Données, data, datamasse : production et conservation de données.

Bertrand Müller, directeur de recherche

CNRS-Centre Maurice Halbwachs
École Normale Supérieure
48 boulevard Jourdan
75014 PARIS

tel: + 33 1 43 13 62 13
bertrand.muller@ens.fr

(version de travail, sans références bibliographiques ni notes de bas de page)

Introduction

Je voudrais donner à mon intervention deux dimensions singulières. D'une part, une dimension prospective, je voudrais réfléchir ici non pas prioritairement à des pratiques ou des situations passées, mais à des situations en devenir qui peuvent faire écho à des éléments lointain mais qui se manifestent aujourd'hui comme des manifestations de recherches émergentes.
D'autre part, je voudrais donner à ma communication une tournure plutôt problématique. Identifier des situations qui me paraissent aujourd'hui au centre de déplacements dans la recherche susceptibles également d'établir de nouveaux rapports entre l'archive et la recherche.
Une remarque encore : je ne suis pas bibliothécaire, ni archiviste, ni spécialiste des sciences de l'information et de la communication, mais historien, pourtant ce n'est comme historien utilisant des archives ou des données que je vais intervenir ici, mais comme un historien qui réfléchit sur ce que signifie, dans des périodes données, produire des données, de la documentation, des archives. J'inscrit ma réflexion dans un cadre plus général que j'appelle une histoire longue des «régimes documentaires» depuis le Moyen âge jusqu'à nos jours.
Je vais aborder la question sous un angle scientifique large incluant les sciences humaines et sociales.

Le CNRS a publié dans son magazine international en janvier 2013, un dossier sur le thème qui est le vôtre aujourd'hui, dossier intitulé «The big data revolution». L'éditorial définissait ainsi le phénomène, je traduis de l'anglais :

« Big Data est une nouvelle discipline scientifique avec des défis sociétaux considérables (incluant le génomie, la santé, la sécurité globale, etc.) qui agit comme une force directrice pour la recherche. Dans ce sens, les données peuvent être considérées dans le cadre d e grandes infrastructures de partage facilitant la recherche au niveau national, européen et international. »

Ces structures européennes sont d'ailleurs en place dans le cadre du forum ESFRI, Forum stratégique sur les infrastructures de recherche. De son côté le CNRS a créé en mai 2012 une mission de 5 ans sur ces questions intitulé Mastodons. 16 projets (sur 37) ont été retenus et ont fait l'objet d'une première journée d'études en décembre 2012.
Pour autant, la production de données à une large échelle dépasse largement le cadre des activités scientifiques elles-mêmes. Ce phénomène, identifié récemment par la formule de Big Data, est en pleine expansion, il résulte de la transformation d'Internet, on parle désormais d'un web des données, de la croissance de l'économie numérique, de la prolifération des réseaux, la multiplication des échanges publics et privés. Aujourd'hui les data, les données sont partout, produites, diffusées,

vendues et consommées comme des biens de consommations, elles sont devenues un enjeu majeur du développement économique, de la politique de sécurité, comme des nouvelle formes de la guerre moderne. On sait par exemple le rôle joué par l'analyse des big data dans la campagne électorale de Barak Obama en 2012. Récemment, dans son édition du 15 janvier 2105, le Monde posait ainsi la question : Le big data profite-t-il vraiment au consommateur ?

Récent, le phénomène est puissant, envahissant, rapide, protéiforme… c'est de tout cela d'ailleurs dont il se nourrit. Démesuré, il échappe encore à une appréhension rationnelle. Il donne le vertige ou laisse indifférent ou incrédule… Sommes-nous confronté à une nouvelle révolution, après la révolution numérique ? Le phénomène BigData n'est-il qu'un phénomème éphémère, condamné à l'obsolescence pré-programmée des destinées numériques comme pourrait l'être le Cloud computing ? Est-ce une simple étape de plus dans la révolution numérique ? Est-ce une révolution numérique qui annonce aussi d'autres révolutions notamment scientifique ?
Questions bien trop ambitieuses pour être même abordées simplement ! Questions surtout trop importantes pour ne pas être abordées dans leur dimension historique. Il est en effet temps, même si le phénomène paraît encore nouveau, d'historiciser les big data, de penser non seulement les formes d'une irruption, mais aussi les moments antérieurs d'une histoire pas nécessairement continue.

