1. Indexation et grammatisation

L'indexation et la recherche d'information qui lui est associée s'inscrivent dans la dynamique globale de la grammatisation des contenus. Ces deux termes méritent d'être précisés.

Par contenus, nous comprenons[1] des objets matériels, physiques, manifestant une forme sémiotique d'expression. Autrement dit, un contenu associe un support matériel de manifestation et une forme sémiotique d'expression. Un contenu est donc toujours matériel, mais il communique un sens reconnu par ceux qui mobilisent le contenu comme tel, et non comme un simple objet physique. La parole associe, par exemple, l'onde sonore comme support physique et le signifiant linguistique comme forme sémiotique. De même, l'onde acoustique peut être seulement un son, un bruit, c'est alors un objet physique et rien d'autre ; mais sélectionné pour être intégré dans une composition musicale, ce son devient un contenu car il prend une valeur esthétique.

La grammatisation[2] est un processus qui s'inscrit dans la longue durée et qui consiste à rendre toujours plus manipulables les supports matériels des contenus. Ce processus repose sur le recours à de nouveaux types d'inscription, transcrivant des contenus sur de nouveaux supports, leur conférant ainsi une nouvelle manipulabilité. Par exemple, l'écriture est une transposition de la parole rendant la langue observable[3], rapportant la succession dans le temps de la parole à la synopsis permanente de l'inscrit. Or, la synthèse spatiale permet, par le truchement du support d'inscription, de manipuler le contenu, de la fragmenter, de le recomposer, et finalement de dégager des structures inédites, car appartenant à l'ordre de l'écrit manipulable et non à celui de l'oral[4].

La grammatisation est inhérente à toute tentative d'expression, car ces dernières font nécessairement appel à un support matériel qui, comme tel, permet certaines

---

1. Bruno Bachimont, *Patrimoine et numérique : technique et politique de la mémoire*, Bry-sur-Marne, INA, 2017 (Médias et humanités).

2. Sylvain Auroux, *La Révolution technologique de la grammatisation : introduction à l'histoire des sciences du langage*, Liège, Mardaga, 1994. Bernard Stiegler, *La Technique et le Temps. Tome 1 : La faute d'Épiméthée*, Paris, Galilée-Cité des sciences et de l'industrie, 1994 (La technique et le temps).

3. Sylvain Auroux, *op. cit*.

4. Jack Goody, Jean Bazin, Alban Bensa, *La Raison graphique : la domestication de la pensée sauvage*, Paris, Les Éditions de Minuit, 1978 (Le sens commun).

manipulations physiques. L'histoire des supports matériels d'expression des contenus renvoie à la manière dont les différents supports se sont succédé dans le sens d'une plus grande manipulabilité, pour arriver aujourd'hui à un support dont le principe même est la manipulation, à savoir le numérique.

Le numérique est singulier car il est à la fois une rupture et une continuité. Continuité, c'est une évidence. Dans la dynamique de la grammatisation, le numérique n'est qu'une étape de plus dans l'histoire des supports apportant de nouvelles possibilités de manipulation. Ainsi que cela a été plusieurs fois souligné, le numérique a fait pour l'audiovisuel ce que l'écriture a apporté à la parole : un enregistrement permettant d'observer et de manipuler le contenu. Mais le numérique fait également rupture avec tout ce que l'on a connu avant. Sans prétendre à l'exhaustivité, on peut relever les points suivants.

En premier lieu, le numérique s'adosse à une théorie, le calcul, dont le principe même est de décrire les limites et possibilités de manipulation : étant donné des éléments fondamentaux et premiers (un alphabet de symboles primitifs, par exemple binaires, les fameux 0 et 1), la théorie du calcul permet d'explorer les algorithmes ou méthodes de résolution de problème que l'on peut construire par la manipulation de ces symboles. Autrement dit, le calcul est la manipulabilité comme résolution de problème : si on possède une méthode pour résoudre un problème donné, on peut la traduire en un algorithme de calcul (c'est la fameuse thèse de Church[5]). De ce point de vue, le numérique n'est pas seulement une étape de la grammatisation, mais son aboutissement. Quand une entité est numérisée, elle est rapportée à du calculable, c'est-à-dire à de la manipulation. Le calcul est la fin et la finalité de la grammatisation.

Mais c'est théorique. Cela signifie qu'en pratique, les contenus ne sont pas réduits à du pur calculable, mais seulement ce que l'on aura retenu pour les coder (les pixels d'une photo avec la quantification et échantillonnage associés) sera calculable. Le contenu est réduit à un code, et c'est seulement ce dernier qui est calculable, non le contenu d'origine lui-même. C'est la raison pour laquelle, en deçà du plan théorique où le calcul est l'aboutissement de la grammatisation, on a une histoire des supports numériques liés aux différentes méthodes, approches, compromis, effectués pour coder les contenus. Si, en théorie, le numérique est l'aboutissement d'une continuité, en pratique, il se manifeste à travers une série de ruptures scandées par des innovations technologiques.

