

Le médium algorithmique*

Par Pierre Lévy[†]

Dans les sciences, l'économie ou la politique, les activités humaines sont de plus en plus fondées sur la gestion et l'analyse d'énormes masses de données numériques¹. Même si nous n'en n'avons pas clairement conscience, notre société devient progressivement *datacentrique*. Parallèlement à cette évolution, nos communications - émettrices et réceptrices de données - reposent sur une infrastructure de plus en plus complexe de manipulation automatique de symboles que j'appelle *le médium algorithmique*. Mais aussi bien la société datacentrique que le médium algorithmique sur lequel elle repose n'en sont encore qu'à leurs timides commencements. L'essentiel de leur croissance et de leur développement reste encore à venir. De plus, les esprits restent fascinés par la puissance de *diffusion* de messages offerte par l'Internet, une puissance qui n'est pas loin d'avoir atteint son terme, alors qu'un immense espace - encore inexploré - s'ouvre à *la transformation et à l'analyse* du déluge de données que nous produisons quotidiennement. A l'avant-garde de la révolution algorithmique, IEML (ou tout autre système ayant les mêmes propriétés) va démocratiser la catégorisation et l'analyse automatique de l'océan de données. Son utilisation dans les médias sociaux va créer un environnement encore plus propice qu'aujourd'hui à l'apprentissage collaboratif et à la production de connaissances massivement distribuée. Ce faisant, IEML (ou quelque autre code sémantique universel) va contribuer à faire émerger le médium algorithmique du futur et permettre une réflexion de l'intelligence collective sur l'édification de la société datacentrique à venir.

1 L'épuisement de la logique de la diffusion

1.1 L'horizon de la communication universelle

De Gutenberg jusqu'au milieu du XXe siècle, le principal effet technique des médias était d'enregistrer, de multiplier et de transmettre mécaniquement les symboles de la communication humaine. On peut notamment prendre comme exemples l'imprimerie (journaux, magazines, livres), l'industrie du disque, le cinéma, le téléphone, la radio et la télévision. Certes, il existait aussi des techniques de calcul, ou de *transformation* automatique des symboles. Mais les cal-

*A paraître dans la revue « Sociétés », Paris.

[†]Professeur au Dept de communication de l'Université d'Ottawa, titulaire de la Chaire de Recherche du Canada en Intelligence Collective, membre de la Société Royale du Canada

1. Je traduis par « masses de données » ou « données massives » l'anglais « big data ».

culatrices automatiques disponibles avant les ordinateurs n'étaient pas très puissantes et d'un usage limité.

Les premiers ordinateurs eurent peu d'impact sur la communication sociale à cause de leurs prix élevés, de leur complexité d'utilisation et de leur faible nombre de propriétaires (essentiellement les grandes entreprises, certains laboratoires scientifiques et administrations gouvernementales de pays riches). Ce n'est qu'à partir des années 1980 que le développement de l'informatique personnelle mit entre les mains d'une part croissante de la population de puissants outils de *production de messages*, qu'il s'agisse de textes, de tableaux de chiffres, d'images ou de musique. Dès cette époque, la démocratisation des imprimantes et le développement des réseaux de communication entre ordinateurs, tout comme la multiplication des chaînes de radio et de télévision, commença à remettre progressivement en question le monopole sur la diffusion massive des messages, qui appartenait traditionnellement aux éditeurs, journalistes professionnels et responsables des grands réseaux de télévision². Cette révolution de la diffusion s'est accélérée avec l'arrivée du World Wide Web au milieu des années 1990. Elle s'est épanouie dans la sphère publique mondiale multimédia d'un nouveau genre qui prévaut au début du XXIe siècle.

Sur le plan de la structure de la communication sociale, la caractéristique essentielle de la nouvelle sphère publique est de permettre à n'importe qui de produire des messages, d'émettre en direction d'une communauté sans frontière et d'accéder aux messages produits par les autres émetteurs. Cette liberté de communication est d'autant plus effective que son exercice est quasi gratuit et ne nécessite aucune connaissance technique préalable. Malgré les limites que je vais bientôt souligner, il faut saluer le nouvel horizon de communication qui s'offre désormais à nous : au rythme où croissent les taux de connexions, presque chaque être humain de la prochaine génération sera capable de diffuser ses messages à la totalité de la planète gratuitement et sans effort.

Il est certain que la manipulation - ou la transformation - automatique des symboles était pratiquée dès les années 1960 et 1970. J'ai également déjà noté qu'une bonne part de l'informatique personnelle était utilisée à produire l'information et non seulement à la diffuser. Enfin, les grandes entreprises du Web telles que Google, Amazon, E-bay, Apple, Facebook ou Twitter traitent quotidiennement d'immenses masses de données dans de véritables « usines informationnelles » entièrement automatisées. Malgré cela, pour la majorité des gens, le médium numérique est encore conçu et utilisé comme un outil de diffusion et de réception d'information, dans la continuité des médias de masse depuis l'imprimerie en passant par la télévision. Un peu comme si le Web distribuait à chaque individu la puissance d'une maison d'édition, d'une chaîne de télévision et d'un réseau postal multimédia en temps réel, tout en lui donnant accès à une bibliothèque et médiathèque mondiale omniprésente. De même que les premiers livres imprimés - les incunables - copiaient étroitement la forme des manuscrits, nous nous servons encore du médium numérique pour achever, ou pousser jusqu'à son terme, la puissance de diffusion des médias antérieurs. Tout le monde

2. Un monopole d'ailleurs bien souvent régulé ou contrôlé par les gouvernements.

peut émettre universellement. Chacun peut recevoir de n'importe où.