«Historiciser les Big Data procure une étendue comparative et une profondeur historique à la discussion actuelle sur le potentiel révolutionnaire des modes de production des connaissances fondés sur des données massives et les défis sociétaux posés par le déluge des données.»

J'emprunte cette citation approximativement traduite de l'anglais à l'énoncé d'un programme d'un groupe de recherche lancé par le Max Planck Institut für Wissenschaftsgeschichte, mais je ne suivrai pas complètement ces propositions.

Je vais donc ici me limiter à deux séries de questions concernant des problèmes sémantiques ou lexicologiques dans un premier temps, puis dans un deuxième temps, je voudrais interroger certaines des reformulations des liens entre les données et leur conservation, leur mise en archive. Enfin je terminerai par des considérations sur les liens entre big data et la transformation des sciences.

I. Questions sémantiques et lexicologiques

Il me faut commencer par mettre en place quelques éléments sémantiques. Ce préalable me paraît important notamment concernant deux notions assez problématiques aujourd'hui : celle de **données**, d'une part, celle **d'archive** de l'autre.

Ces deux notions paraissent avoir désormais un destin lié bien que ce lien soit à la fois paradoxal et récent. Paradoxal, parce que ces deux notions ont envahi l'espace sémantique du web ; elles sont l'une et l'autre omniprésentes. En même temps, ces deux notions désignent aussi des choses ou des phénomènes différents entièrement redéfinies par les pratiques numériques.
Si les enjeux prioritaires des données, des data et des big data sont d'abord la curation et la conservation, les archivistes sont aujourd'hui confrontés à la complexité de la mise en archive de masses non structurées, protéiformes et hétérogènes de données qui ne sont pas des fonds d'archives ni des collections documentaires. D'ailleurs, les enjeux de la conservation durable des données concernent désormais autant l'institution des archives, que les bibliothèques ou les centres de documentation.

A) Donnée, data

D'abord les données. Notion à la fois ancienne et très nouvelle. Quelques rappels étymologiques

pour commencer, puis des éléments informatiques et enfin une définition scientifique.

Le mot «donnée» apparaît vers 1200, sa signification est associée au verbe «donner» et désigne une distribution de biens ou d'argent, une aumône. Ce n'est qu'au milieu du XVIIIe siècle, que le mot prend une signification nouvelle, notamment dans le langage mathématique : données se réfère alors, selon l'Encyclopédie méthodique, à

«certaines choses ou quantités, qu'on suppose être données ou connues, & dont on se sert pour en trouver d'autres qui sont inconnues, & que l'on cherche. Un problème ou une question renferme en général deux sortes de grandeurs, les données et les cherchées, data & quaesita.»

Utilisé également dans d'autres domaines, les arts, la philosophie, la médecine, etc., le mot désigne des «choses que l'on prend pour **accordées** sans avoir de preuves immédiates de leur certitude».

Dans une perspective plus constructiviste, les données ne sont pas simplement données mais bien produites par un ensemble d'opérations spécifiques de définition, de collecte, de classement, etc… opérations qui ont fait dire au sociologue Bruno Latour qu'il conviendrait de parler d'«obtenues» plutôt que de «données». Les «cherchées» ou les «requises» évoquées dans l'Encyclopédies me paraissent également convenables !

Ces «choses accordées» relèvent toujours d'une mise en forme, d'une catégorisation, d'une structuration déterminées. En d'autres termes, les données ne sont pas des signes d'une nature muette mais elles sont des éléments déjà signifiants et organisés, disponibles pour des processus de connaissances possibles. Les données sont donc liées à des configurations cognitives.

Comme le suggère Jean-Michel Berthelot, les données sont à la fois «toujours déjà structurées et multiplement structurables» (les documents pour un archiviste ou les archives pour un historien), leur description et leur explication relèvent d'un jeu tensionnel entre des programmes différents» (analyse scientifique versus description archivistique).

