Cycle et vocabulaire de l' Open Source Intelligence (OSINT)

Ugo Verdi

Cycle et vocabulaire de l'Open Source Intelligence (OSINT)

Le cas des services de renseignement

Résumé

Pour Williams et Blum, l'OSINT suit un cycle opérationnel de quatre étapes : la collecte (*collection*), le traitement (*processing*), l'exploitation et la production (Williams et Blum, 2018).

La collecte

La collecte renvoie directement à la collecte de données (*data collection*) le processus de rassemblement et de mesure d'informations sur des variables d'intérêt, d'une manière systématique établie qui permet de répondre aux questions de recherche énoncées, de tester des hypothèses et d'évaluer les résultats (ORI, 2005). Celle-ci peut être réalisée soit par l'extraction de données (*data mining*) ayant pour objectif d'extraire des informations utiles et cachées au sein de grands volumes de données (Han, Kamber, et al., 2012), soit par le *screen scraping*, une technique de capture d'écran permettant d'enregistrer les informations affichées sur un écran pour les utiliser à d'autres fins (cette technique est originaire du *data scraping)* (MagIt, 2020). Ou encore par le Web *Crawling*, un processus de recueil d'information sur le web qui consiste à mettre en place un robot appelé *crawler* qui parcourt tout ou partie du Web, copie les pages trouvées et les stocke dans une archive (Opendatalab, 2021).

L'ensemble de ces données hétérogènes sont alors transférées dans des entrepôts de données (*data warehouse*) au sein desquels elles sont assemblées (*data aggregation*). Elles sont ensuite complétées et interrogées pour effectuer diverses analyses (Gupta et Mathur, 2012).

Le traitement

L'OSINT a recours à plusieurs techniques pour le traitement de ses données. En premier lieu, concernant le fond, s'applique le nettoyage des données (appelé indifféremment « data cleansing », « data cleaning » ou « scrubbing ») qui consiste en la détection et la suppression d'erreurs ou incohérences au sein des données afin d'améliorer leur qualité (Rahm et Do, 2000). Concernant la forme, l'emploi du *data wrangling* ou *data munging*, à savoir un traitement des données qui restructure, nettoie, et enrichit les données « brutes » disponibles dans un format plus adapté (Sharma, 2021). Le *data wrangling* s'accompagne de six étapes : la découverte, la structuration, le nettoyage, l'enrichissement, la validation et la publication des données (Rattenbury, Hellerstein, et al., 2015). Une autre technique est la réduction de données (*data reduction)* qui réduit la quantité de données produites, traitées et transmises par des capteurs tout en maintenant un bon niveau de qualité des données rassemblées (Moussa, 2018).

L'exploitation

S'ensuit l'analyse des données (*data analysis*), à savoir l'ensemble de procédures d'analyse de données, de techniques pour interpréter les résultats de ces procédures, manières de planifier le rassemblement de données pour rendre leur analyse plus simple, plus précise ou correcte, et tous les mécanismes et résultats des statistiques (mathématiques) qui s'appliquent à l'analyse des données (Tukey, 1962). Parmi les méthodes de classifications s'inscrit la métadonnée, aussi appelée « data dictionnary », « record layout » et « data documentation » (Herzog, 2015), une donnée sur la donnée permettant de mieux décrire et retrouver cette dernière (Buckland, 2006).

Cette description détaillée de la donnée permettant aux utilisateurs de mieux comprendre et travailler avec des sets de données, (Herzog, 2015) ainsi que de former des structures d'organisation au moyen desquelles les documents peuvent être organisés (Buckland, 2017).

La production

La partie finale est la production où les données sont fournies au consommateur sous une forme utilisable (Williams et Blum, 2018). Il peut être fait usage de la *datavisualisation,* aussi appelée « dataviz », qui regroupe un ensemble de techniques de représentation graphique et d'exploration visuelle de données (Arruabarrena, 2017) rendant cohérent un set de données et permettant de comprendre facilement des informations très compliquées et conséquentes.

Le vocabulaire de l'OSINT regroupe un ensemble éclectique de concepts, de techniques et de mouvements politiques. Sa pleine reconnaissance, compréhension et application nécessite une collaboration entre les différents domaines d'expertises universitaires, les agences gouvernementales et les entreprises de renseignements impliquées dans cette démarche.

Bibliographie

1. Arruabarrena, B. (2017). L'expert en dataviz, un métier en transition. *I2D - Information, données & documents*, *54*(3), 7-8. https://doi.org/10.3917/i2d.173.0007
2. Buckland, M. K. (2006). Description and search : Metadata as infrastructure. *Brazilian Journal of Information Science*, 3-14. https://doi.org/10.36311/1981-1640.2006.v0n0.02.p3
3. Gupta, S., & Mathur, S. (2012). Classification of data warehouse testing approaches. *International Journal of Computers & Technology*, *3*(3), 381-386.
4. Han, J., Kamber, M., & Pei, J. (2012). *Data mining : Concepts and techniques* (3e éd.). Elsevier/Morgan Kaufmann.
5. Herzog, D. (2015). *Data literacy, a user's guide*. Sage Publishing.
6. *Lexique autour de l'Open Data*. (2021). opendatalab.fr. http://www.opendatalab.fr/l-opendata/lexique-autour-de-l-open-data
7. Moussa, M. A. (2018). *Data gathering and anomaly detection in wireless sensors networks* [Paris-Est]. https://pastel.archives-ouvertes.fr/tel-01936285/document
8. ORI. (2021). *Data collection*. ori.hhs.gov/. https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html
9. Rahm, E., & Do, H. H. (2000). Data cleaning : Problems and current approaches. *IEEE Data Engineering Bulletin*, 12.
10. Rattenbury, T., Hellerstein, J., Heer, J., Kandel, S., & Carreras, C. (2017). *Principles of Data Wrangling, practical techniques for data preparation*. O'Reilly Media.
11. *Screen Scraping*. (2020). lemagit.fr. https://www.lemagit.fr/definition/Screen-scraping
12. Sharma, A. (2021). *What is Data Wrangling ? Its Tools & 6 steps in Wrangling*. favtutor.com. https://favtutor.com/blogs/data-wrangling
13. Sparks, T. (2014). Why haven't technologies fixed open source intelligence? *JMU Scholarly Commons*, 146.

14. Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 67.

15. Williams, H., & Blum, I. (2018). *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise* (p. 63). RAND Corporation. https://www.rand.org/pubs/research_reports/RR1964.html