1.2 Les limites du médium algorithmique contemporain

Outre la censure et la surveillance malveillante exercée par les régimes dictatoriaux, il au moins trois limites à ce processus de communication mondiale.

1.2.1 Les limites cognitives

La première limite tient aux compétences cognitives des groupes sociaux et des individus : plus leur niveau de formation (primaire, secondaire, universitaire) est élevé, plus leur pensée critique³ est développée et mieux ils sont capables de tourner à leur avantage le nouvel environnement de communication. Au fur et à mesure que les points d'accès et les dispositifs mobiles se multiplient, la fameuse question de la fracture numérique⁴ recoupe de manière croissante les problèmes de l'alphabétisation et de l'éducation. Il faut noter que le simple fait de savoir lire et écrire donne déjà accès à une possibilité d'expression ainsi qu'à une foule de relations sociales et d'informations qui auraient été hors de portée sans le médium numérique⁵.

1.2.2 Les limites sémantiques

La seconde limite est sémantique puisque, si la connexion technique tend à devenir universelle, la communication du sens reste encore fragmentée selon les frontières des langues, des systèmes de classification, des disciplines et autres univers culturels plus ou moins disjoints. Le « Web sémantique » impulsé par Tim Berners Lee dès la fin des années 1990 est fort utile pour traduire des relations logiques entre données. Mais il n'a pas tenu ses promesses en matière d'interopérabilité du sens, malgré l'autorité de son promoteur et l'investissement de nombreuses équipes d'ingénieurs. Comme je l'ai montré dans le tome 1 de *La sphère sémantique*⁶, il est impossible de traiter pleinement les problèmes sémantique en restant dans les limites étroites de la logique. D'autre part, les méthodes essentiellement statistiques utilisées par Google ainsi que par les nombreux systèmes de traduction automatique disponibles fournissent d'excellent outils d'aide à la traduction, mais elles n'ont pas mieux réussi que le « Web sémantique » à ouvrir un véritable espace de communication translinguistique. Les statistiques ne sont pas plus efficaces que la logique pour automatiser le traitement de la signification. Ici encore, il manque un *codage* de la signification linguistique qui la rende véritablement calculable dans toute sa complexité.

3. La pensée critique désigne ici la capacité d'évaluer la transparence d'une source d'information, de vérifier son exactitude par recoupements et de décrypter ses présupposés et ses théories implicites.

4. *Digital divide* en anglais.

5. Voir les travaux de Manuel Castells (par exemple, *Communication Power*, Oxford University Press, 2009) et de Barry Wellman (par exemple, avec Lee Rainie, *Networked: The New Social Operating System*, MIT press, 2012).

6. *La sphère sémantique*, tome 1, Hermès-Lavoisier, Paris-Londres 2011, en particulier au chapitre 8

1.2.3 Les limites du positivisme statistique

L'accès du grand public à la puissance de diffusion du Web ainsi que les flots de données numériques qui coulent désormais de toutes les activités humaines nous confrontent au problème suivant : comment transformer les torrents de données en fleuves de connaissances ? La solution à ce problème va déterminer la prochaine étape de l'évolution du médium algorithmique. Certains observateurs enthousiastes du traitement statistique des « *big data* », comme Chris Anderson, le rédacteur en chef de *Wired*, se sont empressés de déclarer que les théories scientifiques (en général!) étaient désormais obsolètes⁷. Nous n'aurions plus besoin que de flots massifs de données et d'algorithmes statistiques puissants opérant dans les « nuages » de l'Internet : les théories - et donc les hypothèses qu'elles proposent et la réflexion dont elles sont issues - appartiendraient à une étape révolue de la méthode scientifique. Il paraît que les nombres parlent d'eux-mêmes. Mais c'est évidemment oublier qu'il faut, préalablement à tout calcul, déterminer les données pertinentes, savoir exactement ce que l'on compte, et nommer - c'est-à-dire catégoriser - les patterns émergents. De plus, aucune corrélation statistique ne livre directement des relations causales. Celles-ci relèvent nécessairement d'hypothèses qui *expliquent* les corrélations mises en évidence par les calculs statistiques. Sous couvert de pensée révolutionnaire, Chris Anderson et ses émules ressuscitent la vieille épistémologie positiviste et empiriste en vogue au XIXe siècle selon laquelle seuls les raisonnements inductifs (c'est-à-dire uniquement basés sur les données) sont scientifiques. Cette position revient à refouler ou à passer sous silence les théories - et donc les hypothèses risquées fondées sur une pensée personnelle - qui sont nécessairement à l'oeuvre dans n'importe quel processus d'analyse de données et qui se manifestent par des décisions de sélection, d'identification et de catégorisation. On ne peut initier un traitement statistique et interpréter ses résultats sans aucune théorie. Encore une fois, le seul choix que nous ayons est de laisser cette théorie à l'état tacite ou de l'explicitier. Expliciter une théorie permet de la relativiser, de la comparer avec d'autres théories, de la partager, de la généraliser, de la critiquer et de l'améliorer⁸. Cela constitue même une des principales composantes de ce qu'il est convenu d'appeler « la pensée critique », que l'éducation secondaire et universitaire est censée développer chez les étudiants.

Outre l'observation empirique, la connaissance scientifique a toujours eu à voir avec le souci de la catégorisation et de la description correcte des données phénoménales, description qui obéit nécessairement à des théories plus ou moins formalisées. En décrivant des relations fonctionnelles entre des variables,

7. Voir : de Chris Anderson « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », *Wired*, 23 juin 2008.

8. Parmi la très abondante littérature sur le sujet, voir notamment les ouvrages de deux grands épistémologues du XXe siècle, Karl Popper et Michael Polanyi.

