



**HAL**  
open science

# Ethique du numérique, données personnelles et données d'intérêt générale. Préparation à l'agrégation d'Informatique

Roberto Di Cosmo

## ► To cite this version:

Roberto Di Cosmo. Ethique du numérique, données personnelles et données d'intérêt générale. Préparation à l'agrégation d'Informatique. Master. France. 2023. <hal-03341559v2>

**HAL Id: hal-03341559**

**<https://hal.science/hal-03341559v2>**

Submitted on 17 Apr 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

# Ethique du numérique, données personnelles et données d'intérêt générale

Préparation à l'agrégation d'Informatique

Roberto Di Cosmo  
Inria et Université de Paris

17 Avril 2023



## Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des données pour la Science Ouverte
- 10 Conclusion



# Courte biographie: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- 30+ years of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- 20+ years of Free and Open Source Software
- 10+ years building and directing structures for the common good



1999 *DemoLinux* – first live GNU/Linux distro

2007 *Free Software Thematic Group*  
150 members 40 projects 200Me

2008 *Mancoosi project* [www.mancoosi.org](http://www.mancoosi.org)

2010 *IRILL* [www.irill.org](http://www.irill.org)

2015 *Software Heritage* at INRIA

2018 *National Committee for Open Science*, France

2021 *EOSC Task Force on Infrastructures for Software*,  
European Union

Standard disclaimer: IANAL (I am not a lawyer)

## Elargir vos perspectives...

L'Informatique est une science dont les applications ont un impact sur toute notre société.

## ... à travers le programme officiel ...

- Points clés du RGPD.
- Notions de Données d'Intérêt Général (DIG).
- Points clés de la loi n° 2016-1321 du 7 octobre 2016 pour une République numérique.
- Enjeux d'éthique du numérique et valeurs sous-jacentes.

## ... pour mieux vous préparer

Vous avez la responsabilité de former des nouvelles générations plus conscientes des enjeux du numérique.

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte
- 10 Conclusion



## Les bases de la révolution informatique:

- processeurs,
- systèmes de stockage des données,
- réseaux,
- algorithmes et logiciels,
- ...

Cela évolue à une vitesse sans précédent.

# Loi de Moore (in memoriam, 24/3/2023, voir entrée sur Wikipedia)

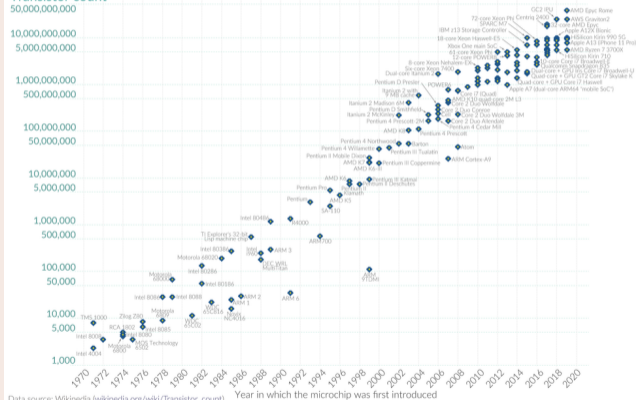
"Le nombre de transistors double tous les deux ans" Gordon Moore, 1965

## Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing - such as processing speed or the price of computers.

Our World  
in Data

### Transistor count



Data source: Wikipedia ([wikipedia.org/wiki/Transistor\\_count](https://www.wikipedia.org/wiki/Transistor_count))  
OurWorldInData.org - Research and data to make progress against the world's largest problems. Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

# Capacité de stockage, vitesse du réseau

## Taille des disques

Année	2.5in	3.5in
1997	3Gb	10Gb
2007	250Gb	1000Gb
2023	8000Gb	22000Gb

Et on peut avoir des SSD!

## Débit du réseau

Année	Rx	Tx
1998	256 kb/s	128 kb/s
2007	28000 kb/s	1000 kb/s (*)
2021	1000000 kb/s	100000 kb/s

(\*) Exception importante!

Trois à quatre ordres de grandeur en 25 ans...

... et sur toutes les dimensions!

