à proximité de l'entité visée par ses recherches. Si les recherches visent la Chine, un serveur situé à Hong Kong, Taïwan ou Singapour peut être sélectionné. Il est également possible d'utiliser un service qui permet de contourner directement la « grande muraille électronique » (le *Great Firewall*). Pour la Russie, les pays baltes sont des options de choix. Pour la Corée du Nord, des serveurs RPV basés à Séoul seront plutôt privilégiés.

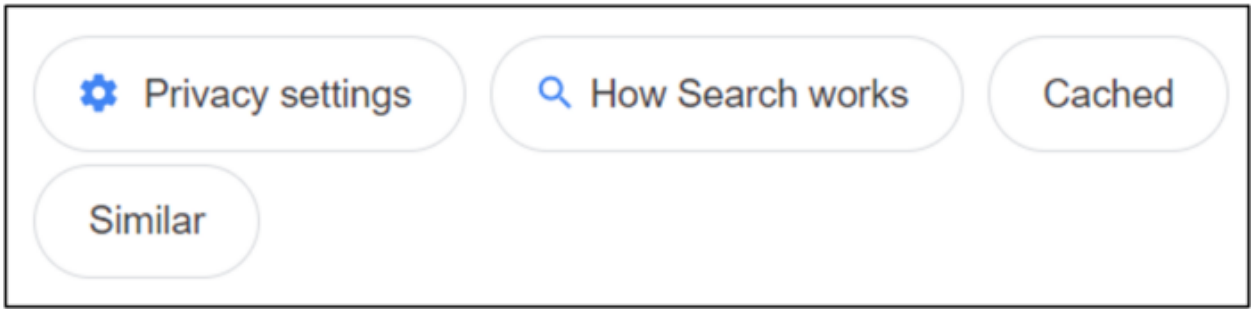
Les options et facteurs à prendre en considération sont nombreux lors du choix d'un RPV, notamment le prix, le nombre de serveurs, la vitesse de connexion, le fait que le service conserve ou non des journaux d'activité de navigation et la saturation, c'est-à-dire si le gouvernement qui fait l'objet de la recherche a bloqué un grand nombre des nœuds de connexion du service consulté.

2. Pages Web stockées dans la mémoire cache

Un moyen sûr d'accéder à une page Web est d'accéder à sa version stockée dans la mémoire cache par Google, plutôt qu'à sa version publiée sur Internet. La page mise en cache est une version antérieure d'un site Web, consulté depuis le moteur de recherche de Google et enregistrée par ce dernier lors de la génération de résultats de recherche et d'aperçus. Si toutes les pages Web ne sont pas mises en cache, la plupart d'entre elles offrent cette option.

Pour accéder aux versions mises en cache, on peut précéder l'adresse URL du mot « cache ». Ainsi, on écrira : **cache:[URL]** directement dans la barre de navigation. On peut aussi cliquer sur les trois points situés à côté d'un résultat de recherche dans Google pour en savoir plus sur cette page.

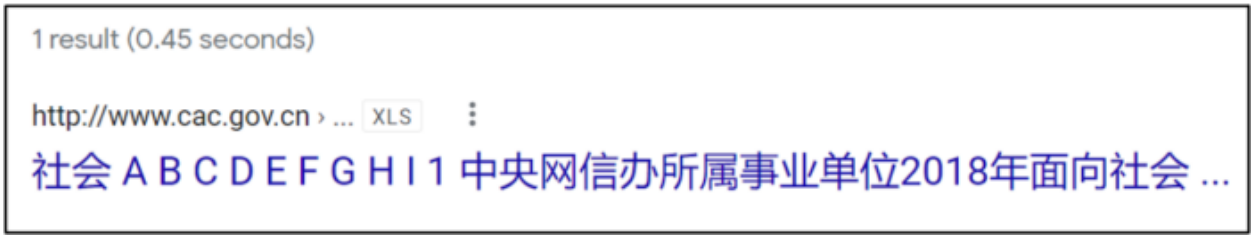
Un bouton « Cached » (en cache) apparaît dans le coin inférieur droit de la fenêtre contextuelle qui s'affiche alors. Il faut cliquer sur ce bouton pour accéder à la page mise en cache.



