Et demain ? Archivage et big data

Marie-Odile Charaudeau, Alexis Fritel, Charles Huot, Philippe Martin, Laurent Prével

Et demain ? Archivage et *big data*

Marie-Odile Charaudeau, Alexis Fritel, Charles Huot, Philippe Martin, Laurent Prével

Et demain ? Archivage et *big data*

Marie-Odile CHARAUDEAU
Alexis FRITEL
Charles HUOT
Philippe MARTIN
Laurent PRÉVEL

« Nous avons mis une grande partie de notre mémoire
et de nos cerveaux dans l'ordinateur et maintenant,
il s'agit d'innover, de créer, et d'inventer autrement »
Michel Serres

La présente réflexion s'intéresse, à l'instar de VITAM (Valeurs immatérielles transmises aux archives pour mémoire – futur socle d'archivage électronique pour les données produites par l'ensemble des administrations[1]), à la problématique posée par la confrontation entre les principes fondamentaux de l'archivage et la révolution des usages soulevée par les *big data*.

Contexte et approche proposée des *big data*

3D, réalité augmentée, tout tactile, géolocalisation, *big data*… autant de nouveautés qui paraîtront évidentes dans une vingtaine d'années, mais qui

1 http://www.modernisation.gouv.fr/ladministration-change-avec-le-numerique/par-son-systeme-dinformation/vitam-vers-un-socle-d-archivage-electronique-commun-toute-l-administration

demandent aux entreprises et aux organisations un véritable effort d'adaptation. L'amplitude de choc, les secteurs impactés, les *business models* traditionnels concurrencés et toutes les applications qui en découleront sont, à ce jour, loin d'être identifiés. Le numérique dévore le monde et, octet après octet, le transforme en profondeur. Les administrations, entreprises, organisations qui ne prennent pas la mesure de cette révolution se fragiliseront inéluctablement. Dans le cas contraire elles gagneront en efficacité et en compétitivité.

Du document aux data

La première vague de la révolution numérique a d'abord introduit les documents électroniques dans les entreprises et les organisations, apportant d'importants gains de productivité et de baisse des coûts de gestion[1].

La seconde vague, celle d'Internet, a bouleversé le fonctionnement des entreprises avec le déploiement des portails et des blogs, de l'e-commerce, des réseaux sociaux, de la mobilité, etc. Du fait de leur importance stratégique, ces flux et contenus sont considérés comme des actifs immatériels et sont de plus en plus souvent évalués financièrement par les entreprises.

Nous voici entrés dans l'ère des *big data*, troisième vague de la révolution numérique. Ces ensembles de données nécessitent de nouveaux outils techniques pour les comprendre et en tirer du sens, et posent des questions profondes sur leur collecte, leur interprétation et leur analyse. Entre les données produites par les internautes, les entreprises et celles mises à disposition par les collectivités, il existe une masse colossale de données non-exploitées, potentiellement génératrices d'énormes gains de compétitivité pour les entreprises et les organisations qui sauront se transformer en *data-driven*.

[1] Outre l'étude « Une nouvelle économie ? Transformation du rôle de l'innovation et des technologies de l'information dans la croissance », publiée par l'OCDE en 2000 (OECD Publishing, 104 pages) il existe de nombreux articles plus récents concernant la dématérialisation et contenant des chiffres sur les gains de productivité et la baisse des coûts de gestion obtenus dans le cadre des projets de dématérialisation, notamment dans les études publiées par *Markess International (http://www.markess.com/etudes-detaillees)* dont « Atouts et bénéfices de la dématérialisation de documents RH » du 9 mars 2011, « Enjeux légaux et réduction des coûts poussent les projets de dématérialisation » du 5 novembre 2011, et « Infographie - Optimisation des processus documentaires » du 5 décembre 2012. En outre, le baromètre 2014 publié par CXP reprend ces mêmes notions. http://www.cxp.fr/content/barometre-cxp-2014-optimisation-des-processus-clients-par-la-dematerialisation-des-flux

La quantité de *data* va encore s'accroître avec l'Internet des objets : il y aura bientôt plus d'objets que de personnes reliés à Internet. L'univers numérique double tous les deux ans. Il pose la complexité d'un nouveau monde : dès 2020, il pourrait compter jusqu'à 80 milliards d'objets connectés et peser 44 000 milliards de gigaoctets, soit dix fois plus qu'à présent !

