HAL open science

Les impacts environnementaux et sociétaux des données : un défi pour l'avenir

Didier Mallarino, Sylvie Le Bras, Cyrille Bonamy

To cite this version:

Didier Mallarino, Sylvie Le Bras, Cyrille Bonamy. Les impacts environnementaux et sociétaux des données : un défi pour l'avenir. Congrès JRES : Les Journées Réseaux de l'Enseignement et de la Recherche, RENATER, May 2022, Marseille, France. hal-03702208

HAL Id: hal-03702208

https://hal.science/hal-03702208

Submitted on 22 Jun 2022

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.

Les impacts environnementaux et sociétaux des données : un défi pour l'avenir

Didier Mallarino
OSU Pythéas / Laboratoire MIO
Université de Toulon, Campus de la Garde

Sylvie le Bras
Aix-Marseille université, CEA, CNRS, BIAM, Saint Paul lez durance, France 13108

Cyrille Bonamy
LEGI – UMR5519, Domaine Universitaire, 38400 Saint-Martin-d'Hères

Résumé

Le numérique nous offre l'opportunité d'approfondir notre compréhension du monde. Cependant, avec une soixantaine de zettaoctets de données créées en 2020, le numérique doit lui aussi ***impérativement s'interroger sur ses usages et participer*** *aux efforts de réduction des impacts environnementaux auxquels il participe.*

Que ce soit avec le streaming vidéo, le cloud gaming*, l'internet des objets (IoT), les réseaux sociaux, l'IA ou les* bitcoins*, la croissance des usages fait croître les volumes de données. Si on ne peut bien évidemment pas regarder et résumer les impacts du numérique à l'aide de ce seul indicateur, il reste largement pertinent de s'interroger sur cette croissance exponentielle et sur les besoins matériels nécessaires à la satisfaire.*

Dans le monde académique et la recherche, la production et la consommation des données s'envolent également (calculs complexes, modélisation, imagerie spatiale, études environnementales et engouement pour le deep learning*). Cette accumulation de données se répercute sur les besoins logiciels, matériels et énergétiques qui peinent, malgré les gains d'efficacité, à compenser l'accroissement des usages. Le paradoxe de Jevons (effet rebond) reste totalement d'actualité.*

Ces usages participent et amplifient les dérèglements environnementaux (changement climatique, baisse de la biodiversité, épuisement des ressources naturelles, pollutions) tout en soulevant des questions d'ordre éthique et moral sur qui bénéficie et qui subit.

*Les actions permettant de limiter nos impacts existent mais les réponses sont complexes. Elles imposent une réflexion approfondie sur nos usages et la prise en compte des aspects sociaux, sociétaux, humains et éco responsables pour être pertinentes. Nous aborderons ainsi l'axe de la sobriété et la notion d'*open data*, un des piliers de la science ouverte qui favorise la mutualisation et le partage des données. En effet, la notion de données FAIR qui propose les principes fondamentaux permettant de rendre les données faciles à trouver, accessibles, interopérables et réutilisables, même si il peut conduire à un impact numérique accru, reste fondamental afin de respecter le coût environnemental de l'acquisition bien supérieur au stockage de la donnée, évitant ainsi des acquisitions à de multiples reprises.*

Mots-clefs

EcoInfo, Open Science, Open Data, *sobriété numérique, environnement, changement climatique*, FAIR

1 Introduction

Après un état des lieux des conséquences de l'action anthropique sur l'environnement puis la part du numérique, nous aborderons les impacts de ce dernier sous l'angle un peu inhabituel de la donnée. Ce léger décalage de regard permet à la fois d'interroger nos usages de l'outil numérique et d'envisager comment, à travers eux, nous pouvons participer à l'indispensable diminution des impacts environnementaux. En effet, les approches classiques par le matériel ou le logiciel négligent en partie la raison de l'existence de ces outils : la manipulation de la donnée pour en extraire de l'information et de la connaissance.

Car c'est bien notre soif sans fin pour la donnée qui se répercute sur les besoins en matériel (réseau de transport planétaire, infrastructure serveurs, course à la puissance de traitement) et qui a de nombreuses conséquences tant environnementales que sociétales. C'est enfin l'approche propriétaire des logiciels de traitement et des données elles mêmes qui impose souvent des incompatibilités et une utilisation parfois abusive qui entraîne des impacts sociétaux catastrophiques (dataveillance – utilisation de la donnée pour surveiller les individus –, datapulation – utilisation de la donnée pour manipuler les individus – allant également jusqu'à la mise en danger des démocraties).

Au sein de l'Enseignement Supérieur et Recherche (ESR), ces données, en croissance exponentielle comme partout, sont d'une incroyable variété. Les modèles numériques, les simulateurs, l'intelligence artificielle, les campagnes de mesures sont des exemples évidents de cette gourmandise en données qui alimentent aussi le « *big data* ».

Afin de réduire les impacts négatifs, des pistes existent, allant d'une forme de sobriété indispensable, eu égard aux enjeux environnementaux, à l'*Open Science* et au *FAIR*.

2 Les enjeux environnementaux et sociétaux

2.1 Changement climatique [1]

Depuis plus de 30 ans, le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC [2]) évalue et synthétise les connaissances scientifiques sur le climat. Le dernier rapport du GIEC, publié en 2021 [3], est vraiment catastrophique et souligne plusieurs éléments importants :

– les causes du réchauffement sont humaines : ce sont nos activités et leur croissance qui impactent le climat planétaire de manière rapide et généralisée par la production des Gaz à Effet de Serre (GES) ;

– il y a urgence à réduire nos émissions de GES très rapidement et fortement afin de limiter les dégâts environnementaux et conserver un environnement viable pour notre espèce ;

– la décennie 2010-2020 a été la plus chaude depuis 125000 ans. Le seuil des 1,5 degrés sera sans aucun doute atteint bien avant 2050 et l'ampleur de ce changement n'a pas été observé depuis plusieurs milliers d'années ;

– la violence et la fréquence des événements extrêmes sont corrélés avec la température globale de la planète. Plusieurs points de bascule catastrophiques pourraient entraîner l'humanité dans un monde partiellement invivable pour notre espèce (modifications des grands courants océaniques,

fonte du pergélisol ou des glaces aux pôles, montée des océans). D'autres impacts plus « discrets » tels que la hausse régulière des températures, les modifications des régimes des précipitations, la baisse du nombre de jours de gel, la fonte des glaciers induisent des feux de forêts, des tensions sur les ressources en eau, accélèrent l'érosion de la biodiversité en modifiant les milieux naturels et en perturbant les organismes vivants. La vitesse exceptionnelle de ces changements n'autorise pas une adaptation graduelle ;

– la société et l'économie sont largement affectées : exposition accrue aux aléas climatiques et aux risques sanitaires, pertes « économiques » et agricoles entraînant des crises géopolitiques (migrations, conflits, etc.) engendrées par des tensions liées à l'accès aux ressources fondamentales (alimentation, eau). La prise en compte par les militaires des menaces engendrées par les dégradations environnementales dans leurs études prospectives en dit long sur la situation [5].