La version mise en cache d'une page Web comporte une bannière en haut de page qui ressemble à ceci :

L'accès à la version mise en cache d'une page Web n'est pas infaillible. Un propriétaire de site Web peut voir les adresses IP des utilisateurs qui consultent une page Web stockée dans la mémoire cache en consultant certaines images intégrées et d'autres éléments. L'accès à la version en texte seul d'une page stockée dans la mémoire cache ou au code source HTML peut réduire certains de ces risques et permettre de trouver plus rapidement de l'information sur des pages Web lentes à charger.

Les pages Web mises en cache sont particulièrement utiles pour prévisualiser des documents qui, autrement, devraient être téléchargés directement sur l'ordinateur, une **pratique qu'il faut éviter dans toute la mesure du possible**. Prenons l'exemple de ce tableur **.xls** hébergé par la Commission des affaires du cyberespace de la Chine.



Le simple fait de cliquer sur ce résultat de recherche Google entraînerait normalement le téléchargement automatique du fichier sur votre ordinateur - un désastre. Les liens de téléchargement automatique représentent un défi permanent lorsqu'il s'agit de recueillir de l'information à partir de sources ouvertes sur des sites Web étrangers.

Un moyen plus sûr (et plus rapide) d'obtenir l'information est d'accéder à la version mise en cache de la page Web où le fichier est hébergé. Plutôt que de télécharger un document et de l'ouvrir dans Excel, les archives Google le transforment en page Web que l'utilisateur peut consulter dans son navigateur.

Download (18k) Link to this page Edit a copy online

Google automatically generates this HTML view of the file http://www.cac.gov.cn/1122610936_15224000328681n.xls as we crawl the web.

Google is neither affiliated with the authors of this page nor responsible for its content.

社会

	A	B	C	D	E	F	G	H	I
1	中央网信办所属事业单位2018年面向社会公开招聘工作人员职位信息表								
2 3	用人单位	岗位类别	职位代码	职位简介	招聘人数	专业方向	学历学位	资格条件	备注
4	网络安全应急指挥中心	人力资源管理	003	负责人力资源管理日常工作	1	人力资源管理、行政管理等相关专业	全日制大学本科以上学历及学士以上学位	中共党员；熟悉人力资源管理各模块工作；具有一定的组织协调能力和文字综合能力	
5		会计	004	负责会计等财务日常工作	1	会计学、财务管理等相关专业	全日制大学本科以上学历及学士以上学位	中共党员；熟悉财务法律制度和财务工作；具有一定的沟通协调能力；有会计中级以上职称优先	

Cette stratégie fonctionne pour tous les types de fichiers courants (.doc, .pdf, .xls et .xlsx, entre autres), mais provoque parfois des erreurs de formatage (en particulier pour les PDF).

3. Services d’archivage

L’archivage des sources est extrêmement important. Il est fréquent que les sources d’information disparaissent, et que les liens vers les sites Web originaux deviennent périmés quelques jours — voire quelques heures — après la publication d’une recherche. Or, il y a plusieurs raisons de vouloir archiver un site Web, outre la volonté de garantir l’accès futur au matériel.

- Les services d’archivage peuvent remplir une fonction semblable à celle d’une page Web mise en cache en permettant de visualiser une version plus sécuritaire de la page. (Il est également possible d’archiver une page Web stockée dans la mémoire cache par Google, plutôt que la source originale, pour une protection renforcée.)
- Certains services d’archivage, comme Wayback Machine (voir ci-dessous), indiquent si *quelqu’un d’autre* a déjà archivé la page, une information qui peut s’avérer utile.
- Certains services d’archivage génèrent des liens uniques et affichent l’heure exacte à laquelle ces liens ont été générés. Cela peut s’avérer utile en cas de litige pour plagiat et/ou pour assurer le suivi de calendriers de projets.

