L’archive et la massification des données : une nouvelle raison numérique

Bruno BACHIMONT

Mesdames et messieurs, je voudrais tout d’abord remercier l’Association des archivistes français pour son invitation dans ce magnifique centre de congrès et saluer son inconscience puisque, dans la mesure où je ne suis pas archiviste de profession, les organisateurs ne peuvent anticiper la nature des propos que je vais tenir.

Je vais cependant revenir sur des thèmes déjà largement évoqués dans les communautés du document et de l’archive, et en particulier sur le numérique, en le considérant non seulement comme un outil, mais surtout comme un milieu dans lequel nous évoluons désormais tous, qui reconditionne notre manière de nous relier au monde, à la culture, à l’environnement. Nous pensions naguère que la relation au monde passait à travers la compréhension de sa physique, l’enseignement des sciences de la nature constituant un socle incontournable des cursus scolaires et supérieurs, et nous devrons désormais penser notre relation au monde à travers la compréhension des artefacts numériques qui nous relient à lui.

Je vais aborder la question du numérique sous l’angle de ses actions sur les contenus en précisant d’où je parle et quelles sont mes interrogations de départ. La matrice initiale est constituée par le document, envisagé comme la permanence et la mémoire de l’événement. Un événement ponctuel correspond à ce qui arrive, au sens étymologique du terme, et la volonté de fixer la mémoire de cet événement dans la pérennité d’un support et la permanence de sa lecture a justifié la mise en place d’outils qui permettent de maintenir les contenus dans la durée.

Un contenu correspond dans ce système à un support matériel, une trace, qui permet de manifester physiquement et perceptiblement une forme interprétable (texte, image, son, etc.) possédant une valeur sémiotique (c’est-à-dire que cette forme est un signifiant qui nous adresse un signifié, qui nous fait signe et qui fait sens) et par conséquent exprimant un sens pour celui/celle qui l'interprète. Cette trace donne lieu à un travail interprétatif et renvoie à la question de la maîtrise des codes sémiotiques présidant à l'interprétation de cette trace, question qui ne se pose qu'à un niveau sémiotique et culturel. Il y a donc bien la nécessité de maîtriser les interprétants de la trace pour être capable d'en assurer l'interprétation. Mais, et c'est là un point important, le numérique et plus généralement les supports techniques documentaires modifient la consultation de la trace ne nécessitant pas de médiation technologique : pour ce que j'appelle les « médiums perceptifs » un outil de décodage n'est en effet pas utile pour la lecture. Face à ces médiums, seule la perception est requise pour accéder à la nature du contenu et laisser place au travail d'interprétation à travers la maîtrise des codes culturels utilisés pour produire cette trace et de ceux nécessaires pour la compréhension de l'interprétation lors de sa consultation. Avec le numérique, et plus généralement les « médiums technologiques », il faudra aussi prendre en compte, outre notre capacité culturelle à écrire et lire, notre capacité technique à coder sur un support (binaire, électronique, etc.) et à décoder.

La situation a évolué avec l'invention, depuis plus d'un siècle, du codage qui permet de décomposer des documents en, d'une part, une ressource codée accessible uniquement à travers une médiation technique et, d'autre part, des vues publiées produites par cette médiation. Le document consulté se réduit alors à une reconstruction à partir d'une ressource codée permanente. Cela constitue des médiums technologiques (audiovisuel, numérique) et le travail de la mémoire documentaire est réparti sur deux objets :

- la ressource codée assurera le rôle de la permanence, mais n'est plus directement consultable – l'accès à cette ressource nécessite une médiation technologique. La lisibilité directe a été perdue au profit de la permanence ;
- la vue publiée, reconstruite à partir des ressources codées *via* une médiation technique afin d'accéder à une lisibilité. Mais la perception intervient dans l'éphémère de la consultation, le temps par exemple de l'ouverture de l'écran : la vue n'est pas permanente.

Après un contenu initial qui assure la permanence et la lisibilité, le complexe documentaire devient une ressource permanente, mais pas directement lisible, et une vue publiée lisible, mais éphémère, les deux étant reliés par des dispositifs techniques de codage/décodage. Seul l'ensemble des deux permet d'assurer la conservation de la permanence et de lisibilité du contenu, c'est-à-dire la mémoire de l'événement.

Outre la maîtrise des conditions d'interprétation, il faut désormais avoir la maîtrise des conditions de lisibilité technique et des moyens de surmonter le fossé d'obsolescence (décrochage progressif entre les conditions techniques de la ressource et le système technique contemporain de lecture permettant d'y accéder et de recouvrer à partir de la ressource une vue consultable et lisible). Désormais, l'éphémère de la consultation doit se construire à partir de la permanence de la ressource, mais celle-ci est devenue illisible sinon par une médiation technologique.

Le numérique introduit en outre un principe supplémentaire de codage universel où il est un support *anonyme* et *gyrovague*.