Une évolution sans précédents, avec un grand impact sur la société

Cela pose des défis éthiques majeurs auxquels on était peu préparés

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte
- 10 Conclusion



## Utilisateurs d'Internet

de 2.6M en 1990 à 3+ milliards en 2016 (voir <https://ourworldindata.org/internet>)

## Abonnements mobiles

de 34M en 1993 à 8 milliards en 2019 (voir <https://www.statista.com/statistics/262950/global-mobile-subscriptions-since-1993/>)

- et depuis 2007 (iPhone), abonnement mobile == accès internet

## Téléphones mobiles

- ordinateurs sans fil, mobiles, connectés
- réseau massif de senseurs
- financé par les utilisateurs

## Cloud

- compute/storage/network à la demande
- data centers massifs
- opérés par des corporations multinationales

## Expansion massive du logiciel libre

- [SourceForge](#) (1999), [BitBucket](#) (2008), [GitHub](#) (2008), [GitLab](#) (2014), etc. etc.

## Réseaux peer to peer

- [Napster](#) (1999), [BitTorrent](#) (2001), etc.

## Développement collaboratif des connaissances

- [Wikipedia](#) (depuis 2001)
- [OpenStreetMap](#) (depuis 2004)

## Cadre législatif du logiciel

- "oeuvre de l'esprit", couverte par le droit d'auteur ([article L112-2 du CPI](#))
  - droit moral
  - droit patrimonial
- la (re)utilisation se fait sur la base d'un contrat, qui peut prendre la forme d'une *licence logicielle*
- des particularités mériteraient une étude approfondie (e.g.: voir [cette analyse d'un cas concret](#))

## Le logiciel libre ...

- logiciel dont la licence autorise sans restrictions
  - *l'utilisation*
  - la *modification* (implique l'accès aux sources et aux outils)
  - la *redistribution*
  - la *redistribution* des versions modifiées
- ces droits peuvent s'accompagner d'obligations (de citation, de préservation de la licence, etc.)
- cela mériterait une étude approfondie, voir par exemple
  - Canevet et Pellegrini, *Droit des Logiciels*
  - Di Cosmo, *Notes du cours sur les logiciels libres*

## ... est un des moteurs de la révolution numérique

- au cœur de pratiquement tous les systèmes logiciels modernes
- réduction massive du coût de développement de nouveaux systèmes informatiques

### L'échange *prime* sur la possession

- moteurs de recherche
- sites web
- information en ligne
- social networks

### Test

Combien de jours pouvez-vous résister/exister sans réseau?

## Evolution d'Internet

de épiphénomène technologique de la guerre froide (DARPA, TCP/IP, 1970): assurer la communication fiable même en cas de guerre nucléaire ... ..à moteur du changement des modes de communication, avec des problématiques multiples

## Supranationalité (connexion entre machines dans différents pays)

- quel droit applicable?
  - MegaUpload, PirateBay
  - GAFAM vs éditeurs
  - ...
- qui pilote les décisions techniques?
  - ICANN et ISOC
  - ...

## Neutralité des réseaux

### La neutralité du net en une image

Les fournisseurs d'accès (Orange, Free, Vodafone, ...)  
doivent me garantir un accès à Internet



**Internet est un droit.**  
Seule la justice peut décider  
d'une privation de droit.

© Sébastien Desbarrats, Internet@moa.fr

C'était le cas à l'origine de l'Internet, mais la technique n'impose rien.

Il s'agit d' un choix *sociétal*, et peut varier selon les pays et dans le temps

- grand firewall de Chine
- debat reglementaire aux US, et ses évolutions
- zero rating