En particulier, deux services d'archivage gratuits ont été adoptés par la communauté de recherche des sources ouvertes. Il s'agit de :

- The Internet Archive (Wayback Machine, en anglais seulement) : https://web.archive.org/save/
- Archive Today (en anglais seulement) : https://archive.vn/

Il vaut souvent la peine d'enregistrer les documents ayant une valeur particulière dans plusieurs services d'archivage.

Veuillez noter que la plupart des services d'archivage transmettent un « ping » au site Web en utilisant une adresse IP établie aux États-Unis. Cela peut nuire à vos efforts visant à demeurer furtif, par exemple si vous avez opté pour un serveur virtuel en Chine ou en Russie. **Il importe également de noter que les propriétaires de sites Web peuvent rompre rétroactivement des liens déjà archivés.** C'est pourquoi les services d'archivage numérique sur le Web ne sont pas toujours la meilleure option.

Pour garantir un maximum de confidentialité, de sécurité et d'accès à long terme, il est souvent avantageux d'enregistrer des copies des pages Web en format PDF sur votre ordinateur, puis de les verser sur le nuage ou un disque dur externe. **Ce n'est pas la même chose que de télécharger un PDF à partir du site Web lui-même, pratique qu'il convient d'éviter dans toute la mesure du possible.** Au contraire, il faut suivre les étapes suivantes au moment de consulter une page Web :

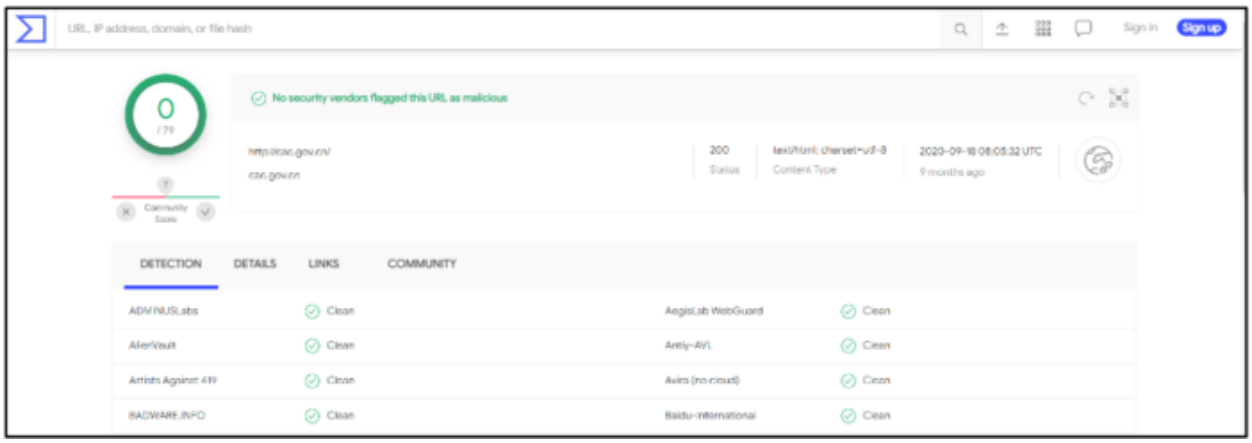
- Il faut d'abord tenter d'« imprimer » la page Web en ouvrant l'interface d'impression (CTRL+P sur un PC ; CMD+P sur un Mac).
- Ensuite, au lieu d'imprimer la page, il faut changer la destination pour « Enregistrer en format PDF ».
- Enfin, l'utilisateur songera aussi à enregistrer le fichier sauvegardé sur une clé USB ou à le verser sur un nuage sécurisé comme Google Drive.