- Anonyme

Le code numérique constitué de « 0 » et de « 1 » est littéralement non sémantique. Une suite de « 0 » et de « 1 » ne décrit jamais l'objet représenté et la lecture de ce code nécessite une convention technique arbitraire et extrinsèque qui peut librement varier : un même code peut être lu de multiples manières. Un même flux binaire produira par exemple du son dans un lecteur audio ou une vidéo dans un lecteur vidéo. L'intérêt de ce caractère non sémantique est double : non seulement c'est une abstraction sémantique car elle permet une relecture arbitraire variable d'un même code, mais c'est également une abstraction matérielle car le même code numérique peut être réalisé sur différents supports physiques. Ainsi, le fichier numérique peut se révéler identique sur des supports sans aucun rapport entre eux d'un point de vue physique comme un support optique (DVD) et un support magnétique (disque dur). On parle souvent d'implémentation matérielle pour caractériser ce type d'abstraction.

- Gyrovague

Le numérique permet de créer une couche conventionnelle arbitraire unifiant la diversité des supports physiques et ouvrant un espace interprétatif inédit et constitue en cela un support à « inertie sémantique nulle » : il n'oppose aucune résistance au changement d'interprétation, seul suffit un autre code ou une autre manière de le lire.

Ces éléments sont bien connus et utilisés entre autres par les artistes. Lors d'un concert de musique contemporaine par exemple, une prestation instrumentale donnera lieu à un enregistrement et une numérisation en temps réel, le code numérique ainsi obtenu donne lieu à son tour à une lecture vidéo projetée sur grand écran ; ce qui fait que ce qui sera visionné coïncidera temporellement avec ce qui sera entendu. Cependant, un arbitraire de la désémantisation du son en flux binaire et de la resémantisation du flux binaire en image s'interposera entre les deux. Malgré l'absence de rapport entre les deux, sinon une trace causale qui ne respecte pas du tout les relations de ressemblance entre son et image, la synchronisation du visionnage et de l'écoute ne peut pas empêcher la création d'un lien entre ce qui est entendu et ce qui est visionné pour les animaux sémiotiques tels que les humains qui ne peuvent pas s'empêcher de donner du sens à ce qu'ils perçoivent et de corréler ce qui est simultané.

Cet effet esthétique de corrélation son-image alors qu'ils n'ont qu'un lien arbitraire est recherché par l'artiste et il joue à plein sur la possibilité technique numérique d'être un support technique sans sémantique propre dépendant uniquement de la manière de le relire pour lui donner un sens appréhendable. Les arts médiatiques se révèlent toujours intéressants dans la mesure où les artistes explorent les possibles techniques inédits et imaginent généralement bien plus loin que les ingénieurs les possibilités des techniques créées. Ils constituent à ce titre un fantastique laboratoire pour comprendre les problèmes du type de l'archive ou de la mémoire posés par le numérique. Une œuvre d'art numérique est un concentré de tous les cauchemars qu'il est possible de croiser dans le monde de l'archive et de la mémoire. En reprenant un mot d'esprit souvent véhiculé par les institutions de conservation d'art médiatique, la méthode de conservation la plus efficace d'une œuvre consiste à conserver le numéro de téléphone de l'assistant musical qui avait permis de concevoir l'œuvre parce que tout le reste correspond à des solutions techniques devenues illisibles ou incompréhensibles au cours du temps. La complexité technique amène à ce que ce qui pourrait être considéré comme relevant du répertoire n'est plus répétable ce qui pose bien évidemment des problèmes importants.

En résumé, après une unicité première où l'objet lisible permanent permet la mémoire de l'événement, il est désormais décomposé en une ressource permanente qui doit donner lieu à une vue publiée.

Cette désarticulation du document en ressources *versus* visualisation s'est concrétisée en trois étapes principales :

- une numérisation du document : des pratiques déjà possédées dans le monde physique ont été transférées sur le numérique (recherche documentaire : c'est le paradigme de la bibliothèque en ligne) ;

- passage du document à la notion de ressource annotée : le paradigme est celui du Web des données ou du Web sémantique. Les documents ne sont plus directement accessibles, mais ce sont des ressources ou des fragments documentaires qui deviennent autonomes les uns des autres et qui n'existent qu'à travers leurs annotations généralement formalisées avec des formats (RDF, etc.) ;

- enfin, la donnée manipulée apparue plus récemment. Selon ce paradigme du *big data*, les données sont insignifiantes de manière isolée mais permettent de construire des effets de sens par leur masse et elles sont restituées à travers des visualisations toujours saisissantes parce que très belles, suggestives et difficiles à interpréter. Le problème des *big data* réside dans le paradoxe ressenti face à des traitements appliqués sur une masse de données produisant des visualisations riches mais difficiles à interpréter.

Dans cette déconstruction du document en ressources puis en données, deux problématiques nouvelles sont apparues, qui correspondent à ce que le numérique fait aux contenus ou aux documents de manière générale :

- Le nominalisme de la variante

Ceci correspond à une tension entre la ressource permanente et la vue publiée. Plusieurs possibilités existent pour publier une ressource avec le numérique ce qui engendre une multiplicité de variantes différentes face à une ressource permanente, intègre dans sa caractérisation numérique mais inaccessible comme telle, et accessible seulement à travers les multiples vues/variantes que l'on peut reconstruire à partir d'elle.