La maîtrise et la valorisation

Dans ce contexte, la maîtrise et la valorisation des données sont les deux composantes essentielles et indissociables de toute stratégie numérique.

Tant pour des raisons opérationnelles que réglementaires, les flux documentaires et leurs contenus informationnels doivent être pris en compte, gérés et conservés de façon rigoureuse et fiable, ce qui est l'objectif prioritaire de la maîtrise des contenus au travers de la dématérialisation, l'archivage, la gestion électronique des documents (GED), la gouvernance de l'information.

Simultanément, l'exploitation de ces mêmes données et contenus afin d'optimiser les procédures, de détecter des tendances de marché, de permettre des analyses fines d'activités, etc. constitue un des fondements de la valorisation des contenus. Pour ce faire, les technologies utilisées sont celles du *content analytics*, du sémantique, du prédictif, de la datavisualisation, etc.

La nécessaire transformation numérique des entreprises

« Tout le monde commence à craindre de se faire "Uberiser" », a récemment déclaré Maurice Lévy, le patron de Publicis, au *Financial Times* [1]. Au travers de cette phrase, il exprime bien la prise de conscience que la transformation numérique est une vraie révolution industrielle touchant tous les secteurs d'activité.

Les entreprises et les organisations n'ont plus le choix aujourd'hui : elles doivent engager leur transformation numérique. Le passage du « non-connecté » au « tout-connecté » est une réalité économique et technique qui s'impose à tous. En France, les entreprises et les organisations sont longtemps restées conservatrices, mais un nombre croissant d'acteurs économiques est désormais conscient des enjeux et prêt à accélérer leur transformation numérique. Plus de la moitié des dirigeants français estiment que le numérique entraînera des changements économiques dans leur secteur et 46 % craignent de perdre des clients au bénéfice de nouveaux acteurs si leur entreprise ne

[1] Décembre 2014, repris par La Tribune : http://www.latribune.fr/technos-medias/20141217tribd1e82ceae/tout-le-monde-a-peur-de-se-faire-uberiser-maurice-levy.html

prend pas le virage numérique dans les douze mois à venir[1]. Anticiper la révolution numérique est une chose, savoir adapter son modèle économique en est une autre. La transformation numérique ne consiste pas tant à numériser ses processus qu'à inventer de nouveaux métiers, de nouveaux services et être « obsédé » par ses clients.

L'enjeu consiste à détecter les mutations et les compétences nécessaires pour les cristalliser. Il ne s'agit pas de prédire le futur, mais d'extraire le sens des tendances observables. La difficulté consiste à rester sans cesse à l'écoute des changements sans confondre les simples bruits, sans conséquence sur la stratégie, avec un vrai signal, précurseur du changement puissance dix. L'atout de l'incertitude est de favoriser l'innovation et la création de valeur.

Innover consiste à se projeter dans l'avenir en quittant sa zone de confort et en acceptant le changement de paradigme.

Enjeux du *big data*

Comme le rappelle l'ensemble des articles consacrés au *big data*, celui-ci se définit par la règle des trois V qui en décrit les dimensions : le volume, la variété et la vélocité.