4. Analyse d’adresses URL et de fichiers

Il arrive parfois qu’une source d’information offrant un potentiel intéressant résiste à l’archivage et ne s’accompagne pas d’une page Web mise en cache. Il est risqué de cliquer directement sur ce type de lien, mais vous pouvez faire preuve de diligence raisonnable. **En cas de doute, il conviendra de procéder à une analyse avant de cliquer**.

VirusTotal (en anglais seulement) est un service gratuit qui recherche les logiciels malveillants dans les fichiers et les adresses URL en les comparant à des dizaines de logiciels antivirus, y compris des marques grands bien connues comme AVG, BitDefender et Kaspersky.

VirusTotal recueille de l’information sur les fichiers et les URL téléchargés dans sa base de données. Il s’agit d’une plateforme d’essai pour les services antivirus. VirusTotal a accès à plus de 79 services antivirus, car il fournit de l’information diagnostique permettant d’améliorer les produits antivirus à partir des analyses générées par les utilisateurs. Les utilisateurs n’ont pas besoin de créer un compte.



5. Bac à sable de navigateur

Si vous vous installez pour une longue séance de recherche d'information, il est préférable d'effectuer toutes vos recherches dans un « bac à sable virtuel » (ou machine virtuelle, VM). Plusieurs applications peuvent créer un pare-feu autour des programmes et applications que l'utilisateur choisit d'exécuter, dont les navigateurs Web comme Google Chrome et Firefox.

Une séance d'utilisation d'un navigateur Web exécutée à l'intérieur du bac à sable se ferme quand l'utilisateur ferme le bac à sable. Tous les fichiers téléchargés à partir du navigateur restent dans le bac à sable et peuvent être supprimés sans être enregistrés sur votre ordinateur quand le bac à sable est fermé. L'utilisateur conserve la possibilité d'autoriser le transfert de fichiers individuels hors du bac à sable.

Il existe diverses options de bac à sable pour les utilisateurs de PC ou de Mac. La plupart sont des applications gratuites, libres et relativement légères. Sandboxie (en anglais seulement) est une application populaire pour les utilisateurs de PC. Les utilisateurs de Mac peuvent se tourner vers l'application VirtualBox d'Oracle (en anglais seulement).

6. Logiciels antivirus

Pour mener des recherches à partir de sources ouvertes, il incombe à l'utilisateur de souscrire un abonnement à un logiciel antivirus de haute qualité. Toutefois, diverses options gratuites qu'il est conseillé de télécharger et d'exécuter régulièrement s'offrent à l'utilisateur qui ne s'est pas encore doté d'un logiciel antivirus.

- Malwarebytes propose notamment des analyses anti-logiciels malveillants gratuites, relativement légères et à la demande. Cet outil peut être utilisé conjointement avec d'autres logiciels antivirus.
- Bitdefender (en anglais seulement) est souvent cité comme un logiciel antivirus gratuit de grande qualité. Des essais gratuits sont également proposés par des fournisseurs d'abonnements payants comme Norton, McAfee, AVG et Kaspersky.

Conclusion

Si un utilisateur enfreint l'une des six règles essentielles énoncées dans le présent guide ou clique accidentellement sur un lien à téléchargement automatique, il convient de lancer une analyse rapide avec Malwarebytes. Il ne faut cependant pas oublier que la meilleure pratique lors de recherches à partir de sources ouvertes consiste à toujours supposer qu'il y a eu compromission. Si un État souhaite suivre vos activités de navigation et de recherche, il n'aura certainement aucune difficulté à le faire.

Les gouvernements et les médias du monde entier commencent à reconnaître la valeur des recherches à partir de sources ouvertes. Même dans les sociétés relativement fermées, il existe un *océan* de données inexploitées pouvant éclairer les décisions commerciales et politiques. Des études récentes ont mis en évidence l'utilité des documents budgétaires, des bons de commande, de l'imagerie géospatiale, des messages sur les médias sociaux, des documents gouvernementaux et des biographies des élites (sites en anglais seulement) pour comprendre les ambitions géopolitiques et les capacités militaires des États. Armé de ces conseils et astuces, où vous aventurerez-vous ?