La difficulté consiste à pouvoir gérer le rapport 1- n, c'est-à-dire la ressource une *versus* les n vues qui permettent de la consulter. La tension documentaire se déplace de la préservation de la ressource vers la prise en compte de la vue éphémère du document et du poids redonné à la variante, c'est-à-dire à la vue publiée, et suppose donc de passer de l'essence du document dans la ressource vers sa manifestation individuelle et singulière (vue publiée). Ce nominalisme implique un renoncement à l'essence documentaire comme exemplaire de référence, pour aborder une anthologie de la variante : le document n'existe plus qu'à travers des variantes protéiformes et possédant chacune leur propre légitimité singulière.

- Le nominalisme de la donnée

Ceci implique une tension entre données et visualisations. La donnée devient le vecteur de l'intelligence et de la compréhension du fait social et du fait humain, et l'interprétation des *big data* permet donc de renégocier un nouveau rapport à la culture et à la mémoire.

Le nominalisme de la variante ne fait que renouer avec des savoirs acquis depuis très longtemps par l'archivistique. Si des gens sont prêts à assumer cette conséquence du numérique, c'est bien le monde de l'archive puisqu'il travaille depuis longtemps la tension émanant des multiples variantes sous lesquelles se manifeste une œuvre, à l'instar des manuscrits antiques et médiévaux par exemple, où la multiplicité des variantes font de l'œuvre une abstraction reconstruite et postulée par la critique de ces dernières (on peut relire à ce propos avec profit et plaisir l'*Éloge de la variante* de Bernard Cerquiglini). Le numérique nous réapprend que ces leçons restent pertinentes. Le nominalisme de la variante ne constitue pas un problème mais représente au contraire un acquis de la profession et de la culture archivistiques. Elles savent parfaitement gérer cette tension propre au numérique, l'archivistique du numérique (le numérique comme problème ou objet) étant davantage l'enjeu que l'archivistique numérique (le numérique comme outil).

En revanche, les *big data* posent le problème du nominalisme de la donnée et du statut du fait humain lorsqu'il est conditionné à travers des traitements statistiques appliqués sur des faits passés au crible du formatage numérique et des données qui ne sont plus lues mais qu'il suffit désormais de calculer – ce problème se pose collectivement au-delà du monde de l'archive. Cette tension entre calcul et interprétation devra certainement être reconsidérée.

Abordons ces révolutions nominalistes avec un interlude puisque le nominalisme fleure bon des mots que vous avez entendus naguère ou jadis dans vos parcours universitaires et scolaires. Je veux parler en effet de cette notion apparue au Moyen Âge, mais qui a connu un acmé chez Guillaume d'Ockham, rencontré notamment avec Umberto Eco (*Le Nom de la rose*). Ce nominalisme représente la critique en règle du rapport entre le langage et le monde existant jusque-là : le réalisme médiéval. L'agencement syntaxique d'une phrase construite selon les règles logiques et syntaxiques reflète l'organisation du monde selon ce réalisme. Ainsi, en exprimant que « l'homme est un animal », cela signifie que l'essence de l'homme possède l'essence d'animal parmi ses propriétés. Cette phrase reflète dans sa structure (sujet, copule, prédicat) l'organisation du monde à travers les essences et les entités qui le constituent. Ceci implique la nécessité de maîtriser les arts du langage (*trivium* médiéval) puisque la compréhension du monde émane de la compréhension du langage. L'organisation des mots reflète peu ou prou l'organisation des choses.

Le nominalisme médiéval a ruiné cette conception. Selon ce paradigme, le monde n'est constitué que d'individus sans essence générale expliquant leur nature et leur ressemblance. Dès lors, l'organisation du langage et des mots ne reflète pas l'organisation des choses et les lois du monde ne sont pas à rechercher dans les lois du langage. Le nominalisme médiéval a ouvert un espace, à travers l'abolition de cette relation classique entre le mot et la chose, pour la recherche d'une nouvelle relation entre la pensée et la nature et notamment ce qui sera occupé par le paradigme calculatoire et expérimental qui a commencé avec les calculateurs d'Oxford mais a été consacré par la révolution copernico-cartésienne. Un texte ancien de Hans Jonas argumente que la véritable révolution scientifique n'est pas celle des XVIe et XVIIe siècles, comme cela est souvent avancé, mais est intervenue plus tôt au XIVe siècle avec ce nominalisme qui a dissout la relation classique du mot à la chose pour ouvrir un autre rapport au monde : le monde de la nature, objet de la mesure et du calcul. Le lien du mot à la chose n'est plus ontologique, mais désormais sémiotique, et le lien de la science à la chose est expérimentale et calculatoire.

Selon moi, nous vivons sans doute aujourd'hui une deuxième révolution nominaliste avec les *big data*. Elle ne concerne cependant plus le monde de la nature comme au XIVe siècle, mais la relation du mot au fait humain ou au fait social qui sera alors reconditionnée. Désormais, il suffit de rassembler les documents en masse suffisante pour les calculer, les interpréter, produire des anticipations ou des visualisations d'éléments extraits de ces informations sans avoir besoin de lire nous-mêmes ces documents pour les interpréter. La relation est désormais fondée sur le calcul au lieu d'avoir un rapport à la culture reposant sur de la médiation de la langue et la compréhension de l'humain. A la fin du Moyen Âge, une relation de calcul a remplacé la relation de l'homme à la nature, désormais ce sera une relation de calcul remplaçant la classique relation de l'homme à la culture à travers sa compréhension du langage.