Dès lors, il est nécessaire de distinguer plusieurs niveaux, si l'on s'en tient provisoirement à la démarche scientifique en sciences humaines et sociales.

Le niveau des **data**, c'est-à-dire les matériaux recueillis à partir des traces ou construits par le chercheur (questionnaire, entretien, expérimentation, simulation, etc). Ces data sont des intermédiaires déjà structurés et souvent commentés : distributions statistiques, règlements, témoignages, registres, comptabilités, correspondances, etc…

Celui ensuite des **événements** plutôt que des choses, c'est-à-dire de l'arrière-fonds sur lequel s'inscrivent des agrégats d'action ou des effets d'action.

Celui enfin des faits, c'est-à-dire des «objets stabilisés d'analyse et d'explication» qui procèdent précisément d'une opération de construction. Cette reformulation de la problématique des données que j'emprunte à Jean-Michel Berthelot me permet aussi de rappeler que **données, documents et archives** sont également le produit d'une chaîne de constructions successives ou simultanées mais aussi de constructions spécifiques, en tension entre des domaines de traitement liés mais distincts (archivistes, bibliothécaires, documentalistes, chercheurs, informaticiens, historiens pour ce qui nous concernent).

Dans cette perspective, non seulement les données sont construites, mais elles sont également chargées de **signification** ou **redéfinies** en fonction de nouveaux usages, mais elles n'ont pas de signification pour elle-mêmes hors des contextes auxquels elles sont associées.

En revanche en informatique, la notion de **donnée**, liée à celle **d'information** à laquelle elle s'oppose, n'a pas de signification : elle est en quelque sorte l'élément qui donne une forme à une information, laquelle est disjointe de sa dimension sémantique.

La théorie de l'information s'est définie sur l'exclusion de la sémantique et dans l'indifférence à la signification du message. Le sens est produit par la ré-articulation d'un ensemble de données et de méta-données.

Cette distinction qui fondait en fait une théorie mathématique de la communication et de l'information, préoccupée d'abord par la quantité et la qualité des données transmises, reposait ainsi sur la séparation stricte des données elles-mêmes et de leur traitement. Cette opération indispensable au développement indépendant des applications informatiques est illégitime du point de vue de la science. Cette double opération de dissociation et de réarticulation est au principe même de la naissance de l'informatique dans les opérations de codage, transmission et décodage des informations secrètes pendant la Deuxième Guerre mondiale.

B) Données, méta-données, records

C'est moins l'émergence de l'informatique que le développement d'Internet dans les années 1990 qui a rebrassé les cartes. En effet, les «systèmes d'information» que sont les archives, les bibliothèques ou les médiathèques, avaient une longue pratique de la codification, du signalement et du contenu des documents placés sous leur circonscription. Mais ici la codification n'a pas le sens de codage.
Les descriptions documentaires ou les notices bibliographiques, d'abord consignées sur des fiches, avaient été informatisées et normalisées et consacraient également une dissociation entre les notices et les documents mais essentiellement à de des fins de classement, de catalogage et d'inventaire.

Il me semble que l'élaboration de nouveaux langages de balisage (d'abord GML, puis SGML et HTML) a provoqué une rupture en réintroduisant ces informations référentielles dans les documents eux-mêmes sous forme de **méta-données**.
Les méta-données, si je ne me trompe pas, sont des données signifiantes sur des données qui ont fait l'objet de normalisation (par exemple le Dublin Core). Elles sont susceptibles d'offrir l'accès au contenu informationnel d'une ressource numérique et de générer également de nouvelles formes de données. Les méta-données sont des éléments essentiels du développement du WEB en particulier du WEB dit sémantique ou encore du WEB des données.

Leur emballage «technique» oblitère en fait des enjeux politiques, économiques et cognitifs considérables puisque désormais, dans l'économie des connaissances et du savoir qui est la nôtre, les méta-données sont les clés d'accès aux ressources numériques c'est-à-dire aux données, à l'information, à la connaissance et à de nouvelles connaissances par leur articulation et par les agencements nouveaux que permettent l'association des méta-données.