Le volume correspond à la quantité considérable de données à prendre en compte. Cette quantité n'est pas absolue mais relative à un contexte donné. Par exemple, une commune rurale aura une vision différente du *big data* relatif aux données sur ses administrés de celle d'un site de vente en ligne international. Bien souvent, on illustre le volume du *big data* en termes de stockage physique : teraoctet (To), petaoctet (Po), exaoctet (Eo) ou zettaoctet (Zo). Mais une telle mesure est un raccourci qui reflète mal la quantité de données à prendre effectivement en compte. Prenons l'exemple d'un site de partage en ligne de photos et de vidéos. Pour celui-ci, les contenus publiés représentent à la fois une quantité de stockage considérable, mais une quantité informationnelle réduite. En revanche, les données relatives à ces fichiers, c'est-à-dire les métadonnées descriptives, les données relatives aux transactions de publication, voire les données produites par des traitements d'analyse numérique faits *a posteriori* sur les contenus, représentent quant à elles un volume de données individuelles considérable pour un volume de stockage beaucoup plus modeste.

[1] http://www.accenture.com/fr-fr/Pages/insight-europes-growth-digital-opportunities-competitiveness-growth.aspx

La variété des données intègre leur forme (structurée, non structurée ou mixte), mais aussi dans le temps puisque de nouveau flux apparaissent, certains disparaissent et beaucoup évoluent au fil du temps. Cette variété des données est la dimension majeure du *big data*. On va chercher à élargir le périmètre des flux afin d'augmenter l'assiette de données à interpréter. On va également faire varier ce périmètre dans le temps et en exclure les sources dont la fiabilité n'est pas avérée.

La vélocité est celle de la fréquence à laquelle les données sont générées, collectées et consommées. Avec le *big data*, on vise un délai de traitement fortement réduit en comparaison de la situation existante. Dans certains cas, il s'agira d'un traitement au fil de l'eau, mais ce n'est aucunement une condition nécessaire. Par exemple, la détection d'épidémies par analyse des flux des réseaux sociaux ne répondra pas à la même exigence de temps que celle relative à la gestion dynamique du trafic routier en ville.

Des objectifs stratégiques

Par le biais de la description des dimensions du *big data*, en aval se trouve l'analyse des données. Et c'est bien le domaine de l'analytique qui vit une transformation majeure.

Mais avant tout, sans prendre en compte les besoins propres aux administrations ou aux organismes de recherche, pour quelles raisons les entreprises analysent-elles leurs données ? On peut distinguer trois objectifs.

Optimiser la façon de produire et de travailler est le premier objectif. Bien souvent, il s'agit d'une optimisation économique relative aux coûts ou aux prix. L'illustration d'une telle optimisation est celle de la tarification différenciée. Initialement pratiquée dans le domaine du transport aérien, elle régit aujourd'hui le prix de votre billet de train, de votre kilowattheure ou de votre course de taxi commandée en ligne.

Identifier des risques est le deuxième objectif. Là encore, il s'agit bien souvent de risques financiers, tels que la fraude, avec une forte exigence de réactivité, ou le taux d'attrition, c'est-à-dire le risque de perte d'un client par résiliation, passage à la concurrence ou changement d'offre. Mais, il s'agit également des risques de non-conformité pour lesquels le nombre et la complexité croissante des réglementations impactent son évaluation par la quantité de données à prendre en compte et la variété des règles à appliquer.

Le troisième et dernier objectif est de se créer de nouvelles opportunités commerciales, par exemple par stratégies de montée en gamme ou de vente additionnelle, mais aussi par l'analyse marketing des tendances du marché.

L'analytique à plusieurs facettes

Aujourd'hui l'analyse des données se fait selon deux approches différentes : l'informatique décisionnelle et la science des données.

L'informatique décisionnelle vise à produire des tableaux de bord, des indicateurs de performances ou des rapports, depuis des données passées. Elle s'appuie sur des données structurées et maîtrisées qui sont collectées et copiées depuis les systèmes internes de l'entreprise, et qui, après de nombreux prétraitements et mises en forme, sont regroupées selon des modèles de données prédéfinis dans des entrepôts de données. Lors de leur collecte, les données peuvent être expurgées des informations personnelles ou confidentielles. Les traitements d'analyse se font ensuite en mémoire, sur des sous-ensembles de données, à l'aide de la technologie OLAP[1]. Développée depuis les années 1960, l'informatique décisionnelle s'est largement diffusée dans les années 1990. Les référentiels de données qui l'alimentent se distinguent à la fois des applications métier qui sous-tendent les activités de l'entreprise et de l'archivage électronique lorsque celui-ci est mis en œuvre. Les données exploitées sont structurées, mais les contenus non structurés ne sont pas exploités par l'informatique décisionnelle.