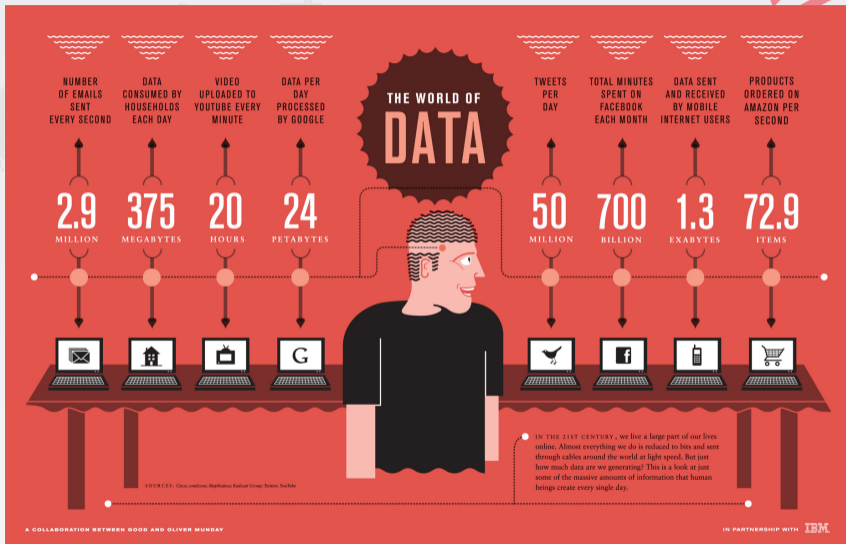
par l'état (dématérialisation des données publiques), par exemple:

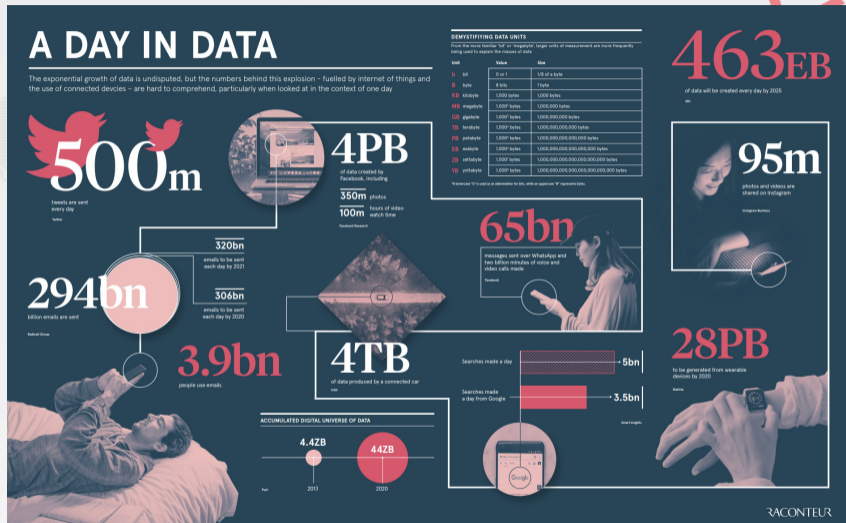
- dossier médical partagé
- dossier fiscal
- casier judiciaire
- cadastre

par des acteurs privés, par exemple:

- préférences de consommation (cartes de fidélité)
- préférences culturelles (cinéma, VOD, moteur de recherche)
- préférences politiques (réseaux sociaux, moteur de recherche)
- réseau de connaissances (téléphone, réseaux sociaux, courriel, banques)
- textes, photos, vidéos (téléphone, réseaux sociaux, moteur de recherche)

avec notre collaboration active





Credits: VisualCapitalist 2019

## ... qui a été longtemps peu ou pas encadrée: quelques exemples

### Cession de droits à Facebook (par simple click), version 2013

For content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to *use* any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account *unless* your content has been shared with others, and they have not deleted it.

### Clearview

Reconnaissance faciale basée sur plusieurs milliards d'images obtenues sur le Web et les réseaux sociaux

### WhatsApp

Collecte de (meta)données personnelles (voir [un résumé du débat](#))

## Les ingrédients

- dématérialisation
- dans des infrastructures *centralisées*
- mises en réseau
- avec des moyens de calcul énormes

## Les résultats

il y en a des positifs et des négatifs

Avec données massives, calcul et avancées algorithmiques on peut avoir

des outils qui font rêver...

- traduction automatique (DeepL)
- détection automatique de cancers
- génération automatique de documents, textes (LLM, ChatGPT, Bard, LLaMa etc.)
- assistants personnels
- ...