La contribution du numérique aux émissions de GES est estimée aux alentours de 4 % de la totalité des émissions [6] (ce qui est du même ordre que l'impact de l'aviation civile). Ce qui inquiète dans le cas du numérique, c'est sa croissance vertigineuse (de l'ordre de +6 % par an) alors même que seule une partie (environ la moitié [7]) de l'humanité l'utilise. Cette croissance est en grande partie portée par l'explosion des usages du numérique et la croissance des données consommées. Par ailleurs, les effets rebonds liés au numérique sont nombreux : obsolescence matérielle et logicielle, multiplication des objets et services numériques plus ou moins incontournables, nouveaux usages, accélération économique, augmentation du volume des données associées, etc. Malheureusement, cette croissance n'est pas compensée par les progrès techniques d'efficacité énergétique qui, bien que destinés à aider à une plus grande sobriété, participent au final à une augmentation des usages [8].

2.2 Biodiversité

Le deuxième risque profond pour notre planète est le changement qui touche à la biodiversité. La surexploitation des ressources végétales, animales et minérales prive le monde vivant de l'espace et de la diversité nécessaires au maintien d'un monde viable et équilibré.

La Plateforme intergouvernementale scientifique et politique sur la biodiversité et les services écosystémiques (IPBES [9]), sorte de « GIEC de la biodiversité », menée par plus de 550 experts, a publié plusieurs rapports préoccupants : nous sommes confrontés à « **une extinction massive d'espèces »**, une première depuis la disparition des dinosaures, et ce déclin menace le bien-être de l'humanité. Les insectes [10], dont la taille est inversement proportionnelle à leur utilité pour le bon état de la biosphère, sont en danger : le taux de disparition des colonies d'abeilles chaque hiver est de l'ordre de 30 % [12] et certaines études scientifiques font état de disparition de 75 % sur les 30 dernières années des insectes volants dans certaines zones [10] tandis que les oiseaux, les poissons et les mammifères disparaissent également de manière alarmante soit par surexploitation de la ressource, soit par manque d'habitat et de nourriture, mais aussi simplement par la pression de sélection que nous appliquons pour ne conserver que les espèces que nous jugeons utile, domestiquées pour notre alimentation ou d'autres usages. Enfin, les ressources en eau par personne ont baissé de en moyenne de 30 % entre 1970 et 1990 alors même que la consommation globale à presque doublé les 50 dernières années [11] tandis que les terres arables sont artificialisées ou

désertifiées. Les scientifiques s'accordent à parler d'une sixième extinction d'origine anthropique et extrêmement rapide. Cette érosion du vivant, des terres arables et des ressources hydriques vitales, entraîne également des tensions géopolitiques (flux migratoires, conflits armés).

Bien que les liens avec le numérique puissent sembler ténus, celui-ci est loin d'être neutre dans ces problématiques. Les outils numériques et la croissance continuelle des données de grande qualité et accessibles (imagerie satellite, prévisions météo, géolocalisation, sonars) facilitent l'exploitation intensive des océans, tandis que les exploitations minières[1] et les déchets électroniques polluent pour de nombreuses années d'immenses zones aquatiques et terrestres. Les outils numériques et leur interconnexion mondiale servent également au système de production et de consommation de masse. Sans le numérique, toutes ces surexploitations seraient probablement moins faciles, moins rapides, et probablement moins destructrices.

2.3 Les indicateurs environnementaux

La situation environnementale peut être analysée au travers de plusieurs types d'indicateurs environnementaux qui permettent soit une vision globale (jour du dépassement), soit une vision sectorielle (limites planétaires). Les deux sont malheureusement très mauvais. Malgré les limites inhérentes à ce type d'indicateurs et la probable sous-estimation des dégâts, ces visions synthétiques restent fortement inquiétantes et on pourra trouver en annexe les deux principaux indicateurs auxquels nous faisons référence ici.

2.4 La part du numérique

Le numérique ne fait malheureusement pas exclusivement partie de la solution. Il participe même assez largement à la dégradation de la situation. La croissance des outils numériques agit sur plusieurs de ces secteurs en accélérant leurs propres potentialités, tandis que le numérique lui-même consomme les ressources de la terre (minerais métalliques, sable, pétrochimie, etc.), l'eau et l'énergie pour la fabrication des puces électroniques, dans une course à la miniaturisation nécessitant paradoxalement des ressources de plus en plus importantes pour fabriquer de plus en plus petit. Le numérique est également une source importante de déchets : plus de 50 millions de tonnes de DEEE (Déchet d'Équipement Électrique et Électronique) par an, dont une grande partie enfouie.

Le lien indiscutable avec notre consommation de données, utiles ou non, conduit le numérique sur une pente de croissance insoutenable environnementalement parlant.

2.5 Les enjeux sociétaux fortement intriqués avec le numérique

Les outils numériques sont positifs à bien des égards. Ils sont un puissant moyen de connecter les personnes, d'améliorer la recherches dans tous les secteurs (médical, sciences fondamentales, etc.), et qui offre de réelles nouvelles possibilités telles que le télétravail dont l'ADEME a conclu dans une étude récente [20] que les externalités positives prédominaient malgré de nombreux effets rebond. Cependant, si le numérique peut parfois aider à améliorer l'efficacité énergétique, les effets rebond peuvent parfois annihiler les gains soit par un accroissement de l'usage, soit par une

1 selon l'organisme *Green Cross*, l'industrie minière serait la 2e industrie la plus polluante du monde, derrière le recyclage des batteries au plomb et devant les teintureries, les décharges industrielles et les tanneries [13])

redirection des économies engendrées sur d'autres secteurs polluants. Il ne sera donc pas question de « faire l'autruche » et de ne pas regarder en face les aspects plus sombres de cette « révolution ». De nombreux aspects sont en effet concernés par cette utilisation de données massives.

Le premier d'entre eux, c'est la révolution de l'intelligence artificielle. En effet, le *deep learning* utilise des volumes de données croissants et introduit de nombreuses interrogations, notamment vis à vis d'usages insensés ou a minima discutables (*fake news*, profilage, etc...).

Les performances associées à ces pratiques se font cependant au prix d'une consommation d'énergie importante afin de manipuler et extraire les gros volumes de données nécessaires [25]. Ces algorithmes entraînent également une forme de pensée unique, très souvent biaisée [26] soit par des bases de données malgré tout limitées, soit par une volonté spécifique liée à un usage ciblé : commercial (réseaux sociaux), manipulatoire (site de commerce) ou dictatorial (processus de notation sociale en Chine [27],[28], procureur numérique pour rendre la justice [29]). Ces usages de datapulation et de dataveillance explosent à travers le monde, rejoignant les pires scénarios de la science-fiction, dont le roman de Georges Orwell « 1984 » avait déjà exploré l'idée. Même les barrières du RGPD en Europe et de la CNIL en France semblent bien fragiles face à la tentation de surveiller et manipuler « pour le bien » des sociétés ou des personnes. La crise sanitaire et la polémique autour de l'usage des « pass sanitaires » en sont un tout petit exemple.

Parmi les effets sociétaux dévastateurs liés à une surconsommation du numérique et des écrans, on n'oubliera pas les effets d'addictions, d'anxiété, d'hyperactivité ou de retards mentaux avérés sur les enfants en bas âge [30].

Enfin, la marchandisation croissante et la grande valeur attribuée à certains types de données (par exemple, les données personnelles) sont associées à une augmentation de l'inégalité de pouvoir et de visibilité entre les différentes nations, les classes de population et les communautés scientifiques. La fracture numérique entre ceux qui peuvent non seulement accéder aux données, mais aussi les utiliser, se creuse : elle devient une « fracture des données ». La privatisation des données entraîne un déséquilibre entre les données accessibles et celles plus complexes et plus valorisables conservées par les acteurs privés, qui ne diffusent que les données de « moindre importance » ou celles pour lesquelles la participation massive d'autres acteurs peut les aider à apporter du sens et de la connaissance, ou plutôt une valeur commerciale [31]. Dans l'exemple du numérique, cette rétention d'information d'accès aux données environnementales de la fabrication des matériels empêche une évaluation juste et précise de leur impact environnemental.