- Karl Popper, *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, 1972 (En Français : *La Connaissance objective*. Traduction intégrale et préface de Jean-Jacques Rosat, Flammarion, collection Champs, Paris, 1998).

- Michael Polanyi, *Personal Knowledge: Towards a Post-Critical Philosophy*, University of Chicago Press, 1974 (publication originale en 1964).

la théorie offre une prise conceptuelle sur le monde phénoménal qui permet (au moins partiellement) de le prévoir et de le maîtriser. Les données d'aujourd'hui correspondent à ce que l'épistémologie des siècles passés appelait les phénomènes. Pour continuer de filer cette métaphore, les algorithmes d'analyse de flux massifs de données d'aujourd'hui correspondent respectivement aux instruments d'observation de la science classique. Ces algorithmes nous montrent des patterns, c'est-à-dire en fin de compte des images. Mais ce n'est pas parce que nous sommes capables d'exploiter la puissance du médium algorithmique pour « observer » les données qu'il faut s'arrêter en si bon chemin. Nous devons maintenant nous appuyer sur la puissance de calcul de l'Internet pour « théoriser » (catégoriser, modéliser, expliquer, partager, discuter) nos observations, sans oublier de remettre cette théorisation entre les mains d'une intelligence collective foisonnante.

2 Les débuts de la société data-centrique

Je vais maintenant décrire les premiers pas de la société data-centrique, telle qu'elle se dessine sur la base de notre nouvelle capacité - massivement distribuée - de diffusion universelle.

2.1 Une communication stigmergique

Commençons par analyser le type de communication désormais dominant au début du XXI^e siècle et qui se déroule dans ce qu'il est convenu d'appeler les « médias sociaux ». Les médias sociaux ne désignent pas ici seulement les blogs et les services classiques de réseautage tels que Facebook, Twitter ou LinkedIn. Ils comprennent aussi une foule de services de publication, de collaboration, de fouille, d'apprentissage ou de transaction en ligne qui permettent à leurs utilisateurs de trouver les informations, les produits ou les personnes qu'ils cherchent grâce à l'exploitation de l'intelligence collective en ligne. Par exemple, Wikipedia repose sur l'activité auto-organisée de millions d'auteurs et d'éditeurs. Google utilise les hyperliens affichés par des millions de sites Web pour classer ses réponses à nos requêtes. Amazon nous suggère des livres basés sur les choix des utilisateurs qui ont des profils d'achats semblables aux nôtres, etc. En fin de compte, la communication a toujours lieu entre des personnes. Mais, dans le médium algorithmique, cette communication a lieu principalement sur un mode *stigmergique*, c'est-à-dire que les personnes communiquent entre elles en modifiant leur environnement commun : les données numériques en ligne. Chaque lien que nous créons, chaque tag que nous apposons sur une information, chaque acte d'évaluation ou d'approbation, chaque « j'aime », chaque requête, chaque achat, chaque commentaire, chaque re-tweet, toutes ces opérations modifient subtilement la mémoire commune, c'est-à-dire la masse inextricable des *relations* entre les données. Notre comportement en ligne émet un flux continu de messages et d'indices qui contribue - parfois directement mais le plus souvent indirectement - à orienter et à informer les autres internautes. Or il en est évidemment

ainsi parce l'information que nous produisons individuellement est traitée par des algorithmes afin d'être transformée en information utile pour la collectivité.

2.2 Le modèle data-centrique de la communication

Dans le modèle data-centrique de la communication, l'interaction des individus avec la masse des données communes - et par cet intermédiaire avec les autres participants - peut se décomposer en quatre étapes logiquement distinctes mais pratiquement interdépendantes : la production, l'acheminement, la fouille et l'analyse.

2.2.1 La production

Les données doivent d'abord être produites avant d'être introduites dans la mémoire collective. Que ce soit pour la rédaction et l'édition de texte, de tableaux, d'images fixes ou animées, de son, de musique, de logiciel ou de paquets multimedia, le producteur de données utilise invariablement un ou plusieurs logiciels, c'est-à-dire en fin de compte des algorithmes. Il en est de même pour les données générées par les transactions économiques, par les capteurs biomédicaux ou par ceux qui permettent d'identifier, de localiser et de mesurer les actions d'objets ou de machines.

2.2.2 L'acheminement

Une fois produites et émises vers une base de données en ligne, les données sont traitées afin d'être acheminées vers les destinataires pertinents. En fonction de nos activités et de notre réseau, les algorithmes des services en ligne nous envoient des sélections d'informations, des recommandations de personnes à suivre, des suggestions d'achats, des publicités, etc. L'analyse de grandes masses de données et le filtrage collaboratif alimentent ainsi - souvent gratuitement - nos demandes d'information quotidienne et de veille sur nos sujets favoris.

2.2.3 La fouille

Une requête explicite sur un moteur de recherche (c'est-à-dire en fait sur une base de données) déclenche une sélection d'information ainsi que le classement et la présentation des données sélectionnées. Dans l'activité de fouille de données, c'est en principe l'utilisateur qui détermine les informations qu'il reçoit. Mais les algorithmes des moteurs de recherche personnalisent leurs résultats en fonction de notre profil (langue, requêtes précédentes, etc.). De plus, leurs algorithmes de sélection et de rangement ont un impact déterminant sur les résultats qu'ils nous renvoient, comme on peut s'en rendre compte facilement en comparant les résultats de la même requête sur différents moteurs de recherche.