Cette révolution nominaliste est entraînée par l'existence de masses considérables de données rendues possibles par le numérique qui donnent au paradigme calculatoire un espace de jeu inédit qu'il n'avait jamais trouvé. Tout le monde connaît les bases de données qui fleurissent, soit institutionnelles (Institut national de l'audiovisuel), soit semblables à YouTube qui engrangent un nombre considérable de données sans évoquer les métadonnées commerciales rassemblées de manière éparse. La révolution nominaliste remplace l'analyse qualitative de la culture à travers la compréhension langagière par une analyse statistique et quantitative. Le rapport à l'humain et au social change totalement de nature par rapport à l'exploitation de données.

C'est la raison pour laquelle je parle de révolution nominaliste. Elle change complètement notre rapport au monde. Ce fut jadis avec la nature et nous négocions désormais un nouveau rapport à la culture à travers ce nominalisme de la donnée, une évolution problématique à mon sens, alors que le nominalisme de la variance n'est pas problématique même s'il s'agit d'un fait massif auquel nous devons nous adapter.

Cette évolution, que j'estime problématique pour ma part même si beaucoup y voient plutôt une promesse et une chance, n'était pas inscrite d'avance dans le paradigme numérique qui s'est progressivement instauré.

Le numérique a commencé plutôt par une belle histoire et une bonne nouvelle. Les chantres du numérique, notamment dans les années 1990, s'appuyaient sur les bonnes propriétés du code binaire pour esquisser les conséquences positives qui en résulteraient pour le monde documentaire. En effet, le substrat binaire permet d'effectuer, d'un support physique à un autre, des copies parfaites car identiques bit à bit : bien que réalisés sur des supports matériels différents, les bits sont les mêmes et donc le contenu binaire est identique : comme je l'ai rappelé plus haut, c'est ce qui nous permet de considérer que nous avons le même document sur notre DVD ou sur notre disque dur, alors que matériellement ce sont des supports aux propriétés physiques totalement différentes. Or, ce principe de la copie parfaite permet de lever deux obstacles quasi mythiques pour la manipulation et l'usage de documents :

- la production de copies parfaites permet de rendre un document accessible à plusieurs endroits à la fois ; on peut avoir désormais un accès *ubiquitaire* au contenu, ou, comme le disent les économistes, un accès non concurrent : le fait que j'accède à un contenu ne prive pas une autre personne d'y accéder ;
- le fait de pallier la corruption des supports en recopiant parfaitement un support avant que sa corruption physique n'altère les conditions de lecture du document : le document devient indépendant de la corruption physique du support pour être (quasi) *éternel*.

Le fantasme du document ubiquitaire et éternel est donc apparu puisque le binaire permet d'être identique malgré la diversité des supports physiques et de pallier les défauts matériels. Dans un DVD, il est seulement demandé à une bosse d'être différente d'un trou et réciproquement, ce qui permet de reconnaître un zéro d'un un. Le DVD peut donc subir quelques tracas avant que sa lecture ne devienne impossible. Mais cela survient malheureusement assez rapidement puisque selon un rapport de l'Académie des sciences un DVD dure entre 10 et 20 ans. Même si le support numérique apporte en théorie la copie éternelle et ubiquitaire, l'intervention de supports physiques, d'une complexité et d'une fragilité inédites dans l'histoire de l'humanité, conduit à estimer que le numérique – pour reprendre l'expression de Churchill – est le pire des supports, même si je n'en connais pas de meilleur, ou à l'exclusion de tous les autres.

Mais ces propriétés d'éternité et d'ubiquité reposent sur la considération du code binaire comme tel, c'est-à-dire la succession des 0 et des 1. Mais on ne fait rien avec des 0 et des 1 : il faut savoir les lire pour reconnaître le code d'un nombre ou d'un caractère ; pour cela il faut entre autres savoir dans quel sens les lire (de droite à gauche ou de gauche à droite : 1010 signifie 10 ou 5 par exemple), bref tout un ensemble de règles et conventions techniques, arbitraires mais indispensables, pour exploiter ces codes. Ce rêve numérique fait de 0 et de 1 a donc été confronté à d'autres éléments, les quatre cavaliers de l'apocalypse des conventions de lecture numérique, conduisant à la vie concrète du numérique, nous chassant du paradis du binaire pour tomber dans la vie réelle du numérique.

Commençons par la question des formats. Personne n'a jamais affaire à des zéros et des uns en dehors des ingénieurs pour lesquels ce sont des trous et des bosses pour les supports optiques par exemple. Ce sera lu à travers des interfaces qui exploitent des *formats*. Les ressources ne sont jamais directement accessibles en tant que telles, mais possèdent quasiment un statut nouménal pour reprendre le vocabulaire kantien : elles sont observées à travers les phénomènes qu'elles manifestent, les interfaces à travers nos écrans d'ordinateur. Le programmeur ne visualise les ressources qu'au travers de la rédaction de son code et les manipulations sur l'interface du programmeur. L'utilisateur ne les visualise qu'à travers l'interface des utilisateurs. Le substrat binaire sur votre disque dur ou ailleurs est inaccessible en tant que tel, sinon par une médiation, et ne propose jamais de rapport immédiat, sans médiation, avec lui. Le cadre de manifestation du numérique intervient sous la forme des formats avec les métadonnées, l'enfer des applications et des plateformes. Le substrat binaire s'exprimera ainsi à travers un empilement de couches introduisant sa variabilité et ses particularités au lieu d'avoir un contenu restant identique à lui-même. Ainsi, le document ne sera jamais visualisé deux fois de suite de la même manière. Nous sommes en fait dans le flux héraclitéen au lieu d'avoir une permanence parménidienne, éléatique du substrat binaire qui traverse le temps de manière inaltérable : le même contenu n'est jamais visualisé deux fois dans le même fleuve numérique.