3 Quelques définitions

3.1 Les données

De quoi parle-t-on quand on parle de données ? Ou d'information ? Ou même de connaissances ? Des mesures de température ou de pression relevées chaque jour sur un site physique sont des données, tandis que la courbe donnant l'évolution dans le temps de ces données en un lieu est une information. Enfin, l'analyse et l'interprétation de ces informations (pour en déduire par exemple que la température sur Terre augmente en fonction de l'activité humaine) est une connaissance. Ces

trois notions sont très proches les unes des autres et peuvent interagir entre elles, positivement comme négativement (pour induire une interprétation souhaitable par exemple).

• Une **donnée** est une description élémentaire, typiquement numérique ici, d'une réalité physique ou parfois mentale. C'est par exemple une observation, une mesure ou le résultat d'un calcul ;

• À partir de données collectées, de **l'information** est obtenue en organisant ces données, en les structurant pour en dégager du sens ;

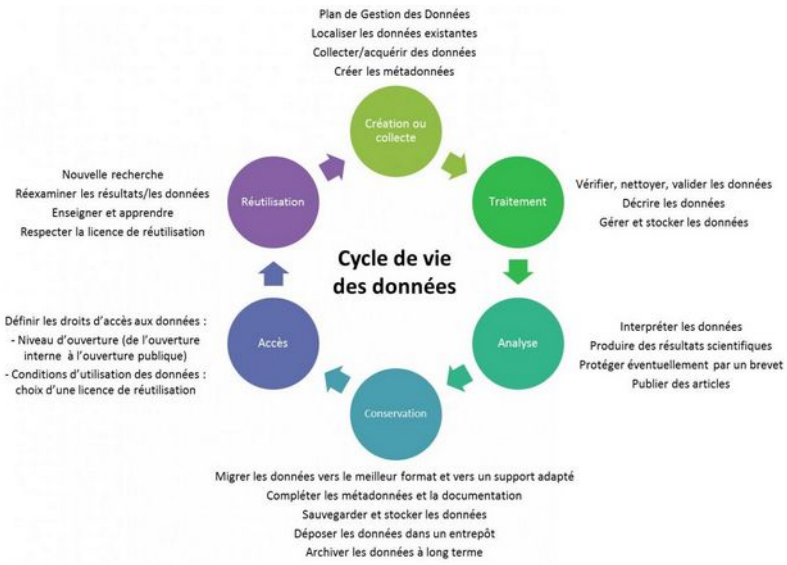
• En interprétant le sens de l'information, nous aboutissons à des **connaissances**, c'est-à-dire à des « faits » considérés comme vrais dans l'univers d'un locuteur, et à des « lois » (des règles logiques) de cet univers. Cette connaissance demande une concordance des analyses et des interprétations par la communauté scientifique notamment.

Enfin, la **métadonnée** est une donnée servant à définir ou décrire une autre donnée quel que soit son support.

3.2 Le cycle de vie de la donnée

La donnée une fois disponible est indéfectiblement liée à un cycle de vie [Figure 1] dont les étapes peuvent légèrement varier en fonction des modèles retenus.

Figure 1: Exemple de cycle de vie



On retrouvera cependant les étapes fondamentales de l'acquisition, de l'analyse, du stockage, du traitement, de l'archivage et de la réutilisation des données. Ce cycle de vie dont l'esprit est assez comparable a celui des objets matériels implique qu'à chaque étape, les questions sont légion pour traiter correctement la donnée. Rares sont cependant les cycles de vie qui intègrent un volet « fin de vie » des données. On présuppose également que la pertinence de créer et conserver ces données est déjà actée. Le cycle de vie sera donc pertinent à utiliser une fois ces deux interrogations levées. On pourra alors également se poser la question des impacts environnementaux et s'interroger sur les points d'amélioration possibles sur ces aspects.

4 La donnée : état des lieux

4.1 Échelle globale

La quantité de données qui circule sur Internet est tout simplement vertigineuse. Les comparaisons sont légion pour estimer ces volumes car les chiffres ne nous sont pas véritablement accessibles intellectuellement. Pour 2020, la masse des données générées, copiées ou consommées sur la toile est estimée à plus 60 zettaoctets [34] avec une croissance là-aussi exponentielle, portée par une explosion des usages. Ce volume devrait atteindre ou dépasser les 180 zettaoctets en 2025 et plus de 600 zettaoctets en 2030 (1 zettaoctet représente 251 030 030 052 003 DVD de 4,38 Go, soit une pile de DVD de 251 millions de kilomètres, suffisante pour dépasser le soleil depuis la terre...).

Cependant, seul un petit pourcentage de ces données nouvellement créées est réellement conservé, puisque seulement 2% des données produites et consommées en 2020 ont été sauvegardées et conservées en 2021. Parallèlement à la forte croissance du volume de données, la « base installée » de la capacité de stockage devrait augmenter, à un taux de croissance annuel composé de 19,2 % sur la période de prévision de 2020 à 2025. En 2020, la base installée de la capacité de stockage a atteint 6,7 zettaoctets.

Cette croissance est bien sûr en grande partie portée par les flux vidéo (streaming, « tubes ») qui occupent une grande part de ce volume et de nos usages. La conséquence de cet usage majoritaire spécifique implique à la fois une croissance quantitative des flux échangés, lié à la fois au nombre croissant de connectés, mais aussi à la croissance des usages et des évolutions technologiques. Cela entraîne également mécaniquement une croissance qualitative matérielle et logicielle. Par ailleurs, les augmentations de la définition des images – 4K – et du volume de données associé entraînent une pression sur la capacité à transporter ces flux de données, générant de nouveaux réseaux physiques – 5G et fibre par exemple – s'empilant sur les technologies déjà très performantes en place. Ainsi, juste en France, les usages individuels, en augmentation constante, sont déjà de l'ordre de 140 Go de données consommés par mois pour un usage fixe et 8,3 Go pour un usage en mobilité.

La croissance de ces volumes est également portée par la progression fulgurante des capacités matérielles à manipuler les flots de données, annihilant ainsi toute velléité d'effort dans l'amélioration des logiciels et des formats de données. Une page web moyenne est maintenant bien souvent plus « lourde » en octets (de l'ordre de 2.3 MB) que le *package* d'installation du jeu Doom [35], un des premiers jeux en 3D de l'histoire informatique. On ne peut que constater la croissance inexorable du poids moyen des pages web depuis 2010 avec une augmentation de presque 400 % en 11 ans (même si cette croissance semble se stabiliser aujourd'hui [36]). Cette croissance malgré l'absence de recherche d'amélioration logicielle et data est rendue possibles par un réseau performant, qu'il soit mobile (4 et 5G) ou fixe (fibre) [37][38] rendant les machines « anciennes » obsolètes car incapables de supporter le poids de la navigation. Cette obsolescence forcée induit des changements de matériel accélérés pour pouvoir gérer les quantités de données stockées (distribuées ou pas) d'une part, et l'accès-même via un réseau de plus en plus performant à ces données.

Même si regarder seulement l'angle du volume de données n'est clairement pas un indicateur suffisant, cela reste un élément de sensibilisation aux impacts du numérique qui d'ailleurs sera utilisé comme tel dans nos prochaines factures de consommation box et mobiles [39].

