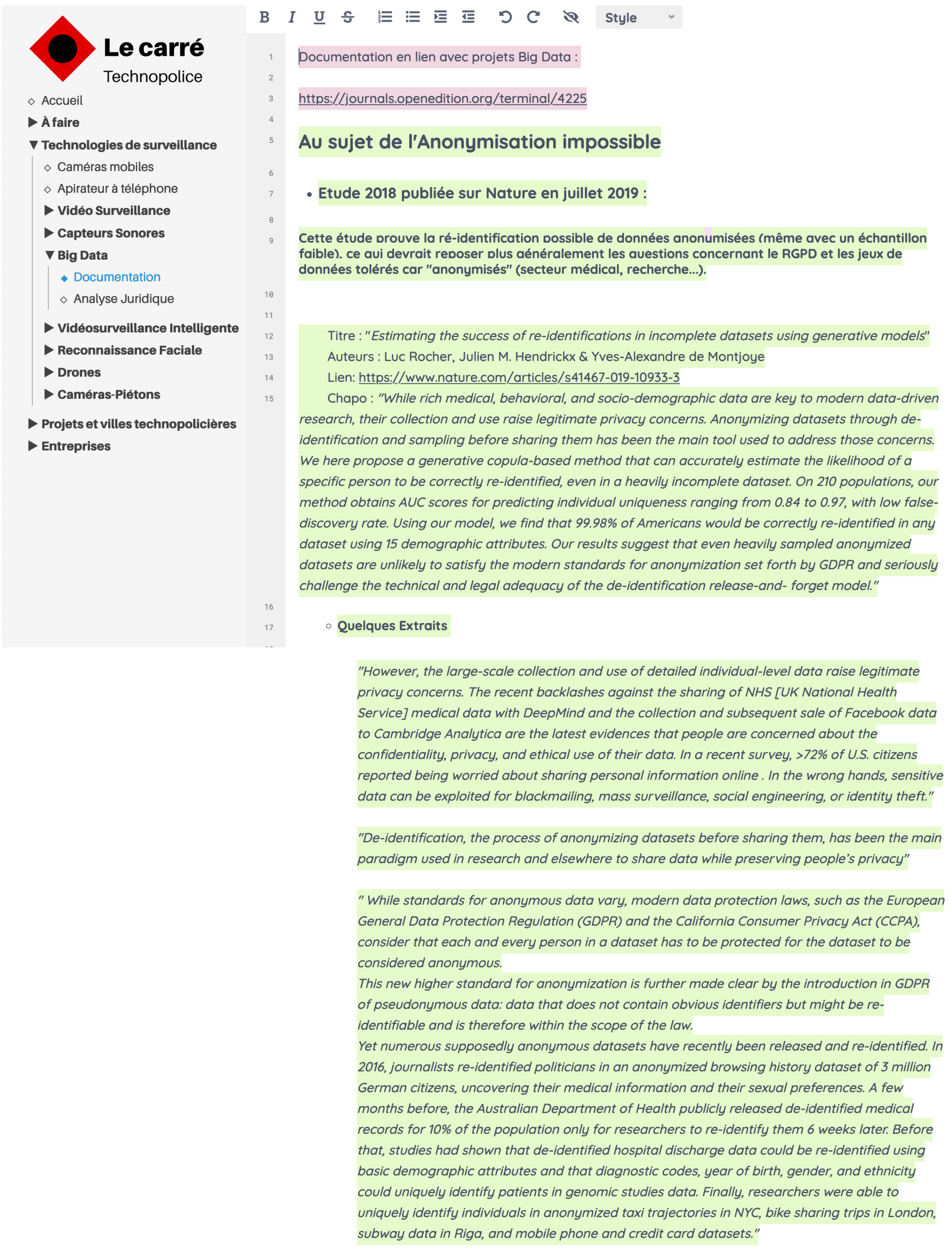

Documentation en lien avec projets Big Data :

https://journals.openedition.org/terminal/4225

# Au sujet de l'Anonymisation impossible

- **Etude 2018 publiée sur Nature en juillet 2019 :**

**Cette étude prouve la ré-identification possible de données anonumisées (même avec un échantillon faible), ce qui devrait reposer plus généralement les questions concernant le RGPD et les jeux de données tolérés car "anonymisés" (secteur médical, recherche…).**

Titre : "*Estimating the success of re-identifications in incomplete datasets using generative models*"
Auteurs : Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye
Lien: https://www.nature.com/articles/s41467-019-10933-3
Chapo : *"While rich medical, behavioral, and socio-demographic data are key to modern data-driven research, their collection and use raise legitimate privacy concerns. Anonymizing datasets through de-identification and sampling before sharing them has been the main tool used to address those concerns. We here propose a generative copula-based method that can accurately estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. On 210 populations, our method obtains AUC scores for predicting individual uniqueness ranging from 0.84 to 0.97, with low false-discovery rate. Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and- forget model."*

  - **Quelques Extraits**

> *"However, the large-scale collection and use of detailed individual-level data raise legitimate privacy concerns. The recent backlashes against the sharing of NHS [UK National Health Service] medical data with DeepMind and the collection and subsequent sale of Facebook data to Cambridge Analytica are the latest evidences that people are concerned about the confidentiality, privacy, and ethical use of their data. In a recent survey, >72% of U.S. citizens reported being worried about sharing personal information online . In the wrong hands, sensitive data can be exploited for blackmailing, mass surveillance, social engineering, or identity theft."*

> *"De-identification, the process of anonymizing datasets before sharing them, has been the main paradigm used in research and elsewhere to share data while preserving people's privacy"*

> *" While standards for anonymous data vary, modern data protection laws, such as the European General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), consider that each and every person in a dataset has to be protected for the dataset to be considered anonymous.*
> *This new higher standard for anonymization is further made clear by the introduction in GDPR of pseudonymous data: data that does not contain obvious identifiers but might be re-identifiable and is therefore within the scope of the law.*
> *Yet numerous supposedly anonymous datasets have recently been released and re-identified. In 2016, journalists re-identified politicians in an anonymized browsing history dataset of 3 million German citizens, uncovering their medical information and their sexual preferences. A few months before, the Australian Department of Health publicly released de-identified medical records for 10% of the population only for researchers to re-identify them 6 weeks later. Before that, studies had shown that de-identified hospital discharge data could be re-identified using basic demographic attributes and that diagnostic codes, year of birth, gender, and ethnicity could uniquely identify patients in genomic studies data. Finally, researchers were able to uniquely identify individuals in anonymized taxi trajectories in NYC, bike sharing trips in London, subway data in Riga, and mobile phone and credit card datasets."*