La science des données regroupe l'analyse prédictive et l'exploration des données. Elle vise à faire des simulations et des hypothèses prédictives (on touche là à l'une des grandes quêtes de l'humanité, celle de la prédiction !). Cette science s'appuie sur des modèles prédictifs et des outils mathématiques d'optimisation et d'analyse statistique. Elle nécessite de nombreuses sources de données hétérogènes et des jeux de données conséquents pour lesquels un traitement en mémoire devient impossible ou nécessite une infrastructure informatique coûteuse. Développée dans les années 1990, la science des données trouve son essor aujourd'hui avec le *big data*, c'est-à-dire la conjoncture de la multiplication des sources de données et des capacités de traitement de données en masse sur une infrastructure informatique standard. En tant que nouvelle approche analytique, elle est potentiellement créatrice de plus de valeur pour ses utilisateurs.

[1] En informatique, et plus particulièrement dans le domaine des bases de données, le traitement analytique en ligne (anglais *online analytical processing*, OLAP) est un type d'application informatique orienté vers l'analyse sur le champ d'informations selon plusieurs axes, dans le but d'obtenir des rapports de synthèse comme ceux utilisés en analyse financière. Les applications de type OLAP sont couramment utilisées en informatique décisionnelle, dans le but d'aider la direction à avoir une vue transversale de l'activité d'une entreprise.

Révolution ou évolution ?

L'évolution des techniques d'analyse est lié à la fois à l'évolution des sources de données, et à celle des architectures informatiques.

Dans les années 1990, les données de l'entreprise se mesuraient en teraoctets et prenaient la forme de données structurées, générées par l'entreprise, et gérées par des bases de données relationnelles ou des entrepôts de données.

Les années 2000, à la faveur de la baisse des coûts du stockage, ont vu l'explosion des contenus non structurés, issus de la dématérialisation des processus, de la mise en place de la GED et de la diffusion de contenus sur Internet. La donnée s'est alors mesurée en petaoctets.

Depuis 2010, les entreprises adoptent de nouvelles techniques de gestion des données basées sur les technologies autres que celles des bases de données traditionnelles. Appelées bases NO-SQL, pour « *not only structured query language* », elles s'appuient sur des couples clé-valeur[1] plutôt que sur des tables structurées. Les volumes se mesurent alors en exaoctets.

En parallèle de cette évolution de la gestion des données et de leur volume, on distingue également trois générations d'architectures informatiques. La première, dite première plateforme, a vu le jour dans les années 1950 et se présente sous la forme de systèmes centralisés auxquels les utilisateurs accèdent depuis des terminaux de saisie. Les systèmes *mainframes* et les terminaux X en sont une illustration. Au début, la première plateforme s'est limitée à un usage interne à l'entreprise avant de s'étendre à des scénarios d'usage entre les entreprises. En fin de compte, elle concerne plusieurs milliers d'applications auxquelles accèdent plusieurs millions d'utilisateurs.

La deuxième plateforme a vu le jour dans les années 1980, d'abord au travers d'architecture client-serveur, puis au travers d'architecture Intranet et Internet. Les applications se comptent alors en dizaines de milliers auxquelles des centaines de millions d'utilisateurs accèdent, essentiellement depuis un ordinateur de bureau, d'abord professionnel, ensuite privé. Avec la deuxième plateforme sont apparues les premières applications gérant les relations entre l'entreprise et ses clients.