4.2 Au sein de l'Enseignement Supérieur et de la Recherche

Si les données utilisées dans l'ESR ont une vocation scientifique et éducative, les volumes en jeu ne sont pas neutres et encore moins négligeables. Par ailleurs, le « devoir » d'exemplarité de l'ESR paraît également un aspect relativement important : en effet, en formant les futurs cadres, chercheurs, enseignants et ingénieurs, une sensibilisation aux enjeux environnementaux et aux impacts spécifiques du numérique paraît le minimum à assurer tandis que les usages autour des données actuelles se doivent d'être réfléchis pour l'avenir tout en intégrant cette dimension.

Que ce soient les sciences techniques avec les simulations environnementales, les calculs mathématiques, la physique théorique ou expérimentale, l'intelligence artificielle, la biologie, ou même les sciences humaines et sociales, les outils numériques sont partout, rendant d'immenses services... mais participant à la croissance des usages et aux pollutions liées au numérique. Il serait peut-être sage au vu de la situation de mettre dans la balance les bénéfices attendus pour l'environnement de ces services par rapport à leurs impacts réels.

En effet, la science a longtemps été dirigée par des méthodes rigoureuses, reproductibles et fiables qui partent d'une théorie et des hypothèses et qui cherchent à valider par l'observation. Cependant, depuis plusieurs années, une science plus exploratoire à vu le jour exploitant les *big data* pour pousser plus loin l'exploration que celle que nous pourrions le faire humainement ou produire des hypothèses initiales qui peuvent être testées à l'aide d'une approche plus déductive. Il existe bien sûr plusieurs dangers liés à l'utilisation de ces méthodes allant des biais à la question fondamentale de savoir si le jeu exploré reste suffisant pour donner une validation scientifique au résultat. Il y a enfin une importante question concernant la compréhension des processus sous-jacents, notamment, par exemple, en ce qui concerne les *deep nets*, qui donnent des résultats « justes » sans qu'on sache, dans certains cas, pourquoi et comment… Et parfois juste parce qu'ils ont appris complètement autre chose que ce à quoi on s'attendait.

De plus, au delà des volumes mis en jeu et nécessaires à cette orientation de la science, on peut constater que ces données massives sont assez fragmentées et **hétérogènes, et ce pour plusieurs raisons :**

– leur nature initiale (calculatoire, génomique, physiologique, biologique, environnementale, sociale, géographique, etc.) ;

– leur format (texte, valeurs numériques, signaux, images 2D et 3D, fichiers audio) avec en plus la multiplicité des représentations disponibles ;

– leur dispersion au sein de plusieurs systèmes d'information (allant du PC de bureau ou pire, de vieux supports devenus illisibles, aux grands centres de calcul et passant par des *clouds*, « souverains » ou pas) ;

– la qualité et la fiabilité de ces jeux de données et la méthodologie appliquée pour gérer cette validité (des jeux de données à fréquence et variétés élevées risquent de comporter de nombreuses erreurs invalidant la qualité des connaissances déduites) ;

– l'accessibilité de ces jeux de données, leur volatilité dans le temps. Cet aspect est fondamental étant donné la tendance des formats et des outils utilisés pour générer et analyser les données à devenir obsolètes, et les efforts nécessaires pour mettre à jour les infrastructures de données afin de garantir l'accès aux données sur le long terme ;

– enfin, la valeur et la signification attribuée à ces données dépend autant de l'utilisation prévue des données que des circonstances historiques, sociales et géographiques.

4.3 Des enjeux de volumes

Partout, les volumes gérés et stockés sont déjà conséquents et les projections indiquent une augmentation significative à venir. Nous allons illustrer ces volumes au travers de quelques exemples significatifs :

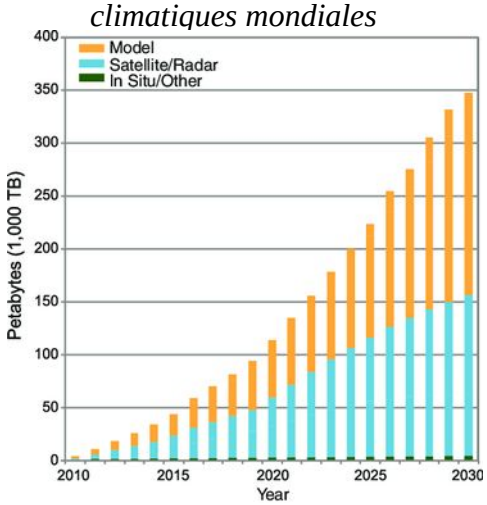
– Le premier exemple est celui du ECMWF (European Centre for Medium-Range Weather Forecasts »[2], qui permet l'accès à plus de 260 millions de fichiers, soit 210 pétaoctets de données (+40 % depuis 2015) et qui accumule plus de 233 téraoctets journaliers supplémentaires sans compter les 46 pétaoctets de données sauvegardées (+40 % depuis 2015).

– Un autre exemple, plus « proche » de nous, est celui du Centre de calcul (CC) de l'IN2P3 dont la stratégie principale est la mutualisation. La spécificité de la discipline est le volume très important de données à gérer, en quantité croissante : 1 pétaoctet par mois soit 12 pétaoctets par an et une conservation à très long terme (> 10 ans). « Avec plus de 100 pétaoctets stockés qui ne nous appartiennent pas sur disques et bandes, le stockage va rapidement devenir un problème prégnant » tandis que le nombre de serveurs (plus de 1000 aujourd'hui) et de calculs explosent (plus de 3 millions de jobs exécutés par mois). La projection à 2030 indique un volume en Exaoctet. Les défis à relever sont multiples : augmentation importante des volumes, calculs complémentaires (IA, ML, etc.) qui s'additionnent à ce volume mais aussi, disposer de bande passante et d'outils de transfert tout en maîtrisant les coûts. « Faire 10 fois plus c'est compliqué, c'est pourquoi il faut impliquer la communauté dans la définition des outils » (Benoît Delaunay, directeur adjoint, CC-IN2P3 Lyon).

– Le centre de stockage des données satellite Allemand : Plus de 60 pétaoctets [43].

– Un dernier exemple concerne la croissance exponentielle du volume des données climatiques mondiales [40] (comprenant à la fois les observations *in situ*, satellitaires et les données de modèles) et les projections pour 2030 : des volumes

Figure 2: volume des données climatiques mondiales



2 ECMWF : https://www.ecmwf.int/en/computing/our-facilities/data-handling-sy

de 350 à 400 pétaoctets attendus et une explosion des volumes des données archivées de la NOAA [41].