Tout cela a introduit une rupture avec le paradis perdu, qu'on peut symboliser par le livre (en particulier imprimé), soit le monde des médiums perceptifs où ce qui est vu correspond à ce qui est conservé et à ce qui peut être lu. Mais pourquoi alors avoir quitté ce paradis, puisque le numérique introduit de la variabilité et de l'éphémère en lieu et place de documents qui pouvaient rester eux-mêmes à travers le temps, pour peu qu'on puisse garantir leur intégrité physique ? La pomme qui nous a chassés de l'Eden du paradigme livresque est l'audiovisuel, du fait de la nécessaire dissociation de la ressource permanente et de la vue republiée à partir de la ressource. En effet, si on veut conserver un objet temporel, par exemple un son ou une mélodie, on ne peut le stocker directement, il faut pour cela l'enregistrer, c'est-à-dire le coder par un objet non temporel, mais spatial, et conserver cet objet spatial pour ensuite le décoder pour reproduire l'objet temporel. Le numérique n'est alors qu'un passage à la limite de ce nécessaire codage des objets temporels, où tout contenu est désormais codé et enregistré, pour être ensuite décodé afin d'en consulter une version perceptible et lisible.

Mais, puisque ce que l'on consulte n'est pas ce qui est conservé mais ce qui est reconstruit à partir d'une ressource préservée, la version consultée dépend alors des conditions de la reconstruction, de l'outillage utilisé et de son paramétrage.

La lecture d'une ressource numérique fera alors appel à la ressource conservée, mais aussi aux informations sur la ressource (ses métadonnées, par exemple le format) et sur le contexte de l'information : la version produite est une reconstruction programmée dépendant de ces informations variables et contextuelles, ce qui aboutira à une multiplicité de vues construites à partir de la même ressource.

Il faut en effet avoir à l'esprit que toute la difficulté consiste dans le fait que le binaire, comme tel, ne signifie rien mais dépend de la manière de le lire. Avoir une ressource binaire est nécessaire, mais pas suffisant. L'accès passe seulement soit par des moyens techniques et physiques en tant qu'ingénieur soit *via* des interfaces en tant que programmeur et utilisateur. L'enjeu de la lecture portera donc sur les différences entre les multiples vues reconstruites à partir d'une même ressource, et entre une vue consultée et la ressource conservée. Ces deux grandes questions visent à déterminer la véracité de la visualisation, le rapport entre la ressource conservée et la vue consultée (en quoi ce que je vois est bien ce qu'il faut voir, ce qu'il y a à voir, ce qui est conforme à la ressource ?), et la commensurabilité des vues entre elles au sens étymologique (quelle est la mesure commune ? comment les comparer ? comme savoir laquelle est la bonne ou laquelle fait foi ?).

Considérons un exemple audiovisuel. Un outil développé à l'Ina permet de collecter des images sur le Web et de les comparer aux archives conservées. On constate alors que les images collectées ne sont pas directement comparables aux archives, mais qu'elles en constituent des variantes : par exemple une image a été améliorée en termes de contraste et des logos y ont été ajoutés ; le fond d'écran a été modifié, une image couleurs archivée a été republiée en noir et blanc pour faire « plus ancien », etc. Il est bien sûr possible de vivre avec de telles variantes. Cependant, je vous propose de vous projeter dans 400 ou 500 ans et d'imaginer la réception de ces deux images, celle en couleurs et sa version en noir et blanc, de provenances différentes. Comment alors déterminer la variante, la copie, l'originale, les règles et principes philologiques pour établir la version de référence entre les deux variantes ? Bref, qui est la copie de qui, et comment le savoir ?

Les règles interprétatives sont plus ou moins simples dans certains cas : la photographie doit être plus ancienne qu'une autre lorsque la personne photographiée est plus jeune sur l'une que sur l'autre. Cependant, l'adoption de principes philologiques devient obligatoire pour faire la part entre les variantes. Les variantes, et non les ressources, seront l'élément premier puisque les ressources, le binaire, sont désormais inaccessibles par elles-mêmes, leur provenance étant inconnue et le binaire étant ininterprétable en lui-même (il faut des règles et conventions de lecture). Par exemple une page Web est une reconstruction dynamique de ressources délocalisées et disloquées en plusieurs lieux : la seule intégrité sur laquelle il est possible de s'appuyer est celle de la vue. La variante possède donc le poids phénoménologique et ontologique du document et constitue ce sur quoi il est possible de s'appuyer. Il ne faut pas rapporter la variante à une origine totalement putative et théorique, inaccessible en tant que telle, mais il faut procéder à une reconstruction à travers une comparaison raisonnée entre ces différentes variantes et des opérations philologiques.