Empruntée au langage de la gestion et de l'information, la «donnée» est définie comme la «plus petite représentation conventionnelle et fondamentale d'une information (fait, notion, objet, non propre, chiffre, statistique, etc.) sous une forme analogique ou digitale permettant d'en effectuer le traitement manuel ou automatique (informatique).» [Ibidem]

A cet égard, les données et plus encore les méta-données qui doivent être conservées absolument sont moins considérées comme des éléments à archiver mais comme des informations désormais indispensables à la mise en archive.

Les méta-données sont des éléments indispensables à la traçabilité d'un document et à l'administration de la preuve qui sont au coeur de la pratique des archivistes mais aussi des systèmes

d'informations.

Les redistributions produites par la conversion numérique ont également brouillé les définitions. C'est le cas entre **données** et **data**, mais c'est aussi le cas pour le terme «**archive**» qui désormais a également envahit le WEB.
Central, mis au singulier, l'archive plus que les archives, le mot a été amputé d'une partie de sa signification et de ses implications. L'archive aujourd'hui désigne plus souvent une opération de sauvegarde ou de conservation plus ou moins durable. Convention qui oblitère évidemment toute la signification et la complexité de la mise en archive.

Paradoxalement, c'est aujourd'hui plus le terme anglo-saxon de «record» lui-même pluralisé (records) qui semble redonner du sens à l'opération archivistique. Le terme désigne les documents sous une double dimension :
leur version définitive et **leur valeur de preuve**, qui se distingue de la notion de document assimilé à son contenu informatif ou encore au document d'archive retenu pour sa dimension historique. La norme qui définit l'organisation du Records management, s'applique à tout type de documents indépendamment de leur support dans le cadre de la gestion courante mais aussi intermédiaire des documents et de leur transfert aux archives mais elle ne concerne pas l'archivage historique.

Le principe du Records management repose sur la notion de records, notion difficile à traduire en français en particulier parce qu'elle convient mal à la tradition française qui retient parfois la notion de «document engageant», en particulier dans le cadre de l'archivistique entrepreneuriale qui a fait sienne d'ailleurs les principes de la gestion documentaire et du records management.

Un document engageant, c'est un document stabilisé, authentifié et validé comme un document définitif, qui conservera précisément sa valeur probante ; il est donc le résultat d'une série d'opérations qui le distingue d'une donnée ou d'un simple document.

II. Archives des sciences et archives scientifiques

Je vous propose encore un détour par les archives des sciences avant d'en revenir aux données scientifiques.
La formule **d'archives des sciences** préférée parfois à celle plus ambiguë **d'archives scientifiques** définit un périmètre d'intervention des archivistes par rapport à l'activité scientifique, privilégiant ainsi des catégories documentaires qui témoignent prioritairement de l'institution scientifique plus que de la recherche elle-même et plus que de la production scientifique notamment la production de données.
Il est vain de refaire le procès de l'impuissance des Archives à préserver ainsi la globalité de l'activité scientifique au seul bénéfice de l'administration de la science. Une interrogation plus historique sur les transformations de l'activité scientifique, de ses lieux de mémoire et d'archives, d'une part, l'examen des mutations également de ce que j'appelle des «régimes documentaires» permettrait sans doute de mieux comprendre comment et pourquoi l'activité scientifique a échappé aux prises de l'institution archivistique.
Je n'ai pas le loisir de développer ici cet argument qu'il fallait pourtant rappeler car il est lié à la problématique de mon intervention. Pour faire court, je dirai que l'un des points obscurs de l'oubli ou de **l'oblitération des archives a été précisément l'archivages des données scientifiques.**

Je voudrais cependant un peu complexifier mon propos en me référant à deux exemples qui sont peut-être aussi des contre-exemple.

A) UK Data Archive

En premier lieu, je voudrais rappeler que l'expression d'archives des données exprime bien une réalité ancienne, puisqu'elle a émergé dans les milieux universitaires britanniques dès les années 1960. Ceux-ci en effet, bien avant le web des données, étaient soucieux d'élaborer des dispositifs qui favoriseraient le partage et la conservation des données issues des grandes enquêtes économiques et sociales. Il s'agissait alors – déjà – d'améliorer la diffusion (médiocre) des données entre chercheurs et de parer à l'exportation de données vendues à des institutions américaines.