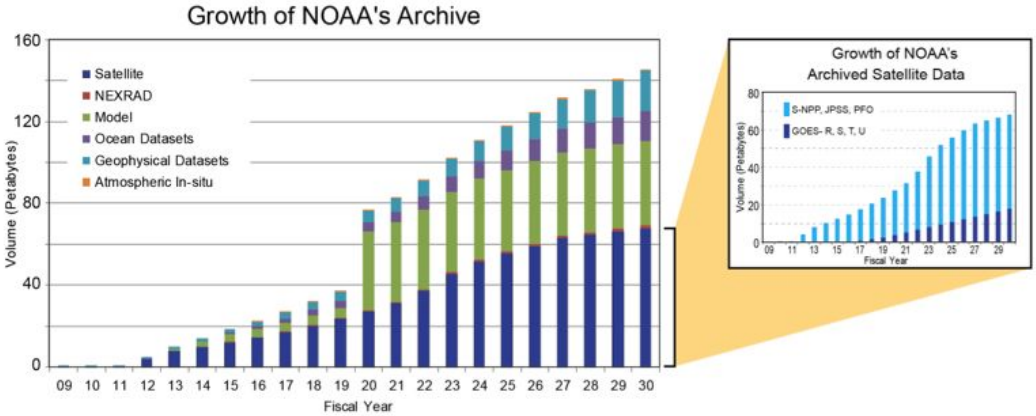


Figure 3 : La croissance des archives de la NOAA

5 La donnée : Les solutions

5.1 Sobriété

La piste la plus simple pour réduire nos impacts est de penser en terme de sobriété et de « ne pas faire ». De fait, l'infrastructure matérielle est installée pour répondre à des pics d'usage et la partie réseau ne consommera pas moins si on ne s'en sert pas à moins de l'éteindre et de la déconnecter. En effet, la consommation d'énergie des infrastructures de transport est quasi indépendante de son usage. Ce n'est pas le cas des terminaux et des centres de données qui eux ont une consommation en utilisation croissante avec l'usage et non nulle quoi qu'il en soit, même éteints. L'idée de la sobriété ici est surtout de ne pas courir après une croissance matérielle qui entraînera à la fois un remplacement et/ou un empilement de technologies. La 5G, si elle rend obsolète certains matériels, n'élimine pour autant ni la 4G, ni même la 2G, dont les antennes émettent encore sur le territoire. Les trois technologies se superposent, permettant ainsi de conserver en état de fonctionnement les matériels non compatibles 5G (ou 4G, etc.), mais augmentant mécaniquement la consommation d'énergie existante, malgré une conception théoriquement moins énergivore. On pourrait donc se poser la question d'arrêter l'émission des 4G et antérieures, mais ceci provoquerait de fait une vague d'obsolescence pour de nombreux matériels. Enfin, la course technologique entraîne

également de nouveaux usages qui, par effet rebond, annihilent trop souvent les améliorations d'efficacité énergétique indéniables des nouveaux matériels. La sobriété reste donc une piste à privilégier en tout premier lieu en s'interrogeant sur l'utilité de produire tel ou tel jeu de données, de l'utilité de le conserver [42] « à chaud », voire de le conserver tout court s'il est produit.

5.2 La science ouverte et le *FAIR*

La science ouverte est une nouvelle approche du processus scientifique basée sur le partage et la coopération. Depuis 2016, elle s'est imposée progressivement dans le cadre des appels d'offres européens (H2020) avec l'obligation selon le principe « aussi ouvert que possible, aussi fermé que nécessaire » qui définit l'accès libre et gratuit aux publications et aux données issues des travaux financés par ses fonds. Ces obligations ont été reprises dans la loi française pour une république numérique et se retrouvent dans les appels d'offres de l'Agence Nationale de la Recherche.

Les enjeux autour de cette ouverture des données vont bien au-delà des enjeux de partage. En effet, l'ouverture des données et des logiciels est la seule possibilité pour sortir des logiques propriétaires et permettre de réduire la profonde fracture numérique qui existe entre ceux qui possèdent les logiciels et les données et ceux qui n'y ont pas accès. Une grande partie des problématiques qui se posent à nous aujourd'hui sera de décider dans quel monde nous souhaitons vivre : le numérique à travers la science ouverte peut jouer un rôle d'ouverture et de fédération ou un rôle d'exclusion et de fermeture.

La dynamique d'ouverture des données en plaçant comme objectifs le partage et la réutilisation des données a exacerbé la recherche de critères standardisés utilisables tant par l'homme que la machine. Les principes *FAIR* ont ainsi été définis. Cet acronyme, qui a émergé en 2016 dans un article de Mark Wilkinson *et al.* [44], est devenu depuis une notion structurante qui met en exergue les 4 principes qui régissent le partage et la réutilisation des données et métadonnées : *Findable* (facile à trouver), *Accessible* (accessible), *Interoperable* (interopérable), *Reusable* (réutilisable). Ces éléments peuvent, intelligemment employés, apporter également de la sobriété dans la gestion des données. Par exemple, l'interopérabilité peut permettre une réduction drastique des formats de fichiers et des outils logiciels développés pour les manipuler, voire même réduire les besoins matériels nécessaires à cette manipulation. Les principes *FAIR* ne permettent pas spécifiquement une approche écoresponsable, mais participent au respect des données, à leur conservation dans des conditions d'interopérabilité et de réutilisation, diminuant par là-même le besoin de nouvelles acquisitions.

41 critères ont été proposés sur ces principes par le *RDA FAIR Data Maturity Model Working Group (Research Data Alliance)*. Une première phase de réflexion, travaux d'appropriation menée dans le cadre international et notamment européen par les nombreuses instances qui rayonnent autour d'EOSC (European Open Science Cloud) , *FAIRSFAIR project* entre autres, en relation avec les parties prenantes (chercheurs, organismes et institutions de recherche, infrastructures et plateformes), se conclut avec le consensus suivant : les principes *FAIR* définis en 2016 et explorés depuis sont une bonne base de travail de laquelle il faut partir et qu'il faut adapter aux spécificités disciplinaires. Ce consensus théorique doit maintenant être confronté à la réalité opérationnelle en prenant en compte les spécificités par discipline notamment via les plateformes ou infrastructures

de recherche. De très nombreux travaux sont initiés sur le sujet dans plusieurs disciplines [45][46][47].

La réflexion et le travail ont été élargis aux autres données ou éléments associés aux données comme les logiciels, les *workflows*, les cahiers de laboratoires électroniques, etc.

On peut prendre pour exemple le travail en cours sur l'adaptation des principes FAIR aux logiciels de recherche par le groupe de travail RDA FAIR for Research Software (FAIR4RS) mais aussi Force11, s'appuyant sur les principes du SIRS (EOSC Scholarly Infrastructures for Research Software[48]) qui sont un peu différents de FAIR.

L'ouverture des données de la recherche est un mouvement qui prend de l'ampleur et se structure. Au niveau français, la science ouverte (SO), c'est notamment le déploiement de la plateforme Recherche data gouv, dont l'objectif est de clarifier le paysage (3 600 entrepôts à travers le monde) avec une offre de dépôt des données pour les thématiques dépourvues d'entrepôts disciplinaires de confiance et le signalement via le moissonnage des données stockées ailleurs.

La structuration du réseau des administrateurs des établissements est en cours ainsi que la cartographie des entrepôts de confiance, institutionnels ou thématique. L'existence du département services et infrastructures numériques du MESRI illustre la volonté de la transformation numérique nécessaire de l'ESRI où il faut dépenser mieux, rationaliser et mutualiser pour plus d'efficacité. Bien s'organiser pour bien s'insérer dans l'organisation européenne, décloisonner, assurer l'interconnexion des centres de calcul..

D'autant plus que tout ce travail sur les données n'est utile que si les données proposées au partage ont une bonne qualité intrinsèque et que si leur acquisition est plus coûteuse pour l'environnement (ou impossible, comme dans le cas de données uniques) que le fait de les « *FAIR*iser » et de les conserver.

Enfin, que ce soit dans l'objectif de réduction de l'impact environnemental ou dans la stratégie de déploiement de la science ouverte, on peut se demander : « À quel moment questionnons-nous nos besoins ? »

Un juste équilibre doit être trouvé entre traçabilité et conservation des données et la question doit être soulevée à chaque étape du cycle de vie de la donnée sachant que cet arbitrage a pour corollaire un impact environnemental.