... ou faire peur

- SciGen: **faux papiers de recherche** (déjà en 2005!)
- générateurs de *fake news* pour les *troll farms*
- *décisions automatiques biaisées* sur la santé, l'emploi ou la liberté des personnes (voir **quelques exemples**)
- ...

## Avec des données massives, on peut

- suivre/prévoir l'avancée d'une grippe ou une épidémie e.g. [Google Flu Trends](#) (2008-2015)
- fournir des prévisions de trafic (Google Maps, Waze, etc.) ou de fréquentation de sites (possiblement utiles en temps de COVID, voir [un autre exemple chez Google](#))

## Mais aussi

- [savoir avant vous si vous attendez un enfant](#):  
*"I had a talk with my daughter. It turns out there's been some activities in my house I haven't been completely aware of"* Un client de Target, en 2012.
- cibler précisément des utilisateurs pour les manipuler (e.g. [Cambridge Analytica](#))

## Traçage des échanges

- NSA (Snowden)
  - recording of conversations, conversion to searchable text (AI)
  - metadata collection (who talks to whom)

## Technologies plus ou moins invasives

- traces GSM
- Deep packet inspection

nous ne sommes pas innocents dans tout cela

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques**
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte
- 10 Conclusion



Ces exemples montrent le besoin d'une éthique de l'Informatique:

- ce qui est techniquement possible n'est pas forcément à faire
- un logiciel qui prend des décisions peut impliquer des questions éthiques

## Principes

Des principes de bases ont été développés par différentes communautés, et notamment:

- code éthique de l'ACM: <https://www.acm.org/code-of-ethics>
- principes d'Asilomar : <https://futureoflife.org/ai-principles>

## Comités

En France, un **comité pilote d'éthique du numérique (CNPEN)** a été créé au sein du Comité National d'Éthique.

Voir aussi: la note sur **logiciels libres et décisions algorithmiques** (LLW Barcelona, 2018)

## Un exemple récent: recherche en sécurité

- soumission de patches "hypocrites" au noyau Linux au IEEE Symposium on Security and Privacy 2021
- réaction des développeurs Linux Avril 2021
  - "you are not welcome here"
- le rapport de comité de programme de S&P Mai 2021
  - violation de deux principes clé
    - Informed Consent / Autonomy
    - Beneficence / Balancing risks and benefits
  - mise en place d'un comité d'éthique

## La question du TDM

- DIRECTIVE (EU) 2019/790, 17 April 2019 on copyright and related rights
- Article 3: Text and data mining for the purposes of scientific research
- Article 4: Exception or limitation for text and data mining

## Essor des LLM

- GitHub CoPilot (OpenAI Codex)
- training dataset = ?
- class action aux Etats Unis

GitHub Copilot litigation [case updates](#) [get updates by email](#) [contact legal team](#)

We've filed a lawsuit challenging GitHub Copilot, an AI product that relies on unprecedented open-source software piracy.

**Because AI needs to be fair & ethical for everyone.**

NOVEMBER 3, 2022

Hello. This is Matthew Butterick. On October 17 I told you that I had teamed up with the amazingly excellent class-action litigators Joseph Saveri, Cadjo Zirpoli, and Travis Manfredi at the Joseph Saveri Law Firm to investigate GitHub Copilot.

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux**
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des données pour la Science Ouverte
- 10 Conclusion



Suite aux dérives multiples et aux vides juridiques constatés, le législateur est intervenu. On s'intéressera ici en particulier aux enjeux suivants:

## Données personnelles

- données liées aux personnes physiques
- encadrées par le [RGPD \(régulation européenne de 2016\)](#) et la [Loi Lemaire \(2016\)](#)

## Données administratives et publiques

- encadrées par [la Loi Lemaire](#) (voir aussi le [rapport Trojette, 2013](#))

## Données d'intérêt générale (Loi Lemaire, rapport Bothorel)

- données produites par des entités publiques ou par délégation des pouvoirs publics (Loi Lemaire)
- données produites ou détenues par des acteurs privés, mais dont l'usage peut être nécessaire pour des fins d'intérêt général ( [rapport Bothorel](#) 23 décembre 2020)