Il en est ressortit notamment la création d'une institution financée conjointement par la London School of Economics (LSE), le Social Research Council, l'industrie et le gouvernement. Cette institution a entrepris un inventaire et une mise en archives des enquêtes susceptibles de fournir des données pour des analyses secondaires et longitudinales. Au début du XXIe siècle l'institution adopta le titre qui est encore le sien de UK Data archive qui a développé une section historique (History Data Unit) intégrée au sein des Archives publiques.

L'un des principes majeurs du partage des données réside dans les conditions qui assurent à une donnée sa compréhension et son interprétation par un utilisateur quelconque. Ce qui exige une description claire et précise des données ainsi que des informations sur le contexte de leur production, soit une bonne documentation et un ensemble de méta-données rigoureux.

Une donnée isolée, c'est-à-dire coupée de son contexte de production mais aussi de son contexte de validité est une donnée in-signifiante. Documenter une donnée revient à noter comment elle a été conçue, ce qu'elle signifie, et préciser son contenu et sa structure. La documentation est également décisive pour assurer une longue conservation des données. Une bonne documentation comprend des informations sur plusieurs éléments : le contexte de la collection de données (historique du projet, ses objectifs, ses hypothèses) ; les méthodes de collecte des données ; la structure des données et les liens entre elles la validation des données, les procédures de test, de preuves, etc. ; les éventuels changements dans le temps ; des informations enfin sur l'accès et les conditions d'usage ou de confidentialité.

Les métadonnées contiennent des éléments de documentation qui ont une signification et un but précis. Ces sont des éléments désormais standardisés qui décrivent l'origine, le but, la référence temporelle, la localisation géographique, le créateur, les conditions d'accès et d'utilisation des données. Les métadonnées sont donc utilisées comme des ressources pour la recherche de données ou comme des références pour les citations.
Ces métadonnées sont généralement structurées et standardisées selon des normes internationalement reconnues (Dublin Core, ISO19115, Data Documentation Initiative (DDI), Metadata Encoding and Transmission Standard (METS) and General International Standard Archival Description (ISAD).

B) Open Archive Information System (OAIS)
Le deuxième exemple est désormais plus connu, et il me sera possible de l'évoquer brièvement. Confronté au double problème de la diffusion et de la conservation de données, la recherche et l'industrie astronautiques ont mis en place un protocole de gestion des données qui a abouti à l'élaboration d'un modèle conceptuel pour l'archivage de données numériques. Ce modèle est l'OAIS qui est aujourd'hui une norme ISO, il est censé répondre à deux problèmes :

1) préserver des données normalisées pour leur diffusion,
2) élaborer un dispositif d'archives qui correspond une communauté d'utilisateurs et s'efforce de répondre aux besoins de cette communauté, en l'occurrence ici la communauté scientifique et industrielle.

Un dispositif OAIS garantit aux utilisateurs des données compréhensibles, accessibles et

disponibles dans la durée. L'OAIS est donc devenu à la fois une norme et un **cadre conceptuel** sur lesquels s'appuient désormais les politiques d'archivage en particulier dans les grandes institutions scientifiques (Le CINES, la BnF, Huma-Num ont développé des plates-formes conformes à l'OAIS).

C'est aussi un cadre de référence pour un ensemble d'outils normalisés, un standard pour l'échange des données pour l'archivage et pour la plupart des plateformes d'archivage électronique récentes.
L'OAIS articule deux éléments : un système d'information et d'échanges des données ; un dispositif d'archivage, définis par rapport à une communauté cible : les utilisateurs.

Ce sont donc ici les usages et l'interprétation des contenus qui entrainent la conservation des données. Dans un univers numérique dans lequel l'archivage n'est plus un stock mais un flux, la conservation des contenus est illusoire mais surtout inutile.

Désormais, on peut dire avec B. Bachimont que «c'est l'appétence au contenu, entretenue par l'animation et l'alimentation des communautés savants et culturelles, qui suscite la succession des lectures et des commentaires comblant le fossé d'intelligibilité et préparant la lecture des générations futures».