2.2.4 L'analyse

Finalement, les données amassées ou réunies à partir de sources diverses peuvent être analysées afin d'en extraire des patterns, des régularités ou des tendances autrement indiscernables. Les résultats de cette analyse de grandes masses de données servent à mieux comprendre des processus complexes, à prévoir le futur (lorsque c'est possible) avec plus d'exactitude et à prendre de meilleures décisions. Les données qui font l'objet de telles analyses peuvent être produites « en interne » par les organismes qui les traitent (gouvernements, laboratoires scientifiques, entreprises...). Mais elles peuvent aussi être récupérées sur le Web, être mises à la disposition du public par des administrations gouvernementales de diverses échelles, ou bien encore par des entreprises pratiquant la mise à disposition publique de leurs données pour des raisons philanthropiques. Contrairement aux trois phases précédentes (production, acheminement et fouille) l'analyse automatique de grandes masses de données se trouve rarement entre les mains des utilisateurs finaux de manière gratuite et intuitive. En 2013, ces analyses automatiques sont encore le plus souvent réservées à des institutions riches et dotées de fortes compétences techniques.

2.2.5 Le rôle des algorithmes dans la communication data-centrique

Les résultats de l'analyse de grandes masses de données sont eux-mêmes des données, qui peuvent servir à produire des documents. Ces documents sont à leur tour acheminés, fouillés et ainsi de suite. Il s'agit d'un cycle. Réexaminons les étapes de ce cycle : les individus collaborent pour produire des données, reçoivent des données pertinentes d'autres groupes ou individus en fonction de leurs profils et de leurs activités, récoltent des données en réponse à des requêtes explicites et analysent les masses de données reçues. Dans tous les cas, la relation des utilisateurs avec les données est médiée par des algorithmes. Nous avons vu que, dans la société data-centrique, les personnes communiquent de manière indirecte, en modifiant la structure de relations entre les données qui constitue leur contexte commun. Or nos interactions avec la masse des informations stockées dans les « nuages » de l'Internet est médiée par des algorithmes. Ces algorithmes gèrent nos interactions sensori-motrices avec les ordinateurs et les gadgets portables, nos accès aux bases de données, nos actes de fouille, de tri, etc. Bien mieux, les algorithmes organisent le bouclage personnel et collectif de l'émission et de la réception des données, ils recommandent et gèrent les contacts personnels sur Internet, ils aménagent l'environnement économique, social et cognitif de l'intelligence collective. En somme, ce sont bel et bien les algorithmes qui forment désormais le milieu de communication au sein duquel les réseaux humains construisent et modifient collaborativement leur mémoire commune.

2.3 Les avancées multiformes d'une culture data-centrique

Plusieurs phénomènes techno-sociaux contemporains témoignent éloquentement de l'éclosion d'une nouvelle ère de la communication sociale.

Le mouvement de la « science ouverte » vise à la constitution d'un bien commun scientifique mondial comprenant non seulement les publications traditionnelles (articles, livres, rapports) mais également les données brutes et les outils logiciels ayant servi à les exploiter. La communauté scientifique internationale communique ainsi de manière de plus en plus transparente en se rassemblant autour du bien commun numérique qu'elle produit et dont elle se nourrit, comme les universités d'antan se rassemblaient autour de leur bibliothèque.

La méta-discipline émergente des *digital humanities* travaille à numériser et à mettre en ligne l'ensemble des archives, textes, documents divers et commentaires accumulés qui constituent le matériau de son activité. Elle construit de plus autour de cette masse documentaire une myriade de réseaux sociaux qui se partagent leurs outils en vue de l'exploitation collaborative des données⁹.

Des journalistes repèrent les ensembles de données prometteuses fournies par les gouvernements, les organismes de production de statistiques ou d'autres institutions. Ils analysent ces données au moyen d'algorithmes appropriés, puis visualisent et expliquent les résultats à leurs lecteurs.

Dans le domaine de la communication organisationnelle et du « *knowledge management* », on considère de plus en plus qu'une bonne gestion *sociale* des connaissances émerge, sur le mode *bottom-up*, des activités *personnelles* des membres de l'organisation pour gérer leurs propres connaissances. L'entreprise ou l'administration devient ainsi un médium social facilitant trois processus complémentaires. Premièrement, ses membres, ses clients et ses partenaires accumulent des données communes codant le savoir qui leur est utile. Deuxièmement, chacun peut accéder à ces données au moment opportun. Troisièmement, des conversations ouvertes entre les participants au réseau facilitent aussi bien la codification du savoir à destination du collectif que son appropriation personnelle en situation de travail.

Dans leur version connexionniste, les cours ouverts en ligne « massifs » (MOOC¹⁰) transcendent l'éducation à distance classique pour stimuler l'apprentissage collaboratif en réseau, les étudiants contribuant à produire le matériel pédagogique commun sur une multitude de plate-formes interconnectées.

La plupart des formes de *crowdsourcing*, tout comme la communication et le partage de fichiers « pair à pair » (*P2P*) illustrent ce modèle de relation sociale dans lequel une foule d'individus se rassemblent virtuellement autour d'un trésor de données communes pour - simultanément - l'enrichir et l'exploiter.

2.4 Les institutions data-centriques

Au-delà de ces tendances d'avant-garde, les institutions humaines comme les administrations gouvernementales, les collectivités locales, les universités, les écoles, les entreprises grandes ou petites et les associations de toutes sortes reposent, pour leurs performances quotidiennes (administratives, budgétaires,

9. Voir sur ce point, de Michael Nielsen, *Reinventing Discovery: The New Era of Networked Science*, Princeton University Press, 2012

10. *Massive open online course*

légal, stratégiques, etc.) sur la gestion informatisée de données numériques. On peut distinguer deux grands types de données à cet égard.