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles**
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte
- 10 Conclusion



# Le règlement général sur la protection des données (RGPD)

## Objectifs du Règlement (UE) 2016/679, en vigueur depuis le 25 Mai 2018

- "protection des personnes physiques à l'égard du traitement des données à caractère personnel"
- "protection des données à caractère personnel"
- "règles relatives à la libre circulation de ces données"

## Structure

- 11 chapitres, 99 articles
- définit les "données à caractère personnel"
- fixe les droits des personnes, et les obligations de qui traite leur données
- fixe le cadre des autorités de contrôle

Nous allons explorer les notions clés dans une présentation préparée par Anne Combe, la Data Protection Officer d'Inria.

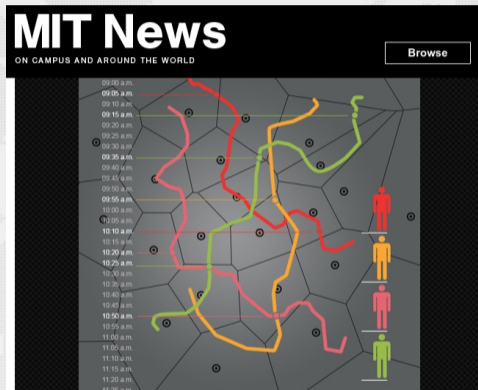
## La Commission Nationale Informatique et Liberté (CNIL)

- créé par la **Loi n° 78-17 du 6 janvier 1978** (Loi Informatique et Liberté)
- avis systématique sur les nouvelles lois (Art. 8, 4.a)
- contrôle du respect du cadre legal des traitements touchant à des données à caractère personnel
  - automatisés
  - non automatisés mais amenant à constituer un fichier
- pouvoir de sanction (Art. 20 à 23)

## à noter

- **il n'y a plus de déclaration préalable de fichiers à la CNIL**

Est-ce facile d'anonymiser *vraiment* les données personnelles?



*“you are identified by just 4 points, over a year”*

*Scientific Reports, 2013*



- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale**
- 8 Données d'Intérêt Générale
- 9 Le cas des données pour la Science Ouverte
- 10 Conclusion



## Répères

- loi n° 2016-1321 du 7 octobre 2016 pour une République numérique
- aussi connue comme "Loi Lemaire", de la Ministre Axelle Lemaire qui l'a portée
- voir le résumé dans le dossier de presse officiel

## Structure

- Titre I: circulation des données et des savoirs (39 articles)
- Titre II: protection des droits dans la société numérique (29 articles)
- Titre III: accès au numérique (41 articles)

## Transparence des algorithmes

Forts enjeux sociétaux (voir [APB](#), [Parcoursup](#), IA, etc.)

- De l'administration ([Article 4](#))
- Des opérateurs économiques (Titre II)

## Données Personnelles

- Protection des données personnelles (Titre II)
- Droit à la récupération et à la portabilité de données (Titre II)

## Internet (Titres II et III)

- neutralité
- très haut débit
- maintien de la connexion

## Données publiques et d'intérêt général: ouverture par défaut (Titre I)

- documents administratifs (y compris les codes sources! [Article 2](#))
- données d'intérêt générale (activités liées aux pouvoirs publiques)
  - cadastre
  - jugements
  - informations issues de la gestion d'un service public délégué (électricité, gaz, ...)
- information scientifique: promotion de l'accès ouvert ([Article 30](#))
- MOOCs

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale**
- 9 Le cas des données pour la Science Ouverte
- 10 Conclusion

En Europe: **Open Data Directive** 2019 (art. 14): "high value dataset" qui doivent être

- (a) *available free of charge, subject to paragraphs 3, 4 and 5;*
- (b) *machine readable;*
- (c) *provided via APIs; and*
- (d) *provided as a bulk download, where relevant.*

il s'agit essentiellement de *données produites par des entités publiques*

En France, **Rapport Bothorel, Décembre 2020** Section 5, données d'Intérêt Générale

- Données d'Intérêt Générale comme "donnée d'un acteur privé" dont la mise à disposition peut se justifier par un "motif d'intérêt général"
- Deux cas identifiés
  - Les données provenant du secteur privé mises à disposition des acteurs publics (B2G).
  - L'échange de données entre acteurs privés (B2B).