Ainsi le mouvement de la science ouverte, s'il participe à la transparence de nos travaux par la traçabilité des recherches en donnant des gages à l'intégrité scientifique, nous questionne aussi sur nos pratiques, leur sens et leurs objectifs, notamment en matière de partage et de conservation. Le challenge devant cette avalanche de données est de trouver des solutions structurelles et organisationnelles à long terme qui ont du sens.

6 Conclusion

L'*open data* (et moins largement l'*Open Science*), en rationalisant la gestion des données et en permettant le partage et la réutilisation, pourrait peut-être également permettre une maîtrise des impacts environnementaux. L'objectif est de ne pas refaire, d'être capable de réutiliser (notamment

pour avoir une science reproductible) et à terme de minimiser les copies, ce qui demande une réflexion approfondie sur les infrastructures à mettre en place. N’oublions pas cependant que cette mise à disposition pourrait entraîner un effet rebond d’usage qui annihilerait rapidement les gains environnementaux acquis de cette rationalisation.

Aujourd'hui, la situation nationale et internationale reste assez confuse. On observe plutôt une augmentation des volumes de données notamment avec l'ajout de métadonnées mais aussi un foisonnement des entrepôts et des solutions techniques qui peinent encore à émerger et se stabiliser.

À terme, cette rationalisation pourrait être bénéfique si elle permet d'avoir une donnée de qualité, interopérable et réutilisable, ce qui permettra in fine que le fruit de nos travaux de recherche soient de qualité et réexploitables. Mais ces objectifs ne sont pas si triviaux et devront être réfléchis pour être atteints avec intelligence et efficacité. On cherchera ainsi à ce que l'investissement en temps et en argent ainsi que l'impact environnemental d’acquisition ou de fabrication des ces données ne soient pas vains.

De nombreux guides existent et permettent de s’approprier et de déployer les concepts du *FAIR*, de l’Open Data et de l’Open Science [49][50][51] tandis que des progrès technologiques sont attendus pour réduire les impacts du stockage[52][53]. Il faudra bien sûr leur prêter une grande attention, rester attentif aux éventuels effets rebond tout en refusant fermement d’ attendre pour agir au vu du contexte environnemental et de sa dégradation rapide.

Si l’Open Data est une voie rationnelle, institutionnellement valorisée, et à priori prometteuse, il apparaît malgré tout plutôt clairement que la simplicité, la sobriété et le questionnement de nos usages restent cependant, de l’avis des auteurs, les priorités à mettre en place. Ces principes peuvent réellement conduire à une diminution concrète et réelle des infrastructures et des outils numériques, ce qui est la seule voie accessible rapidement face à l’urgence environnementale qui nous menace dès aujourd’hui.

Annexe : Les indicateurs environnementaux

Le jour du dépassement [14]

D'après Wikipédia, « Le **jour du dépassement**, ou **jour du dépassement de la Terre** (*Earth Overshoot Day* ou *EOD*) correspond à la date de l'année, calculée par l'ONG américaine *Global Footprint Network* [15] à partir de laquelle l'humanité est supposée avoir consommé l'ensemble des ressources que la planète est capable de régénérer en un an. Passée cette date, l'humanité puiserait donc de manière irréversible dans les réserves « non renouvelables » (à échelle de temps humaine) de la Terre. En 2021, l'ONG a estimé cette date au 29 Juillet [16] ». Les pays en développement sont sans surprise bien moins impactants que les pays riches et un rééquilibrage des niveaux de vie de l'ensemble de la population mondiale vers un état décent devra se faire en diminuant celui des pays riches pour aller vers une moyenne acceptable pour tous.

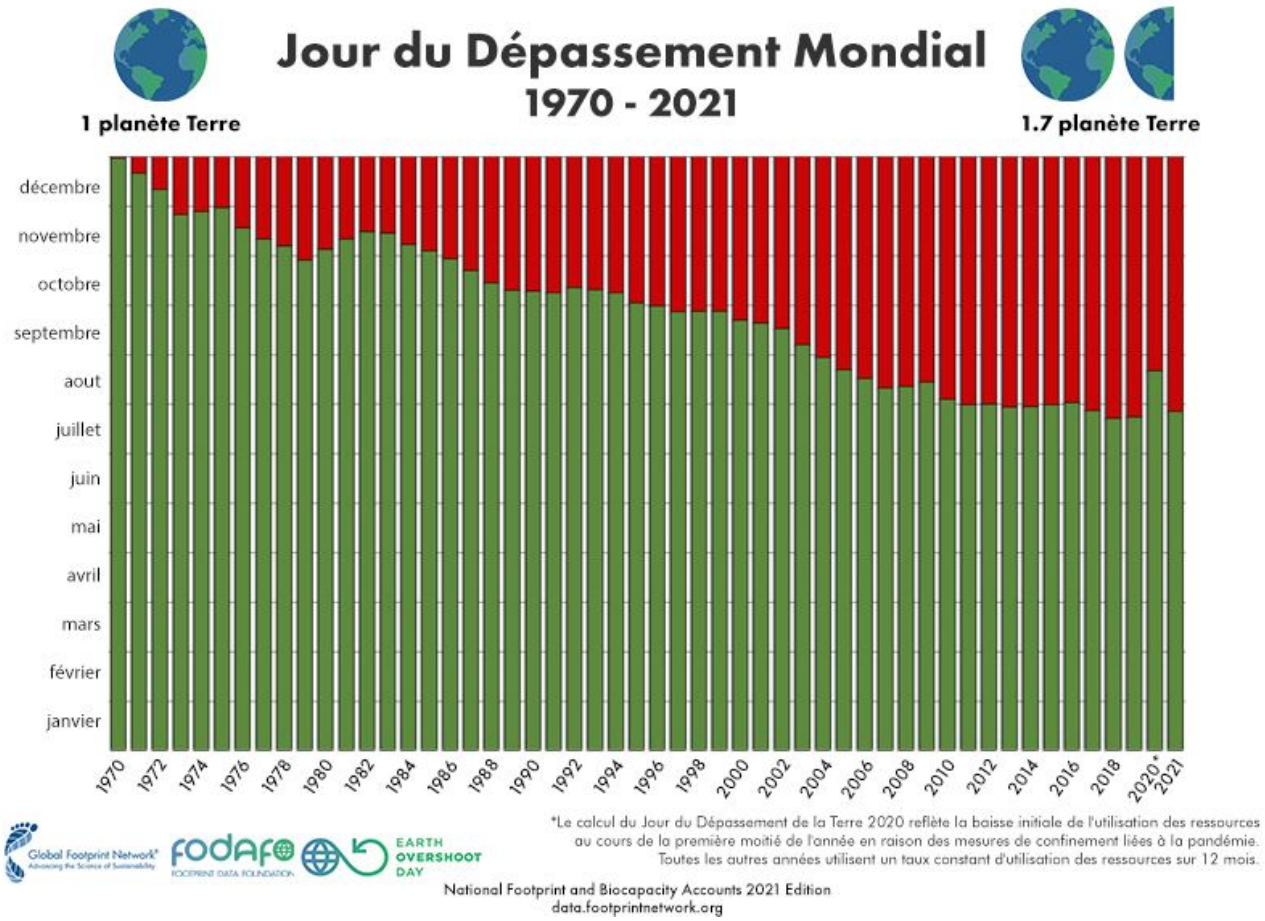


Figure 4 : Jour du dépassement mondial

Les limites planétaires [17]

D'après Wikipédia, « *Les limites planétaires sont les seuils que l'humanité ne doit pas dépasser pour ne pas compromettre les conditions favorables dans lesquelles elle a pu se développer et pour pouvoir durablement vivre dans un écosystème sûr, c'est-à-dire en évitant les modifications brutales et difficilement prévisibles de l'environnement planétaire* ».