La troisième plateforme est apparue ces dernières années avec l'essor des terminaux mobiles, smartphones et tablettes, de l'Internet des machines, et enfin les applications *cloud*. Les applications sont proposées soit sur Internet par des prestataires externes à l'entreprise (*cloud* public), soit déployées sur des

[1] Exemple de couple clé-valeur : clé = prénom (la clé est le Prénom), valeur = Lucie (la valeur de la clé ou du prénom est «Lucie ») ; [clé-valeur] = [prénom-Lucie].

infrastructures informatiques internes (*cloud* privé), qui sont immédiatement disponibles et accessibles depuis l'intérieur ou l'extérieur de l'entreprise et qui garantissent un niveau de service quels que soient la localisation ou le nombre d'utilisateurs connectés. La troisième plateforme a démultiplié le nombre d'applications, qui se comptent dorénavant en millions, et le nombre d'utilisateurs, qui se comptent en milliards. Elle établit surtout une relation quasi permanente entre les individus et les entreprises, soit par le biais d'une relation directe et explicite, soit par le biais des données collectées par des acteurs intermédiaires et revendues aux entreprises.

La conjonction de l'explosion du nombre d'utilisateurs connectés, des flux de données qu'ils génèrent, de la mondialisation de l'économie, de la concurrence qu'elle induit et des progrès des capacités de calcul de l'informatique fait que l'analyse des données a étendu son domaine de l'informatique décisionnelle à la science des données.

Et la science des données induit deux changements opérationnels majeurs vis-à-vis de l'informatique décisionnelle : le besoin d'effectuer des analyses en temps réel, le fait de ne plus prendre seulement en compte des données structurées, mais aussi des données non structurées ou semi-structurées.

L'écosystème du *big data*

Le *big data* induit de nouveaux concepts, aussi bien en termes d'acteurs qu'en termes de technologies.

Pour ce qui est des acteurs, on peut distinguer d'une part les services qui génèrent les données. Ceux-ci existaient bien sur déjà, simplement ils se sont multipliés et sont de plus en plus externes à l'entreprise, comme les distributeurs de billets, les dispositifs de navigation GPS ou les puces RFID[1]. Ensuite viennent les organisations qui collectent les données, telles que les opérateurs téléphoniques, les grandes surfaces ou les administrations. Entre ces derniers et les utilisateurs finaux des données est apparu un nouveau rôle : celui de l'agrégateur de données. Un agrégateur de données peut être un site Web,

[1] La radio-identification, le plus souvent désignée par le sigle RFID (de l'anglais *radio frequency identification*), est une méthode pour mémoriser et récupérer des données à distance en utilisant des marqueurs appelés « radio-étiquettes » (*RFID tag* ou *RFID transponder* en anglais).

une régie publicitaire, un bureau d'analyse crédit ou une compagnie d'assurance. L'agrégateur va vendre aux utilisateurs les données qu'il a agrégées.

D'un point de vue technique, le *big data* nécessite de nouvelles technologies. D'une part, les données induisent des volumes bien plus conséquents auxquelles les techniques de stockage et de calcul traditionnelles ne savent plus répondre. D'autre part, par leur variété et leur vélocité, les données ne peuvent plus s'inscrire dans des structures prédéfinies, telles que celles des bases de données relationnelles ou des cubes OLAP. Ainsi, le *big data* stocke la donnée brute, sans structure *a priori*, en la décrivant simplement à l'aide de couples clé-valeur sur lesquels l'analyse sera effectuée.

On peut faire une analogie entre l'approche traditionnelle des structures prédéfinies et celle des modèles de métadonnées en matière de gestion documentaire. De même, on peut comparer l'approche clé-valeur avec celle qui consisterait à décrire une base documentaire à l'aide d'autant de couples label de métadonnée-valeur que nécessaire, selon les informations dont on dispose sur le moment. C'est ensuite l'analyse de ces couples qui permettrait d'identifier la nature du document et le contexte dans lequel il s'inscrit, pour un moment donné.