Les outils du calcul ayant été mobilisés pour coder toutes sortes de contenus, on assiste à une unification et accélération des innovations techniques. L'unification renvoie au fait que toutes les étapes de cycle de vie (élaboration, transmission, consultation, conservation) des contenus, ainsi que les différentes modalités de leur utilisation, sont désormais fondées sur le calcul, à savoir des codes. L'accélération concerne le fait qu'il suffit de nouveaux codes (par exemple les différents formats utilisés pour coder les images, sons, textes, etc.) pour faciliter certains usages et utilisations ; or, un code est une convention d'écriture, son élaboration ne coûte rien, sinon les relais de son adoption par le marché et les usages. En outre, le code pouvant

5. René Lalement, *Logique, réduction, résolution*, Paris-Milan-Barcelone, Masson, 1990 (Études et recherches en informatique).

être réalisé par différents principes physiques, les outils numériques peuvent bénéficier de toutes les innovations concernant les supports physiques de codage. Enfin, le calcul étant lui-même mis à contribution pour élaborer l'innovation, il contribue lui-même à l'accélération de ses mises en œuvre.

De manière synthétique et nécessairement arbitraire, on peut se souvenir des principales étapes ayant scandé cette petite histoire du numérique. On peut distinguer le point de vue des données, celui des machines, et enfin celui des algorithmes.

Du point de vue des données, le numérique suit les étapes du codage. Initialement, les symboles numériques codèrent des nombres : on s'en servit pour faire du calcul scientifique, l'ordinateur étant un *number cruncher*. Puis, on réalisa que les symboles numériques pouvaient coder autre chose que des nombres, mais n'importe quel symbole, par exemple ceux de l'alphabet ou de n'importe quel répertoire donné. Cela donna lieu à l'informatique de gestion. Puis ce fut le tour des contenus non alphabétiques, les sons, les images, l'audiovisuel. Cet élargissement du périmètre du codage provoqua à chaque fois une mutation du système industriel et bureaucratique associés. Nous vivons actuellement la troisième étape, inaugurée dans les années 1990.

Du point de vue des machines[6], l'évolution s'ancre dans les machines à calculer pour évoluer vers celles pourvues de mémoire, marquant ainsi la possibilité pour l'ingénierie de répliquer la structure des machines théoriques conçues par les logiciens[7].

Enfin, du point de vue des algorithmes, tout dépend du paradigme de programmation adopté, où il faut allier à la fois un type de manipulation (programmation fonctionnelle, résolution logique, réduction du lambda calcul, etc.) et une métaphore de programmation (comme le cerveau – réseaux de neurones formels, comme l'évolution – les algorithmes génétiques, comme la réflexion logique – la résolution en Prolog, etc.).

À ce jour, c'est dans ce paysage que nous sommes plongés. L'indexation est insérée et enserrée dans ces différentes tendances qui l'encadrent, mais aussi qui la débordent. Il convient de revenir à ses fondamentaux pour mettre en perspective à la fois les invariants qui la constituent et les approches qui se succèdent.

Documentation, information, indexation

De la documentation à l'information

L'indexation appartient aux métiers de la documentation. Ces derniers ne sont pas si anciens, rejoignant ceux du document qui leurs préexistaient. En effet, la documentation est la fille de la révolution industrielle et des besoins en accès à la

6. Vernon Pratt, *Machines à penser : une histoire de l'intelligence artificielle*, traduit par Christian Puech, Paris, Presses universitaires de France, 1995 (Sciences, modernités, philosophies). Pierre Wagner, *La Machine en logique*, Paris, Presses universitaires de France, 1998 (Science, histoire et société).

7. Alan Turing, « Théorie des nombres calculables, suivie d'une application au problème de la décision », dans Jean-Yves Girard (éd.), *La Machine de Turing*, Paris, Le Seuil, 1936/1995 (Sources du savoir).

connaissance scientifique et technique permettant d'encadrer, soutenir et accompagner la mise en place des processus industriels. Si cet usage entraîna l'émergence de nouveaux métiers, c'est qu'il correspond à un nouveau rapport au contenu ou au document. En effet, le rapport aux documents s'est construit sur deux piliers fondamentaux : la mémoire de la preuve (les archives) et la mémoire de l'œuvre (les bibliothèques)[8]. Dans le premier cas, le document est la trace de l'événement dont il est une conséquence quasi causale : le document est par lui-même la preuve de ce dont il parle car il en est le résultat. Le juriste et l'historien sont les principaux utilisateurs des archives comme preuves de ce qui s'est passé – l'archiviste devant par conséquent en assurer l'authenticité pour qu'elles puissent fonctionner comme telles. Dans le second cas, à l'instar de la bibliothèque d'Alexandrie et de l'ordonnance de Montpellier (1537) ultérieure, l'enjeu est de garder une trace du génie humain, littéraire, artistique, scientifique. C'est donc l'œuvre comme création spirituelle ou intellectuelle qu'il s'agit de conserver.