La pérennisation de l'archive est ainsi assurée par un double processus :

1) la conservation d'une lisibilité technique qui concerne notamment les formats, la qualité des méta-données, mais aussi

2) la conservation d'une lisibilité intellectuelle ou scientifique qui dépend de l'engagement des communautés d'utilisateurs. Ainsi de la même manière que des fichiers simplement sauvegardés ne résisteront pas à l'obsolescence techniques, les contenus qui ne seront plus sollicités seront également condamnés à l'obsolescence.

Ces contraintes nouvelles impliquent donc plus activement les communautés concernées par la mise en archive et en premier lieu, en collaboration avec les archivistes, la communauté savante et culturelle. Ces contraintes nouvelles générées par les impératifs de la numérisation mais aussi par les exigences de la recherche et du partage des données introduisent une plus grande plasticité dans la conservation des données et des contenus qui est assurée par la transformation de leur ressources. L'archive ne peut plus se penser à partir du seul point de vue de son origine (le principe de provenance) mais en fonction de l'exploitation qui en sera faite. Ne devrait-on pas dès lors parler de l'archive comme une réinvention permanente à partir des ressources conservées mais aussi transformées, enrichies, réinterprétées ?

III. Les données à l'épreuve de la datamasse (plutôt que Big Data et par analogie avec biomasse)

L'émergence récente et invasive de la notion de Big Data change-t-elle la donne dans ces transformations ? Il existe une loi – loi de Moore – qui prédit que les capacités de stockage numérique double tous les 18 mois, cette loi pourrait s'appliquer également à la production exponentielle de données qui est encore plus forte et plus rapide. Le phénomène a pris des proportions quantitatives spectaculaires et exige non seulement de nouveaux moyens de représentation et de diffusion des données. Le phénomène identifié dès le début des années 2000 avait été caractérisé par la règle dite des 3v (volume, vélocité, variété).

Pourtant le phénomène n'est pas nouveau, l'histoire de l'humanité est balisée par des moments de très forte croissance d'information et de documentation, mais l'évolution du WEB documentaire

vers le WEB sémantique et le WEB des données constitue non seulement un changement spectaculaire d'échelle, il transforme également notre rapport aux données.

La puissance des moyens de calcul, la précision des instruments de mesure, la capacité des dispositifs d'enregistrement bouleversent non seulement le monde de l'information mais aussi l'économie, les sciences, ainsi que les modes de gouvernement. Certes, cette expression de Big Data est encore aussi nébuleuse que celle du Cloud computing, on peut déjà cependant mesurer quelques unes des transformations majeures qui sont à l'oeuvre hors le constat de la croissance et de l'accélération en cours, en particulier dans la pratique scientifique qui change également d'échelle.

On parle désormais également de Big Science. Les deux phénomènes étroitement liés redéfinissent en profondeur les modèles organisationnels, économiques, méthodologiques des sciences et annoncent l'émergence de nouveaux paradigmes scientifiques.

Dans le monde des sciences de l'information, certains auteurs parlent de la formation d'un quatrième paradigme centré sur l'analyse de très gros volumes et flux de données collectées automatiquement par des dispositifs informatiques et techniques très complexes comme le synchroton par exemple.

Les expériences du Large Hadron Collider représentent environ 150 millions de capteurs délivrant des données 40 millions de fois par seconde. Il y a autour de 600 millions de collisions par seconde, et après filtrage, il reste 100 collisions d’intérêt par seconde. (Wikipedia)

De la règle des 3v, je retiens un terme : celui de variété. Plus encore que la quantité et la vélocité, c'est la variété eet partant la complexité des données enregistrées qui fait aujourd'hui bouger les repères traditionnels.

La datamasse n'est pas ou plus composée de données relationnelles, structurées, d'obtenues, de cherchées ou de requises, nous ne sommes plus confrontés à des «choses accordées», mais à des masses de données brutes, peu structurées voire non structurées. Ce sont des données complexes provenant du web (Web Mining), dans des formats divers (texte, images, son). Elles peuvent être publiques (Open Data, Web des données) ou privées. Ces données surtout qui posent de nouveaux problèmes qui dépassent les problématiques habituelles des bases de données, elles exigent de nouvelles analyses, de nouveaux protocoles d'interprétation.