Face à ce constat, les archivistes témoignent d'un savoir constitué sur un tel problème, déjà ancien, et ils en maîtrisent les différentes arcanes depuis leurs études. L'approche est en effet semblable à la situation où plusieurs manuscrits médiévaux évoquent un manuscrit perdu ou supposé inexistant. Le nominalisme de la variante, un fait massif imposé par le numérique, ne représente pas un problème si déstabilisant par rapport à la pratique des archivistes, mais bien par rapport aux mauvaises pratiques et habitudes prises depuis l'imprimé et la gestion du papier (un bon-à-tirer fait foi). En dehors de cette exception de l'histoire qu'est l'imprimé depuis trois siècles, les archivistes sont habitués à gérer la prolifération des variantes. L'ouvrage de Bernard Cerquiglini, que j'ai déjà cité, rappelle combien l'archivistique est à même de traiter le nominalisme de la variante qui ramène l'archivistique à ce qu'elle accomplit fort bien depuis ses origines.

La situation différera avec le nominalisme de la donnée puisque le phénomène de massification des données change la nature même de ce qui sera consulté et visualisé. Le phénomène des *big data* ne doit pas être confondu avec la masse de données. Les masses de données existent depuis très longtemps et sont homogènes (vidéos, textes, etc.). Les informaticiens et statisticiens s'escriment depuis 40 ans pour interpréter ce type de documents.

Deux caractéristiques s'ajoutent à la masse avec les *big data* :

- un aspect dynamique afin d'opérer une mise à jour cyclique ;
- la possibilité de gérer des données hétérogènes (logs de consultations, vidéos, textes, etc.). Le caractère sans sémantique du numérique peut permettre d'homogénéiser sur un même support binaire des informations hétérogènes. Il est possible de tout accueillir en numérique parce qu'il oublie tout et n'a pas besoin de se souvenir de l'origine de la donnée. Par son action (génération de « 0 » et de « 1 »), il les rend homogènes entre elles et manipulables selon les mêmes algorithmes.

Ces masses de données répondent au slogan des 4V souvent utilisé pour caractériser les *big data*, soulignant à la fois des caractéristiques et des problèmes :

- volume ;
- vélocité : elles évoluent tout le temps ;
- variété : elles sont en effet hétérogènes ;
- véracité : en soulignant par ce terme que les données ne sont pas toutes vraies ou certifiées. Le manque de véracité ne pose pas de problème eu égard au nombre élevé de données : il est compensé par la redondance statistique qui introduit un filtre de pertinence sur l'aléa singulier de la donnée aberrante.

Les *big data* sont un phénomène important sur le plan scientifique mais aussi médiatique, comme en témoigne une action entreprise par Google portant sur une étude épidémiologique. Google a utilisé les médias en procédant à une analyse *big data* pour observer la propagation de l'épidémie de grippe. Une comparaison entre l'estimation de la propagation par Google et les données épidémiologiques réelles du *Center for Disease Control* américain laisse apparaître une superposition quasi parfaite des deux courbes : l'épidémiologie ne serait plus nécessaire, savoir collecter les données relatives à l'apparition de nouveaux termes afin d'évaluer la propagation de phénomènes serait suffisant. Or, Google mesure les gens qui parlent de la grippe et devient un amplificateur de la rumeur tandis que le *Center for Disease Control* mesure les gens qui ont contracté la grippe. La coïncidence ne traduit pas nécessairement un rapport exploitable essentiel. Cette polémique toujours ouverte se révèle d'autant plus intéressante que cette étude fut publiée dans la revue *Nature*, référence dans le monde scientifique. Selon le paradigme qui se met en place, les sciences expérimentales et théoriques sont devenues inutiles, seules seraient utiles les sciences de données qui remplaceraient tout le reste.

De nouvelles prédictions, plus ou moins dévastatrices, naissent quant aux manières d'exercer les sciences. Lev Manovich, professeur en Californie, le gentil, évoque un nouveau paradigme scientifique pour les sciences humaines et sociales : les *Cultural Analytics*. Chris Anderson, le provocateur, explique que la science classique et la pratique savante critique disparaîtront puisqu'elles pourront être remplacées par un empirisme direct absolu semblable à celui exposé par John Locke. Avec les seules données, il serait possible de tout prédire et ce serait enfin objectif puisque calculé – un sophisme lie le calcul à l'objectivité alors qu'il se révèle tout à fait possible de calculer des éléments subjectifs.

Les *Cultural analytics* consistent à collecter les données, à les analyser avec des mathématiques évoluées. Dans les colloques de l'ANR sur les *big data*, les restitutions de données permettent d'observer un résultat à travers des tableaux de bord habituels ou des visualisations – le site *visualcomplexity.com* permet de visualiser les métaphores graphiques existantes pour envisager des bases de données massives et des *big data*.

Beaucoup parlent d'un nouveau paradigme, sous l'impulsion de Jim Gray, défunt leader de Microsoft et personnage mythique ayant procédé à une conférence sur le quatrième paradigme avant de disparaître en bateau.