Premièrement, les données « internes » représentent les institutions à elles-mêmes. Ceci inclut leurs archives, leurs compétences ou savoirs clés, les opérations en cours, ainsi que les informations sur leurs membres, clients ou bénéficiaires. Les institutions sont généralement responsables de la production et de l'entretien des données sur lesquelles reposent leur mémoire et leurs activités quotidiennes. De telles données sont évidemment produites et gérées en fonction du sens que leur attribue l'institution, un sens qui tient à ses finalités et à ses traditions.

Deuxièmement, les données « externes » sont produites par d'autres institutions, y compris les médias traditionnels. Ces données externes représentent l'environnement de l'institution et font souvent l'objet d'un travail de veille scientifique, industrielle, économique, politique, stratégique et ainsi de suite. Pour l'institution en question, le but de cette activité de renseignement est de comprendre son environnement, de prévoir autant que possible son évolution et de s'y adapter - ou d'y intervenir de manière proactive. Comme les données externes sont produites et organisées en fonction de finalités ou d'angles interprétatifs qui ne sont pas les siens, l'institution doit les recoder - ou les re-catégoriser - à sa manière.

Il est clair que la frontière entre les deux types de données, interne et externe, est hautement perméable. En outre, les institutions sont de plus en plus organisées et interconnectées en réseaux par l'intermédiaire desquels elles échangent leurs données et accomplissent un grand nombre de transactions en ligne. Les institutions deviennent data-centriques dans la mesure où elles s'identifient de moins en moins à leurs infrastructures matérielles (y compris d'ailleurs les ordinateurs eux-mêmes) et à leur localisation physique alors qu'elles s'identifient de plus en plus au noyau dur de leurs données stratégiques ainsi qu'au mode de codage et de traitement des données - internes et externes - qui sont nécessaires au maintien de leur intégrité. Ces données fondamentales - et les algorithmes qui les traitent - peuvent être virtualisées dans les nuages de l'internet, permettant ainsi aux collaborateurs de l'institution d'y accéder à tout instant et de n'importe où.

2.5 Les conflits data-centriques

Les conflits politiques ou militaires tournent également de manière croissante autour du contrôle des données numériques en ligne. Les récentes campagnes électorales se sont certes largement jouées dans les médias sociaux, sur le mode classique de l'agitation et de la propagande. Mais on a aussi vu apparaître une tendance complémentaire : les équipes gagnantes ont massivement investi dans l'analyse automatique de grandes masses de données sur les électeurs et leurs opinions. Elles ont de plus réussi à exploiter en temps réel sur le terrain les résultats de ces analyses ¹¹.

11. Voir par exemple : « Beware the Smart Campaign » de Zeynep Tufekci, in *New York Times*, 16 Nov. 2012

Les révoltes arabes de 2011 ont utilisé les téléphones intelligents et les médias sociaux pour leur organisation interne et la diffusion de leurs messages¹². Symétriquement, les dictatures concernées ont tenté de couper ou de brouiller les réseaux, elles ont infiltré les communautés en ligne pour obtenir des renseignements sur les révolutionnaires et elles se sont livrées à des activités de contre-propagande et d'intoxication dans le medium numérique. Il n'est plus un mouvement social sur la planète qui n'utilise désormais la puissance de coordination et de diffusion des réseaux sociaux. Mais aussi bien les gouvernements que les services de renseignement et de police utilisent les mêmes outils de diffusion, d'exploitation, de manipulation et d'analyse des données.

Quels que soient les camps en présence, il s'agit toujours d'analyser et d'influencer la mémoire à court terme (l'actualité) et à long terme (l'histoire) des communautés concernées, une mémoire qui s'incarne désormais dans des flux et des stocks de données numériques. Du côté de l'analyse, on tente de percevoir les signaux faibles et les patterns d'évolution qui annoncent le futur bien qu'ils n'attirent pas l'attention de la majorité. Du côté de l'influence, on tente de faire oublier, de passer sous silence ou de discréditer certaines données et l'on veut en revanche attirer l'attention sur les données « favorables » (voire les créer artificiellement) et leur conférer un maximum de crédibilité. Les médias classiques (grands journaux et télévision) ne sont plus que des sources de données comme les autres. Ces sources médiatiques sont certes particulièrement puissantes du fait de leurs moyens et de leur crédibilité, mais leurs effets se mesurent désormais à leur capacité différentielle d'influencer la connexion affective des personnes et des groupes avec la masse planétaire des données.

Jusqu'au XIXe siècle, les armées se divisaient en armée de terre et en marine. Le XXe siècle a vu le développement d'une troisième arme majeure : l'aviation. Dès la seconde décennie du XXIe siècle, une quatrième arme a fait son apparition dans les forces de défense de la plupart des grandes et moyennes puissances : *l'arme informatique*. La cyberguerre comprend évidemment l'espionnage, la désinformation et les jeux de simulation. Mais elle ne s'y limite pas car, dans ce cas, elle aurait pu rester confinée à la fonction de renseignement, de propagande et de réflexion stratégique des autres armes. En plus des fonctions classiques qui viennent d'être nommées, les cyberattaques visent le blocage des réseaux et la destruction des données-clés de la force militaire, du gouvernement et de l'économie des puissances ennemies, ainsi que la prise de contrôle d'installations et d'infrastructures à distance. Cette prise de contrôle peut aller jusqu'au sabotage physique des installations et à la disruption complète des infrastructures. Du côté de la défense, la nouvelle arme a pour mission de protéger les réseaux et les données stratégiques des institutions militaires, gouvernementales, industrielles et économiques d'un pays. Il va sans dire qu'une des compétences majeures attendues des cyber-ninjas est la conception et la programmation des algorithmes. Il faut noter en outre que toutes les branches des armées se sont

12. Voir par exemple, de Reda Benkirane, *The Alchemy of Revolution: The Role of Social Networks and New Media in the Arab Spring*, Geneva Centre for Security Policy GCSP Policy Paper 2012/7 http://www.archipress.org/reda/index.php?option=com_content&task=view&id=133&Itemid=1

dotées de capacités de communication, de coordination et de renseignement (embarquées, interactives et en temps réel) qui reproduisent - en les augmentant - les grandes fonctions des médias sociaux civils.