Le stockage *big data* utilise des bases de données NO-SQL et le système de gestion de fichiers HDFS (*Hadoop Distributed File System*) qui se caractérisent par une distribution des données et des traitements, de telle sorte que leurs performances restent stables et ne sont quasiment pas impactées par la quantité de données à stocker et à traiter. Le traitement des données se fait à l'aide de la technologie de calcul parallélisé MapReduce.

Une fois mis en œuvre et alimenté, le *big data* permet à l'entreprise de capter et d'analyser les signaux faibles qui prédisent les tendances à venir. Toutefois cette étape reste complexe, d'autant plus que les données ne sont pas structurées, et elle nécessite une bonne connaissance de ces dernières.

Big data et *data lake*

Depuis peu est venu se greffer au *big data* le terme de *data lake*. Cette image d'un lac de données mutualisant les flux bruts de données issus de sources diverses et variées vient illustrer le principe du *big data*, présenté en amont, d'une analyse des données appliquée à des sources hétérogènes et sans structuration *a priori*.

Cependant le rôle du *data lake* va au-delà de celui du réservoir de données analytiques. Il n'a pas pour vocation de remplacer les systèmes structurés traditionnels utilisés par les applications métier et l'informatique décisionnelle. En revanche, ces dernières peuvent exposer leurs propres données au *data lake* (à l'aide d'Hadoop), ou puiser dans le *data lake* des données supplémentaires (par exemple un *data warehouse* qui utilise le *data lake* pour ses traitements récurrents). Enfin, le *data lake* peut aussi jouer le rôle d'un terrain d'expérimentation flexible pour les experts de la science de ces données, lors de leurs phases de découverte et d'idéation. D'une certaine manière, le *data lake* se positionne comme le nouvel infocentre.

La mise en œuvre d'un *data lake* peut d'ailleurs se faire de façon progressive, sans bouleverser l'organisation du système d'information. La majorité des entreprises et des organisations ayant déployé un *data lake* l'ont fait par introduction de la technologie Hadoop afin de construire un nouveau référentiel de données, alimenté par les applications métier et exploité par les *data warehouses*. D'autres, encore minoritaires, sont allées au-delà. Le *data lake* est devenu le stockage par défaut des nouvelles applications. Il est aussi une source de données pour l'ensemble des applications, nouvelles ou anciennes.

L'objectif à long terme, et que personne n'a encore atteint, est celui d'un *data lake* intégrant les notions de sécurité et de gouvernance et fédérateur entre les applications *cloud* et les applications internalisées.

On le voit, une des conséquences du *big data* et du *data lake* est que les données ne sont plus séparées entre systèmes de production, systèmes analytiques et systèmes d'archivage électroniques (SAE). Si toutes les données deviennent sujettes à analyse et à exploitation par le biais du *data lake*, sont-elles toutes sujettes à l'archivage ?

Data lake **et archivage**

Actuellement le mode de fonctionnement du *data lake* est fondamentalement différent de celui d'un SAE puisqu'afin d'appliquer les règles de gouvernance *ad-hoc* (sécurité, conservation et élimination), mais aussi permettre une recherche et une description pertinente, ce dernier s'appuie sur une gestion structurée des archives. Les contenus ou les données y sont décrits, organisés et classifiés selon des schémas définis *a priori*.

D'autre part, le SAE vise à conserver l'information sous une forme pérenne et indépendante des systèmes producteurs, là où le *data lake* stocke la donnée

brute. Le SAE gère aussi la sécurité, les durées de conservation, les règles d'élimination, les gels en cas de litige et la traçabilité, c'est-à-dire des fonctionnalités que les *data lakes* commencent à peine à prendre en compte.

S'il est évident qu'un SAE peut exposer ses données à un *data lake*, ou bien utiliser le *data lake* comme son propre stockage, voire devenir le destinataire des données du *data lake* pour lesquelles une conservation rigoureuse s'impose, un certain nombre d'interrogations demeure.