Dans la dernière conférence qu'il donna, il présente l'histoire de l'humanité et de la connaissance à travers quatre paradigmes :

- le plus ancien, l'empirisme direct, décrivait les phénomènes naturels ;
- la science théorique au XVI^{e} siècle, c'est-à-dire la modélisation des phénomènes naturels par des lois mathématiques ;
- les sciences numériques inventées depuis environ cinquante ans. Les équations de la physique peuvent être utilisées avec des algorithmes numériques sur les ordinateurs, prédire des courbes, dessiner des ailes d'avion, etc. ;
- l'exploration des données depuis quelques années, c'est-à-dire la possibilité d'abandonner la branche théorique et de ne pas demander aux calculs d'appliquer seulement des lois. Le réel se manifeste à travers les

données et des schémas de compréhension et d'interprétation se dégagent de l'exploitation statistique, probabiliste, etc.

Ce paradigme est proposé et pose des questions. En particulier, deux problèmes se posent selon moi.

- Le problème de la donnée et de la mesure

Un sophisme s'est introduit soulignant qu'une donnée ne correspond qu'à une mesure insérée dans un ordinateur. Or, ceci se révèle totalement faux. Le problème de la mesure et de la donnée est bien connu. Pourquoi dire que les données sont « données » ? Le chercheur de données les trouvera parce qu'elles auront été construites par d'autres (récupération de logs d'un serveur) sans qu'il maîtrise leur mode de construction. Il ne sait pas comment elles sont produites ou se régulent les unes par rapport aux autres. Ces données sont des « construits » selon Bruno Latour, et même des « capta » comme le souligne Joanna Drucker.

Deuxièmement, les traitements statistiques s'appliqueront sur les données de manière totalement indépendante de la manière dont les données ont été obtenues à l'inverse d'un physicien menant sa thèse de physique au CERN (Organisation européenne pour la recherche nucléaire) qui observera des particules obéissant aux mêmes lois physiques que celles qu'il mobilisera pour les interpréter. Dans ce cas, les outils de captation et de mesures construits pour observer ces particules obéissent en effet aux mêmes lois physiques. La même compréhension du monde sert à expliquer la donnée, la manière dont elle est obtenue et interprétée.

La force et la faiblesse des *big data* est constituée par le fait que le traitement des données n'a aucun rapport avec l'obtention des données. Les algorithmes utilisés permettent seulement d'observer la récurrence statistique des données dans la même base sans avoir besoin de comprendre la manière dont les données sont construites. Ceci aboutit pour la première fois à une rupture radicale entre le traitement de la donnée et l'origine de la donnée. Un tel décalage n'a jamais existé en sciences expérimentales, définies par l'homogénéité de l'interprétation théorique de la donnée avec les conditions expérimentales de son obtention. Les outils expérimentaux sont des théories matérialisées (Bachelard), c'est-à-dire que les mêmes lois sont utilisées pour la construction d'un accélérateur de particules et pour analyser ce qui se déroule dans cet accélérateur. Or, dans le serveur d'Amazon, les statistiques sont totalement indépendantes des lois qui ont permis de le concevoir.

Cette rupture des *big data* constitue leur force parce que cela leur permet d'agréger des *data* hétérogènes – ce ne serait pas possible autrement, la science physique ne sait pas le faire. Le lien à l'origine de la donnée a été perdu mais cela n'importe pas puisque seule la récurrence statistique importe. Cependant des éléments tout à fait essentiels seront négligés. Ainsi, dans l'analyse de texte, un mot n'a pas le même sens, le même poids ou le même rôle sémantique dans un titre ou dans un paragraphe. Pourtant, le document sera transformé en un sac de mots par l'opération des *big data* et le même sens sera donné au mot, qu'il se trouve dans le titre ou dans le paragraphe. L'accueil des données de manière hétérogène nous apporte donc à la fois un instrument de comparabilité (comparer grâce à l'homogénéisation par le numérique des informations hétérogènes) et un nivellement de la donnée : l'information devient un sac informe de fragments de données élémentaires, incommensurables entre elles dans leur sémantique, ce qui n'est pas grave parce qu'elles sont ramenées à des informations binaires.

Tout ceci engendre un nouveau problème, l'enjeu des *big data* : comment passer de l'épistémologie de la mesure, héritage d'un temps ancien jusqu'en 1950, à l'épistémologie de la donnée qui permet de prendre au sérieux ce que les *big data* nous permettent de faire ? Mon propos ne vise pas à discréditer les *big data*, mais elles ne correspondent pas à ce qui nous est promis. Elles sont beaucoup plus importantes et constituent non seulement un nouvel outil, mais surtout un nouveau paradigme dans la mesure où l'interprétation des données n'est pas du tout homogène avec l'interprétation de nos mesures il y a quelques décennies.

Selon un sophisme répandu, les données ne seraient que des mesures massives (logs, mots dans les documents) qui sont rendues comparables. Ce n'est cependant pas tout à fait exact parce que ces données seront traitées aveuglément eu égard à leur origine, ce qui n'est jamais le cas avec les sciences de la nature.