3 Le développement du médium algorithmique

Sans doute faut-il avoir épuisé les possibilités logiques de la *diffusion* automatique - à savoir la vertu médiatique des quatre derniers siècles - afin de rencontrer et de commencer à assimiler sur un plan culturel le potentiel encore presque inexploité de la *transformation* automatique : la vertu médiatique des siècles qui viennent. C'est pourquoi je parle ici de médium *algorithmique* : afin de souligner la capacité de transformation automatique de la communication à support numérique. Bien entendu, la puissance de transformation ou de traitement du nouveau médium ne peut s'actualiser que sur la base de l'accomplissement irréversible de la mission du médium précédent, à savoir la diffusion universelle ou l'ubiquité de l'information. Cette ubiquité étant quasiment atteinte au début du XXI^e siècle, les générations qui viennent vont progressivement apprivoiser le traitement automatique du flot océanique des données mondiales, avec toutes les conséquences culturelles imprévisibles que cela va entraîner.

Aujourd'hui, la plupart des algorithmes qui gèrent l'acheminement des messages et la fouille des données sont opaques, puisqu'ils sont protégés par le secret commercial des grandes compagnies du Web. Quant aux algorithmes d'analyse ils sont, pour la plupart, non seulement opaques mais aussi hors d'atteinte de la majorité des internautes pour des raisons à la fois techniques et économiques. Or il est impossible de produire de la connaissance fiable au moyen de méthodes secrètes. Il faut évidemment considérer l'état contemporain du médium algorithmique comme transitoire. On ne pourra apprivoiser culturellement la croissance exponentielle des données - et donc transformer ces données en connaissance réfléchie - que par une mutation qualitative du médium algorithmique. Mais pour comprendre comment nous en sommes arrivé là, et surtout pour prévoir la manière dont la société data-centrique et le médium algorithmique vont continuer à se développer dans l'avenir, il nous faut préalablement explorer l'essence abstraite de la manipulation automatique des symboles.

3.1 La structure algorithmique

Afin de bien saisir la nature du nouveau médium, nous devons nous représenter le plus clairement possible ce qu'est un algorithme et comment il fonctionne.

3.1.1 Codage

Pour que des traitements automatiques puissent être appliqués à des données, il faut évidemment que les données en question aient été préalablement codées de manière adéquate et uniforme. Il ne s'agit pas seulement ici du codage binaire (zéro et un), mais de types de codages plus spécialisés. On peut donner comme exemple le codage des nombres (base deux, huit, dix, seize, etc.) celui

des caractères d'écriture, celui des images (les pixels), celui des sons, et ainsi de suite. Je souligne à ce sujet qu'IEML se présente comme *un système de codage de la signification linguistique* propre à la rendre calculable, exactement comme le système des pixels a rendu les images manipulables par des algorithmes.

3.1.2 Opérateurs

Il faut ensuite imaginer un ensemble d'outils ou de micro-machines spécialisées dans l'exécution de certaines tâches. Apellons ces outils spécialisés des « opérateurs ». Les opérateurs sont précisément identifiés et ils agissent de manière entièrement déterminée et mécanique, toujours de la même manière. Il doit évidemment exister une correspondance ou une adéquation entre le codage des données et le fonctionnement des opérateurs.

Les opérateurs ont d'abord été identifiés à l'intérieur des ordinateurs : ce sont ici des circuits électroniques élémentaires. Mais on peut considérer n'importe quel processeur de données - aussi complexe soit-il - comme une « boîte noire » faisant office d'opérateur. C'est ainsi que le protocole de l'internet, en adressant les ordinateurs dans le réseau, a du même coup ouvert un système universel d'adressage des opérateurs.

3.1.3 Conteneurs

Il faut de plus se représenter un entrepôt de données dont les cases ou « conteneurs » élémentaires sont parfaitement adressées : un système d'enregistrement logique offrant une surface lisse pour l'écriture, l'effacement et la lecture. Il est clair que le codage des données, les opérations qui leur sont appliquées et leur mode d'enregistrement doivent être harmonisés pour optimiser les traitements.

Le premier système d'adressage des conteneurs est interne aux ordinateurs, et il est géré par leur système d'exploitation. Mais au-dessus de cette couche d'adressage interne, les URL du World Wide Web ont instauré un système d'adressage des conteneurs qui est universel.

3.1.4 Instructions

Le quatrième et dernier aspect d'un algorithme est un ensemble ordonné de règles - ou un mécanisme de contrôle - qui organise la circulation récursive des données entre les conteneurs et les opérateurs. La circulation est initiée par un flux de données qui part des conteneurs vers les opérateurs appropriés puis dirige les résultats des opérations vers des conteneurs précisément adressés. Un ensemble de tests (si... alors...) détermine le choix des conteneurs où puiser les données à traiter, le choix des conteneurs où inscrire les résultats et le choix des opérateurs. La circulation de données s'achève lorsqu'un test a déterminé que le traitement était terminé. Dès lors, le résultat du traitement - à savoir un ensemble de données codées - se trouve à une adresse précise du système des conteneurs.