Nous assistons peut-être à une inversion du processus de la découverte scientifique. Au modèle théorico-déductif qui caractérise les sciences expérimentales, fondé sur des théories, des jeux d'hypothèses ou des modèles interprétatifs soumis au test et à l'épreuve des données empiriques, succède un modèle qui s'efforce de découvrir dans les masses et les flux de données parfois très hétérogènes des relations, des corrélations, des régularités significatives, imperceptibles et imprévisibles.
L a dimension exploratoire de la fouille de données fait que les scientifiques ne savent pas nécessairement ce qu’ils cherchent. (Wikipedia)
De nouveaux outils d'analyse sont nécessaires : la statistique descriptive fait place à une statistique inférentielle, mieux adaptée à la lecture inductive et non plus déductive de données à plus faible densité d'information.

Dans ce nouveau paradigme, la donnée n'est plus la preuve mais elle est le matériau potentiel de la découverte scientifique. Ce paradigme se caractérise par l'accumulation de très gros volumes de données qui circulent sur le WEB, qui s'échangent entre de puissants serveurs, de données interconnectées, qui génèrent ainsi de nouvelles données et méta-données, mettent en oeuvre des scénarios analytiques complexes qui cherchent à découvrir les systèmes de relation qui les structurent.

La gestion des données scientifiques a fait émerger une nouvelle science des données (data science) qui mobilise les compétences multiples de mathématiciens, statisticiens, informaticiens, mais aussi de designers et de spécialistes de la visualisation des données.

De nouvelles divisions du travail s'établissent entre la collecte des données, leur gestion, et leur analyse et interprétation différée dans le temps, mais de nouvelles formes de collaboration et de mise en commun de compétences multiples sont désormais nécessaire pour assurer l'élaboration de nouveaux modèles d'édition, de diffusion et de préservation des données.

Autant que les modèles analytiques complexes proposés pour traiter les données, ce sont les modèles informatiques qui assurent la conservation, la lisibilité, l'accessibilité et l'intelligibilité de ces données qui deviennent des enjeux de la recherche scientifique elle-même.

En d'autres termes, dès lors que le travail scientifique est assujetti à la conservation des données, ce sont les archives numériques qui deviennent enjeu et objet de recherche.

Les big data remettent en cause le processus même de la mise en archive qui ne se découpe plus en phase (les cycles de vie) mais devient un processus continu de données lesquelles peuvent coexister durablement dans des bases de données constamment réactualisées ou alimentées. En effet, conserver des données numériques aujourd'hui ce n'est plus les mettre à part pour les stocker, c'est à l'inverse les laisser circuler et les faire migrer constamment d'un serveur à l'autre.
Les données ne sont plus stockées sur un seul serveur. Les données, les fichiers sont "découpés" en morceaux et chaque morceau est envoyé sur un serveur local précis.

Le processus d'archivage est ainsi transformé. L'archive n'est plus un reste, une trace, un vestige, d'une connaissance déjà formalisée ailleurs, elle ne suit plus la connaissance mais elle la précède, les données sont produites et mises en archives indépendamment et antérieurement aux grilles d'interprétation.

Epilogue
C'est ainsi – pour conclure – tout à la fois la conception de l'archive et des données qui sont à redéfinir. La «donnée» qui induit une stabilité, un déjà-là, qu'il suffirait de faire parler, de calculer est tout aussi obsolète que l'idée d'une archive comme trace, relique, voire témoignage ou preuve.
Plus que jamais, la valeur des données dépend aujourd'hui des usages actuels ou venir et leur préservation ne dépend pas seulement de la qualité de leur conservation mais surtout de la continuité des interprétations qui en sont et seront faites.

Un nouvel impératif s'impose ainsi aux archivistes qui n'est plus d'entreprendre la mise en archives illusoire de flux de données instables et hétérogènes mais d'assurer la conservation des systèmes d'intelligibilité et d'interprétation qui leur donne sens à un moment donné.
L'archivage devient alors un processus dynamique qui ne se contente pas de fixer des contenus, mais qui assure la préservation des usages et des interprétations qui transforment et enrichissent ces contenus, produisant ainsi de nouvelles données.