Pas de cadre normatif de référence pour le moment

## Agences fédérales

Une oeuvre produite par (des employés de) une agence fédérale n'est pas couverte par le droit d'auteur (section 105 du Copyright Act), et est donc dans le domaine public.

Grâce à cela on a accès à:

- le code source de l'Apollo 11
- le CIA world factbook
- les données de l'USGS

## Municipalités

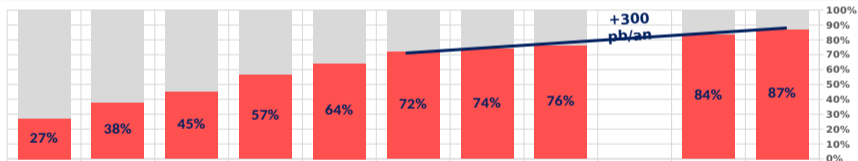
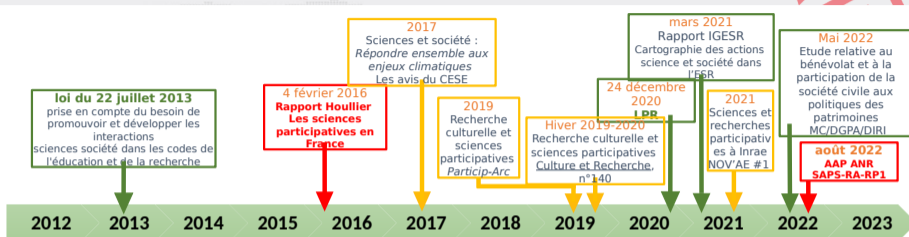
Ouverture des données comme stratégie pour un meilleur service

- Trimet (~ RATP de Portland, OR, USA) fournit des APIs ouvertes aux développeurs depuis plus de dix ans, ce qui a conduit à des dizaines d'applications tierces spécialisées

- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte**
- 10 Conclusion



# Un aperçu rapide



## Taux d'équipement - smartphone

ARCEP - Enquête sur la diffusion des technologies de l'information et de la communication dans la société française en 2022

<https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/barometre-du-numerique/le-barometre-du-numerique.html>

K. MAUSSANG – Livrable Données et recherches participatives – SPSO 13 avril 2023 4



- 1 Introduction
- 2 Accélération technologique
- 3 Impact sur la société
- 4 Enjeux Éthiques
- 5 Enjeux Légaux
- 6 Données personnelles
- 7 Loi Lemaire: données publiques et d'intérêt générale
- 8 Données d'Intérêt Générale
- 9 Le cas des donnés pour la Science Ouverte
- 10 Conclusion

## Perte de l'innocence

l'Informatique *n'est plus* une science exempte d'impact sociétal:

- logiciel et algorithmes font partie du tissu constitutif de nos sociétés
- évolution massive des capacités de calcul, échange et stockage
- dématérialisation des données (en particulier personnelles)
- explosion du domaine d'application des décisions automatiques

Reflection sur l'éthique: ce qui est techniquement possible n'est pas forcément à faire

*"move fast, break things"* n'est plus un credo acceptable, **meme pour les VC**

Un cadre législatif qui se structure (et évolue!)

droit du logiciel (non traité dans ce cours); traitement des données personnelles; données publiques et d'intérêt générale; traitements algorithmiques

## Textes de loi et règlements

- Loi n° 78-17 du 6 janvier 1978 (Loi "informatique et liberté", actualisée en 2021)
- RGPD (règlement général sur la protection des données), 2016
- Loi pour une république numérique (Loi Lemaire), 2016
- Open Data Directive, 2019

## Rapports

- rapport Trojette, 2013
- rapport Bothorel, 2020

## Codes d'éthique

- code éthique de l'ACM
- principes d'Asilomar pour l'IA

## Organismes

- Commission nationale informatique et liberté (CNIL)
- Comité national d'éthique