3.1.5 Les instructions dépendent du jeu d'opérations ainsi que du codage et de l'adressage des données

Dans les explications populaires ou vulgarisées de la programmation, on réduit souvent l'algorithme à une série d'instructions ou à une « recette ». Mais aucun algorithme ne peut jouer son rôle sans disposer préalablement des trois éléments qui suivent : premièrement, un code adéquat pour les données ; deuxièmement, un ensemble bien déterminé d'opérateurs réifiés - ou de fonctions se comportant comme des boîtes noires ; troisièmement, un système de conteneurs précisément adressés capables d'enregistrer les données initiales, les résultats intermédiaires et le résultat final. Les règles - ou instructions - n'ont de sens que par rapport au code, aux opérateurs et aux adresses-mémoire. Je voulais souligner ici un point bien connu des spécialistes, mais dont l'importance est rarement appréciée par le public : le type de traitement, la portée et la performance des algorithmes dépendent étroitement de l'identité des opérateurs, ainsi que du codage et de l'adressage des données.

Je vais maintenant m'appuyer sur l'analyse qui vient d'être faite de la structure intime des algorithmes pour analyser l'évolution du médium algorithmique. Nous allons voir que les grandes étapes de la croissance de ce médium sont précisément reliées à l'apparition de *nouveaux systèmes d'adressage et de codage*, aussi bien pour les conteneurs de données que pour les opérateurs.

3.2 L'oeuf (1940-1970)

De quand peut-on dater l'avènement du médium algorithmique ? Nous pourrions être tentés de lui donner comme date de naissance 1937, puisque c'est cette année-là que Alan Turing a publié son fameux article introduisant pour la première fois le concept de machine universelle, c'est-à-dire le schéma formel d'un ordinateur. L'article en question représente les fonctions calculables par des *programmes* de la machine universelle, c'est-à-dire au fond par des algorithmes. Nous pourrions également choisir 1945, puisque von Neumann publia au mois de juin de cette année le « *First draft of a report on the EDVAC* » où il présente l'architecture fondamentale des ordinateurs : 1) une mémoire contenant les données et les programmes (ces derniers codant les algorithmes), 2) une unité de calcul arithmétique et logique, 3) une unité de contrôle capable d'interpréter les instructions des programmes contenus dans la mémoire. Puisque les textes séminaux de Turing et von Neumann ne représentent que des avancées théoriques, nous pourrions dater la nouvelle ère de la construction et de l'usage effectif des premiers ordinateurs, dans les années 1950. Il est clair cependant que (malgré la préscience de quelques visionnaires), jusqu'à la fin des années 1970, on pouvait encore difficilement parler d'un médium algorithmique. L'une des principales raisons est que les ordinateurs étaient encore à cette époque des grosses machines coûteuses et refermées sur elles-mêmes, dont les interfaces d'entrée et de sortie ne pouvaient être manipulées que par des spécialistes. Quoique déjà dans l'oeuf, le médium algorithmique n'avait pas encore de prévalence sociale. On notera que, entre 1950 et 1980, les flux algorithmiques de données circulent

majoritairement entre *des conteneurs et des opérateurs aux adresses locales*, encloses dans une seule machine.

3.3 L'éclosion (1970-1995)

Une nouvelle tendance se manifeste au cours des années 1970 et prend le dessus pendant les années 1980 : l'interconnexion des ordinateurs. Le protocole de l'Internet (inventé en 1969) s'impose par rapport à ses concurrents pour adresser les machines dans les réseaux de télécommunication. C'est aussi la période pendant laquelle l'informatique devient personnelle. Le numérique est désormais perçu comme un vecteur de transformation et de communication de tous les symboles, et non seulement des nombres. Les activités de la poste, des télécommunications, de l'édition, de la presse et de la diffusion radio-télévisée se mettent à converger.

A ce stade, les données traitées par les algorithmes sont toujours abritées par des conteneurs aux adresses locales mais - en plus de cette adresse - *les opérateurs ont désormais une adresse physique universelle dans le réseau mondial*. En conséquence, les opérateurs algorithmiques peuvent « collaborer » et l'éventail des types de traitement et d'applications s'élargit considérablement.

3.4 La maturation (1995-2020)

L'Internet n'est cependant devenu un support de communication majoritaire - au point de remettre en cause irréversiblement le fonctionnement des médias traditionnels et celui de la plupart des institutions économiques, politiques et culturelles - qu'à partir de l'arrivée du Web autour de 1995. La révolution du Web s'explique essentiellement par *la création d'un système d'adresses physiques universelle des conteneurs*. Il s'agit bien entendu des URL. Notons que - comme dans le cas du protocole de l'Internet pour les opérateurs - ce système universel vient s'ajouter aux adresses locales des conteneurs de données, il ne le supprime pas.

Dès lors, la puissance effective et la capacité de collaboration - ou d'interopération - des algorithmes s'accroît et se diversifie prodigieusement, puisque ce sont maintenant aussi bien les opérateurs que les conteneurs qui possèdent des adresses universelles. La machine programmable fondamentale devient le réseau lui-même, comme le montre la généralisation du « *cloud computing* ».