Proposé par une équipe internationale de 26 chercheurs en 2009, ce regard plus sectoriel mais qui prend en compte des indicateurs fondamentaux et interconnectés de notre environnement planétaire, à été mis à jour en 2015 en ajoutant une dixième limite : la diffusion « d'entités nouvelles » dans l'environnement. Neufs processus naturels ont étés retenus pour leur importance dans la stabilité de notre environnement : le changement climatique, l'érosion de la biodiversité, la perturbation des cycles biogéochimiques de l'azote et du phosphore, les changements d'utilisation des sols, l'acidification des océans, l'utilisation mondiale de l'eau, l'appauvrissement de la couche d'ozone stratosphérique et l'augmentation des aérosols dans l'atmosphère.

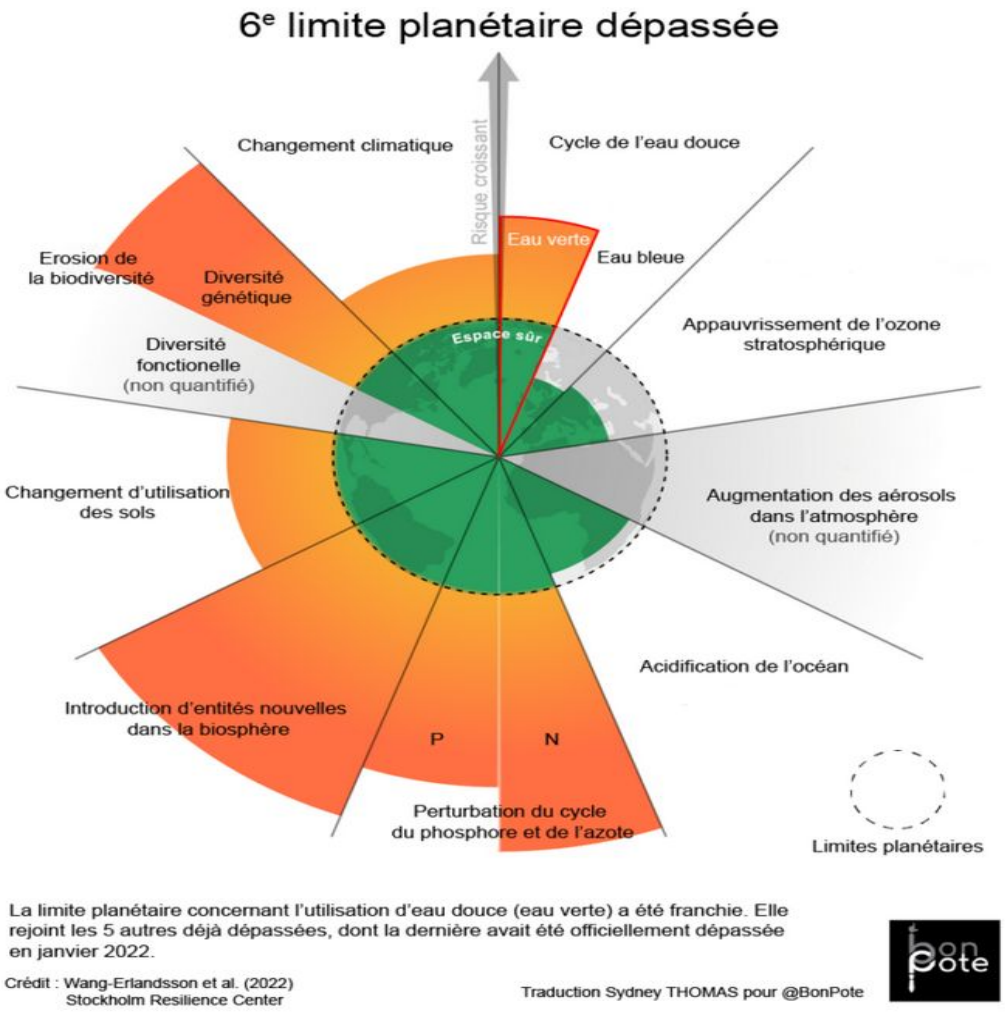


Figure 5 : Limites planétaires

La dernière mise à jour indique qu'une cinquième limite, celle qui concerne les pollutions chimiques, ou les nouvelles entités introduite par l'humanité, vient d'être franchie (Janvier 2022) tandis que la sixième limite à été dépassée officiellement en avril 2022. Quel que soit l'angle de vue sur ces indicateurs, la croissance de la population et le mode de vie consumériste prônant une croissance permanente sont bien les causes principales de la situation planétaire qui se dégrade rapidement mettant en danger le climat, la biodiversité de notre monde et par là même la survie de notre propre espèce qui est totalement dépendante d'un biotope équilibré. Les estimations d'empreinte carbone acceptable pour la planète et par humain font état de 2 tonnes équivalent carbone par personne alors qu'en France, selon l'ADEME, la moyenne actuelle est de l'ordre de 12 tonnes). Ces chiffres montrent clairement les limites pour un mode de vie soutenable pour l'ensemble de la planète et la population actuelle : alimentation, habillage, eau chaude et chauffage modéré, déplacements limités, peu de consommation de produits manufacturés et bien sûr, très peu de numérique puisque la fabrication d'un PC portable à lui tout seul produit par exemple de l'ordre de 400kg de CO^2, sachant également qu'il y a d'autres indicateurs problématiques que le CO^2 à regarder. Ces indicateurs ne font que confirmer les conclusions du rapport *Meadows* (*The limits to Growth* [18]) issu du groupe de réflexion « le Club de Rome », qui avait déjà étudié et montré dès 1972 les dangers du modèle de croissance économique occidental face aux limites physiques de notre planète [19].

Bibliographie

[1] Site du gouvernement Français : « Comprendre le changement climatique » : https://www.notre-environnement.gouv.fr/themes/climat/article/comprendre-le-changement-climatique

[2] Le site du GIEC : https://www.ipcc.ch/

[3] 6ieme Rapport du GIEC : https://www.ipcc.ch/report/ar6/wg1/

[4] « Une étude alerte sur la fonte 'exceptionnelle' des glaciers himalayens » : https://sciencepost.fr/une-etude-alerte-sur-la-fonte-exceptionnelle-des-glaciers-himalayens/

[5] « Les implications du changement climatique sur les armées, les missions de maintien de la paix et la prévention des conflits » : https://www.airuniversity.af.edu/Portals/10/ASPJ_French/journals_F/Volume-07_Issue-3/scottkhan_f.pdf

[6] rapport de *The Shift Project* publié en juillet 2019 : https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf

[7] Digital Report, Social Network & Hootsuite; https://wearesocial.com/fr/blog/2020/01/digital-report-2020/

[8] Site du collectif du CNRS EcoInfo : Effet Rebond et Numérique : https://ecoinfo.cnrs.fr/thematiques/consequences-du-numerique/effets-rebond-du-numerique/

[9] Le site de l'IPBES : https://www.ipbes.net/

[10] « La disparition des insectes » : https://lejournal.cnrs.fr/articles/declin-des-insectes-lurgence-dagir

[11] « Faits et chiffres, 3e rapport mondial des Nations Unies sur la mise en valeur des ressources en eau: l'eau dans dans un monde qui change » : https://unesdoc.unesco.org/ark:/48223/pf0000374903_fre & https://www.unesco.org/reports/wwdr/2021/fr/download-the-report