Le *data lake* n'ayant *a priori* aucune idée de la nature des données qu'il stocke, le risque est élevé de voir des données personnelles y subsister au-delà de la durée légale, ou de voir des données sensibles être accessibles à tous. Ce risque est d'autant plus élevé que, à la différence des *data warehouses* historiques, le *data lake* contient une majorité de contenus non structurés, tels que des documents, plus riches en informations et sans doute plus engageants pour l'entreprise.

Inversement, un système d'archivage électronique a-t-il vocation à archiver l'intégralité du *data lake* ? D'une part, la plupart des SAE actuels sont incapables de gérer la quantité de données du *data lake* car ils s'appuient sur une architecture traditionnelle ; d'autre part, l'intégralité du *data lake* n'a sûrement pas vocation à être conservée.

Si le SAE peut apporter de la structure au *data lake*, et donc de la valeur en termes d'analytique, on peut aussi imaginer que le *data lake* puisse apporter de la valeur à l'archivage.

En effet, autant les SAE savent conserver les données connues, c'est-à-dire celles que l'on sait faire correspondre à des classes d'archives prédéfinies, autant l'archivage des sources de données non gérées (poste de travail, espaces collaboratifs, emails, serveurs de fichiers, réseaux sociaux, etc.) est complexe. La génération de structure à partir d'une information non structurée nécessite alors soit une saisie manuelle, avec toutes les limites qu'elle induit en termes de scalabilité, soit l'utilisation d'outils de type *eDiscovery* qui analysent à la fois les sources et le contenu des documents.

Une fois correctement outillé avec des fonctions de gouvernance, ne pourrait-on pas attendre du *data lake* une capacité à identifier automatiquement lesquelles de ses données sont des archives ou des données à archiver ? Et, ne pourrait-il pas identifier par lui-même quel profil d'archivage doit s'y appliquer ? Devra-t-on conserver l'approche selon laquelle une archive est nécessairement décrite par une structure définie *a priori*, ou bien se construira-t-elle progressivement ?

Une chose est sûre, l'archivage électronique ne pourra pas faire abstraction du *big data* et du *data lake* qui le sous-tend.

Big stockage ou big archivage ?

Dans le cadre posé ici, l'archivage est souvent perçu comme une approche plutôt dédiée aux documents (ou données non structurées), tandis que le *big data* est souvent perçu comme une approche dédiée aux données structurées. Or, il semble qu'aujourd'hui ces deux approches doivent nécessairement fusionner, ou, *a minima*, converger. Pour autant, à l'image de ce qui est exposé précédemment, persiste également la question de savoir si une telle démarche doit viser une solution de type « stockage » ou une solution de type « archivage ».

Sur ce plan, la mise en œuvre de SAE depuis déjà quelques années peut apporter des éléments de réponse. En effet, il apparaît qu'un angle d'attaque pertinent consiste en premier lieu à déterminer quel est le besoin réel, ou tout du moins, le besoin prioritaire ou majoritaire. S'agit-il notamment de la haute disponibilité des informations ? l'intégrité des informations ? la confidentialité des informations ? la pérennité des informations ? etc.

Ainsi le socle de la solution pourra être défini en fonction de l'objectif majeur recherché : un archivage patrimonial, un archivage à valeur probatoire, la préservation d'un savoir-faire, etc. sans oublier de mener une analyse circonstanciée en matière de gestion des risques. Et l'archivage est alors plutôt à considérer comme un « service », à décliner en fonction des objectifs visés : preuve, ré-utilisation, allègement des flux à manipuler, etc.

Pour l'APROGED[1]

Marie-Odile CHARAUDEAU
Alexis FRITEL
Charles HUOT
Philippe MARTIN
Laurent PRÉVEL

[1] À compter du 10 septembre 2015, l'ensemble des activités de l'association ont été arrêtées suite à la décision du Tribunal de Grande Instance de Paris qui a prononcé la liquidation judiciaire sans maintien d'activité.