La décennie 2010-2020 voit le début de la transition vers une société data-centrique. En effet, à partir de cette phase du déploiement social du nouveau médium, les interactions entre personnes empruntent majoritairement le canal de l'Internet, que ce soit pour la pure et simple sociabilité ou pour l'information, le travail, la recherche, l'apprentissage, la consommation, l'action politique, le jeu, la veille et ainsi de suite. Parallèlement, les algorithmes interfacent de manière croissante les relations entre les gens, les relations entre les données et les relations entre les gens et les données. La montée des conflits autour de la propriété et du libre accès aux données, ainsi qu'autour de l'ouverture et de la transparence des algorithmes, sont des signes sûrs de la transition vers une

société data-centrique. Cependant, comme je le disais au début de ce chapitre, malgré leur rôle déjà déterminant, les algorithmes ne sont pas encore perçus dans la conscience collective comme le nouveau médium de la communication et de la pensée humaine. On reste encore fasciné par la logique de la diffusion des médias antérieurs.

3.5 L'envol (2020-...)

Ce qui n'existe pas encore est notoirement difficile à observer ou à reconnaître, et plus encore l'absence de ce qui n'existe pas encore. Or ce qui bloque aujourd'hui le développement d'un véritable médium algorithmique - et du même coup l'avènement d'une nouvelle civilisation - est précisément l'absence d'un système universel et calculable de méta-données sémantiques. Je rappelle que la sphère sémantique fondée sur le langage IEML est la première et (à ma connaissance) la seule candidate à ce rôle de système de coordonnées sémantique des données.

Nous disposons déjà d'un système d'adressage physique universel des données (le Web) et d'un système d'adressage physique universel des opérateurs (l'Internet). Dans sa phase de plein déploiement, le médium algorithmique comprendra en plus *un code sémantique universel : IEML*. Ce système de méta-données - conçu dès l'origine pour optimiser la calculabilité du sens tout en multipliant sa différenciation à l'infini - ouvrira le médium algorithmique à l'inter-opérabilité sémantique et donnera prise à de nouveaux types de manipulation symbolique.

Si les *données* d'aujourd'hui correspondent aux *phénomènes* de la science classique, alors nous avons besoin de *métadonnées* calculables et interopérables qui fassent le pendant aux *théories et aux modèles explicatifs* de la science classique. IEML se présente justement comme un outil algorithmique de théorisation et de catégorisation capable d'exploiter la puissance de calcul du « *cloud* » et d'apporter un indispensable complément aux récents outils algorithmiques d'observation de patterns.

IEML propose une méthode automatisée de définition de concepts et de relations entre concepts. Les données catégorisées en IEML pourront être traitées de manière beaucoup plus efficaces qu'aujourd'hui puisque les catégories et les relations sémantiques entre catégories deviendront alors non seulement calculables mais automatiquement traductibles d'une langue à l'autre¹³. De plus IEML permettra de *comparer* les résultats fournis par l'analyse du même ensemble de données selon des règles de catégorisation (des théories!) différentes.

Lorsque ce système symbolique d'analyse et de synthèse conceptuelle se retrouvera démocratiquement entre toutes les mains, se traduira automatiquement dans toutes les langues et se manipulera aisément à partir d'une simple tablette, alors l'océan des données pourra être apprivoisé et le médium algorithmique sera directement expérimenté comme un outil d'augmentation cognitive - personnelle et sociale - et non seulement comme un instrument de diffusion.

13. Pour être plus précis, les réseaux sémantiques hypertextuels pourront être traduits automatiquement pour toutes les langues figurant dans le dictionnaire multilingue d'IEML.

Sur la base de ce métalanguage ouvert de génération et de reconnaissance de réseaux sémantiques, un cycle auto-générateur d'expérimentation collective et de création d'outils mènera au décollage du médium algorithmique du futur.

4 La révolution de la communication n'est pas terminée

Entre 1950 et 1980, l'oiseau du médium algorithmique dormait dans l'oeuf des gros ordinateurs. Il a écloé comme réseau entre 1980 et 1995 avec l'Internet et les PC. Il a accompli sa maturation multimédia, interactive et sociale grâce au Web depuis 1995. Mais il n'a encore jamais véritablement décollé. Il s'envolera à la fin de la seconde décennie du XXI^e siècle, à partir d'un seuil critique dans l'adoption d'IEML (ou de tout autre système universel de codage computationnel du sens). Cela signifie d'abord que la « révolution de la communication » est tout sauf terminée et que nous n'en avons encore vu que les premiers débuts. Plus la société humaine et ses institutions seront data-centriques, et plus la maîtrise des données - à commencer par leur maîtrise intellectuelle - deviendra un enjeu scientifique, politique, social et économique majeur. La capacité universelle d'émission étant acquise, l'évolution techno-culturelle va se déplacer vers le perfectionnement des capacités collaboratives de production, d'analyse, de synthèse et de transformation des flux et stocks de données. Aujourd'hui, seuls les gouvernements, les grandes entreprises et les grands centres de recherche académiques ont les moyens financiers et les compétences pour extraire des informations pertinentes d'énormes masses de données. Demain, des outils accessibles gratuitement à partir de toutes les tablettes permettront de produire, d'acheminer, de fouiller et d'analyser collaborativement l'océan des données. La maîtrise théorique et pratique de ces outils sera enseignée à l'école primaire. Les algorithmes sur lesquels ils reposeront seront distribués, ouverts, transparents, inter-opérables, personnalisables à volonté et utiliseront toute la puissance du *cloud computing*. IEML se situe à l'avant-garde de ce mouvement indissolublement technique, scientifique et culturel.

Le médium typographique supportait la reproduction et la diffusion automatique des symboles. Il a permis le développement de formes culturelles inédites avant lui telles que la science expérimentale et mathématique de la nature, l'économie industrielle et l'état-nation. Le médium algorithmique supporte la manipulation et la transformation automatique des symboles. La société data-centrique basée sur le médium algorithmique du futur sera sans doute aussi différente de la société industrielle que l'Égypte pharaonique basée sur l'écriture hiéroglyphique l'était des tribus préhistoriques.