[12] « La disparition des abeilles » : https://www.nationalgeographic.fr/environnement/disparition-des-abeilles-hecatombe-involontaire

[13] Rapport de Green Cross et de Pure Earth (Blacksmith Instute) : https://greencross.ch/en/news-info-en/environmental-reports/ten-most-dangerous-sources-of-environmental-toxins-2016/

[14] Le jour du dépassement, Wikipédia, https://fr.wikipedia.org/wiki/Jour_du_d%C3%A9passement

[15] Global Footprint Network : https://fr.wikipedia.org/wiki/Global_Footprint_Network

[16] https://www.wwf.fr/vous-informer/actualites/jour-du-depassement-2021

[17] Les limites planétaires, Wikipédia, https://fr.wikipedia.org/wiki/Limites_plan%C3%A9taires

[18] Le Rapport Meadows, « the limits to growth » : https://www.clubofrome.org/publication/the-limits-to-growth/

[19] « Lecture et Analyse du rapport Meadows, » Jean-Marc Janconvici : https://jancovici.com/recension-de-lectures/societes/rapport-du-club-de-rome-the-limits-of-growth-1972/

[20] https://librairie.ademe.fr/mobilite-et-transport/3776-caracterisation-des-effets-rebond-induits-par-le-teletravail.html

[21] «Génération de visages réalistes par une IA » : https://www.journaldugeek.com/2018/12/18/visages-realistes-artificiels-ia-nvidia/

[22] Un article de vulgarisation sur GPT3, « OpenAI GPT-3 : tout savoir sur l'IA de langage la plus avancée du monde » : https://www.lebigdata.fr/openai-gpt-3-tout-savoir

[23] « Private traits and attributes are predictable from digital records of human behavior » : https://www.pnas.org/content/110/15/5802

[24] « Des chercheurs chinois auraient construit un modèle d'IA de 100 000 milliards de paramètres, soit autant de paramètres que le cerveau humain : https://intelligence-artificielle.developpez.com/actu/329908/Des-chercheurs-chinois-auraient-construit-un-modele-d-IA-de-100-000-milliards-de-parametres-soit-autant-de-parametres-que-le-cerveau-humain-il-serait-571-fois-plus-puissant-que-GPT-3/

[25] « La consommation énergétique de l'IA » : https://ecoinfo.cnrs.fr/2021/06/12/consommation-energetique-de-lutilisation-de-lia/

[26] « Biais de sur-interprétation : qu'est-ce que l'overfitting ? » : https://larevueia.fr/biais-de-surinterpretation-ou-overfitting/

[27] https://sciencepost.fr/surveillance-de-masse-la-chine-lance-une-super-camera-couplee-a-lia/

[28] « Intelligence artificielle en Chine : surveiller et punir », Charles Thibout : https://www.april.org/intelligence-artificielle-en-chine-surveiller-et-punir-charles-thibout

[29] « Dystopique : la Chine introduit un "procureur IA" capable d'accuser automatiquement les citoyens d'un crime » : https://www.etresouverain.com/2021/12/28/dystopique-la-chine-introduit-un-procureur-ia-capable-daccuser-automatiquement-les-citoyens-dun-crime/

[30] « Troubles sur les enfants face à la surconsommation d'écrans » : https://www.lemonde.fr/sciences/article/2019/10/21/michel-desmurget-la-multiplication-des-ecrans-engendre-une-decerebration-a-grande-echelle_6016350_1650684.html

[31] Leonelli, Sabina, « Scientific Research and Big Data », *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2020/entries/science-big-data.

[32] Cycle de vie des données : https://www.univ-brest.fr/drive/menu/DRIVe/Science-Ouverte/Plan-gestion-de-donnees

[33] Cycle de vie des données : https://www.w3.org/2011/gld/wiki/GLD_Life_cycle

[34] https://www.statista.com/statistics/871513/worldwide-data-created/

[35] https://www.wired.com/2016/04/average-webpage-now-size-original-doom/

[36] « Poids moyen des pages web » : https://httparchive.org/reports/page-weight?start=2017_05_15&end=latest&view=list

[37] Données ARCEP : https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/observatoire-des-marches-des-communications-electroniques-en-france/marche-des-communications-electroniques-en-france-t2-2019.html

[38] Données ARCEP : https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees.html

[39] « Consommation de données et impact carbone sur les factures des opérateurs » : https://www.francetvinfo.fr/internet/telephonie/les-donnees-generent-des-emissions-l-empreinte-carbone-de-votre-consommation-internet-desormais-inscrite-sur-votre-facture_4902575.html

[40] « Climate data are dramatically increasing in volume and complexity » : https://www.researchgate.net/figure/The-volume-of-worldwide-climate-data-is-expanding-rapidly-creating-challenges-for-both_fig2_49826559

[41] « Assessing Stewardship Maturity of the Global Historical Climatology Network-Monthly (GHCN-M) Dataset: Use Case Study and Lessons Learned », D-Lib Magazine November/December 2016, Volume 22, Number 11/12 : http://www.dlib.org/dlib/november16/peng/11peng.print.html

[42] Une BD pour réfléchir, Ugo Bienvenu, « Préférence système » : https://www.babelio.com/livres/Bienvenu-Preference-systeme/1170956

[43] « 60 Petabytes for the German Satellite Data Archive D-SDA » : https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12632/22039_read-51751

[44] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI : 10.1038/sdata.2016.18

[45] FAIR Data Maturity Model. Specification and Guidelines : https://zenodo.org/record/3909563#.YGRNnq8za70

[46] https://rdafrance2021.sciencesconf.org/data/pages/PaNOSC_FDMM_RDA_France_15102021.pdf

[47] https://doi.org/10.15497/RDA00050

[48] FAIR Principles for Research Software (FAIR4RS Principles) : https://doi.org/10.15497/RDA00065

[49] Guide de bonnes pratiques sur la gestion des données de la Recherche. Christine Hadrossek, Joanna Janik, Maurice Libes, Violaine Louvet, Marie-Claude Quidoz, Alain Rivet, Geneviève Romier, 2021, https://hal.archives-ouvertes.fr/hal-03152732/document

[50] Guide de bonnes pratiques Gestion et valorisation des données de la recherche. Pierre-Yves Arnould et Marie-Christine Jacquemot. [Rapport de recherche] OTELo ; INIST-CNRS. 2016 – https://hal.archives-ouvertes.fr/hal-01275841

[51] Ouverture des données de recherche, Guide d'analyse du cadre juridique en France. Nicolas Becard, Céline Castets-Renard, Gauthier, Martin Dantant, Laurence Freyt-Caffin Nathalie Gandon, Caroline Martin, Andrea Martelletti, Alexandra Mendoza-Caminade, Nathalie Morcrette, Claire Neirac. Décembre 2017

[52] Futura Sciences : « Stockage données sur des molécules d'ADN » : https://www.futura-sciences.com/tech/actualites/informatique-stockage-adn-nouvelle-dimension-95353/

[53] Futura Sciences : « Stockage données en 5 dimensions sur du verre » : https://www.futura-sciences.com/tech/actualites/technologie-ce-disque-verre-peut-stocker-donnees-pendant-14-milliards-annees-94564/#xtor%3DRSS-8

[54] https://www.researchgate.net/figure/Data-volume-growth-by-year-in-zettabytes_fig2_313400371

[55] https://international-soil-radiocarbon-database.github.io/SOC-Hub/2018/06/27/Big-Data/