- Que faire de ces données et le problème de la visualisation

On arrive au deuxième projet que je voulais aborder. Personne ne consulte les données elles-mêmes. Des algorithmes fonctionnent selon une lecture globale – Alain Giffard parle de lecture industrielle – : les outils lisent les informations pour nous et les reconstruisent à travers des visualisations difficiles à interpréter du fait de la difficulté de comprendre ce qui est visualisé. Les humains, animaux sémiotiques et compulsifs de l'interprétation, ne peuvent s'empêcher de vouloir y attribuer du sens sans savoir s'il est correct ou cohérent. Ainsi la carte de la « politicosphère » américaine produite par *Linkfluence* se révèle presque impossible à interpréter si ce n'est à retrouver des savoirs préalables : par exemple les sites républicains américains sont en lien avec les sites de vendeurs d'armes et les sites démocrates sont en lien avec les mutuelles de santé.

D'autres problèmes existent. Des cartes sont construites selon des principes esthétiques tout à fait différents des principes de la nature des données. Une carte de vulgarisation scientifique (observation d'un télescope spatial) représente une configuration d'énergie qui permet la naissance d'une étoile. Une photographie correspond en fait à des flux électromagnétiques captés par un télescope et transformés et colorisés. Un angle de vue saisissant a été choisi pour améliorer l'aspect esthétique de cette image. Cette image renvoie à l'imaginaire de la nouvelle frontière américaine. Le discours de Kennedy quant à la nouvelle frontière de l'espace se répercute dans les interprétants sémiotiques de la vulgarisation scientifique. Ce qui arrive à la vulgarisation scientifique de l'astronomie est analogue à ce qui peut arriver à la visualisation scientifique de données dans les *big data* : si de très belles choses sont montrées, elles ne correspondent pas forcément à ce qui est calculé et les interprétants sémiotiques ne coïncident pas forcément avec ceux qui gouvernent la compréhension des données.

Ceci se révèle raisonnablement problématique parce que la construction de ces visualisations à partir des données oublie le fait humain, ce qui nous permet de comprendre ce qui s'est passé et ce qui nous est raconté à travers les documents.

J'apporte une distinction entre le fait humain, soit le fait en tant qu'humain, c'est-à-dire l'histoire de nos communautés humaines, et le fait au sens de l'« événement » avec des moyens empiriques et logiques de les analyser, de les confronter, de les valider. Comme je l'ai évoqué tout à l'heure, le phénomène des *big data* reconfigure la manière traditionnelle et classique, humaniste de considérer et traiter le fait humain : le nominalisme de la donnée remplacera les outils habituels de la compréhension du fait humain. Ces derniers reposent sur le témoignage, la mise en indices, la mise en séries, des lois plus ou moins théoriques ; ces outils nous permettent d'interpréter le fait humain au niveau sociologique, historique, psychologique tel qu'il est arrivé à d'autres afin d'en avoir une compréhension interprétative, voire empathique : l'enjeu de la compréhension du fait humain est en effet de le comprendre comme étant « humain », c'est-à-dire ce qui peut arriver, ce qui est arrivé à d'autres humains, dans le passé (histoire) ou dans d'autres cultures (anthropologie). Pour reprendre le mot de Térence, « je suis un homme ; je considère que rien de ce qui est humain ne m'est étranger ». Or, ces faits humains ne sont plus considérés *via* la compréhension humaine que l'on peut en avoir, mais comme des faits mesurables et quantifiés, considérés *via* leur récurrence statistique totalement anonymisée par rapport au fait humain d'origine.

Ceci aboutit à un paradoxe. L'utilisation des *big data* interdit la compréhension du fait humain derrière les données collectées, en dehors de ceux déjà connus. Le paradoxe du *big data* se rapproche du paradoxe du Ménon dans lequel Socrate éduque un esclave et lui fait redécouvrir un théorème de géométrie en traçant des lignes dans le sable. Selon lui, cela signifie qu'il le connaissait d'une vie antérieure. Platon explique que la connaissance est la réminiscence, soit le fait de se souvenir. Il n'est pas possible de connaître quelque chose de nouveau parce que la reconnaissance d'un élément nouveau implique que cela était déjà connu et ce n'est pas reconnu si on ne le connaît pas. Les *big data* sont similaires : il n'est possible d'y reconnaître que ce qui était déjà connu, mais il se révèle très difficile d'avoir une connaissance nouvelle parce que les paradigmes, épistémologique et sémiotique, qu'elles imposent, impliquent que la nouveauté est indécelable.

En conclusion, deux nominalismes sont imposés par le numérique :

- le nominalisme de la variante, massif, mais non problématique en soi – c'est une réalité documentaire que les archivistes connaissent très bien depuis des années. La variante représente la réalité documentaire et pas l'essence documentaire dont elle serait la manifestation ; il faut travailler sur ce qui est montré, fruit d'une reconstruction dynamique utilisant des conventions et règles techniques, et non sur l'hypothétique ressource numérique sous-jacente qui n'a de sens qu'à travers de telles reconstructions ;

- le nominalisme de la donnée, plus récent et plus problématique, qui affecte le monde global de la culture et de la compréhension de nous-mêmes au-delà du monde du document, avec un effacement de la singularité au profit de la répétition statistique et l'enjeu d'une nouvelle épistémologie et d'une nouvelle sémiotique à concevoir. Elle doit être prise au sérieux parce que notre imbécillité épistémologique ne doit pas nous chasser du paradis que les *big data* pourraient construire pour nous.

Bruno BACHIMONT
Directeur de recherche
Université de Compiègne