Dans la documentation, comme le montre dans son traité séminal Suzanne Briet[9], l'enjeu n'est pas la preuve historique ou juridique, ni l'œuvre, mais l'information : ce qu'il faut connaître, en guise de savoir établi à l'appui d'une action à entreprendre ou d'une décision à prendre. Le document comme objet n'est pas ce qui est important, ni son auteur, mais l'information comme telle. Les métiers de la documentation sont les premiers à assumer une vision abstraite du document comme ressource d'information, indépendamment de son support, véhicule, auteur. Il s'agit donc de préciser ici la notion d'information qui nous intéresse.

Information, format et forme sémiotique

Pour cela, revenons à la notion de contenu telle que nous l'avons caractérisée : un support matériel de manifestation d'une forme sémiotique d'expression ; un objet qui fait signe, s'adresse à nous et auquel nous répondons par une interprétation.

L'interprétation permet de dégager ce qui fait signe, les unités signifiantes du contenu. Nous parlerons en ce qui nous concerne d'« unité sémiotique d'interprétation » (USI) pour désigner les plus petites entités signifiantes dégagées par l'interprétation. Définies *a posteriori*, elles ne conditionnent pas le sens, mais en résultent : c'est seulement parce que l'on mène l'interprétation que l'on dégage leur signification et isole leur unité, mais elles ne la précèdent pas. Autrement dit, c'est l'interprétation d'un contenu qui nous dit de quoi il est fait, se compose : quels mots, quelles formes sonores ou visuelles, etc. L'interprétation dépendant du contexte, un même contenu ne recèle pas les mêmes signes selon les situations de consultation : les USI dépendent du contexte.

Par ailleurs, l'inscription est fixée sur un support fixe et permanent. Ce support est la plupart temps le fruit d'une élaboration technique particulière. Or, comme le support apporte à la forme d'expression sa manipulabilité matérielle, on retrouvera

8. Bruno Bachimont, *Patrimoine et numérique*, *op. cit.*

9. Suzanne Briet, *Qu'est-ce que la documentation ?*, Paris, Éditions documentaires, industrielles et techniques, 1951 (Collection de documentologie).

dans la manipulabilité du contenu les possibilités offertes par le conditionnement technique du support. Si on prend l'exemple du livre, les techniques du papier permettent ainsi d'écrire de chaque côté d'un même feuillet, et de les coller et relier par la tranche, offrant ainsi une navigation interpaginale facilitée. Le même livre en numérique sera représenté techniquement par une suite de caractères codés, Unicode en général, plus des codes spécifiques de mise en forme. On pourra alors manipuler le livre au caractère près, et le transformer facilement.

La notion de format découle de cette manipulabilité technique du contenu. On appellera « format » la définition des unités techniquement manipulables dans le support et des opérations de manipulation qui y sont associées. En particulier, nous distinguerons ce que nous nommerons les « unités techniques de manipulation » ou UTM qui sont les plus petites unités que l'on peut manipuler techniquement dans un contenu :

– pour une photo numérique, l'UTM est le pixel ;
– pour un texte numérique, l'UTM est le caractère ;
– pour un texte imprimé sur des feuilles volantes, l'UTM est la feuille ;
– pour un livre imprimé, l'UTM est le livre lui-même.

On constate une dissymétrie entre la production du contenu et sa lecture : les UTM ne sont pas forcément les mêmes. Ainsi, lors de l'impression classique, les UTM sont les caractères de la casse, ce qui n'est plus le cas lors de la lecture, où le lecteur ne peut que manipuler le livre et tourner les pages. De même, la production audiovisuelle manipule une vidéo à l'image près, mais le spectateur a rarement cette possibilité, qui est certes de plus en plus offerte par les outils de lecture.

On remarque également que ce qui est une UTM, par exemple le pixel d'une photo, n'a pas obligatoirement une signification selon la forme sémiotique d'expression associée au contenu, par exemple la forme graphique ou photographique. En effet, un pixel n'a pas de signification particulière, de même le caractère dans un texte (parfois, un caractère peut être un morphème et avoir un sens – le « s » du pluriel – ou non).

Ce qui est techniquement manipulable n'a pas obligatoirement une signification assignable selon la forme sémiotique d'expression associée au support formaté. Cela entraîne des conséquences importantes pour notre propos : cela signifie en particulier que ce qu'on a appelé les *unités sémiotiques d'interprétation* (USI) ne sont pas définies *a priori* dans le contenu mais *a posteriori* à travers le parcours interprétatif. Or, les unités techniques de manipulation (UTM) sont définies *a priori* du fait du formatage technique de l'inscription. Si, dans le cas général, les UTM sont définies *a priori* par le formatage de l'inscription et les USI *a posteriori* par la mise en œuvre de l'interprétation, il ne peut y avoir de correspondance réglée entre les UTM et les USI, entre ce que l'on fait du contenu et ce que cela veut dire.

En conséquence, dans le cas général, il existe un arbitraire entre la structure technique du support et les structures sémantiques que l'interprétation dégage : s'il y a bien une influence réciproque, celle-ci ne se laisse pas décrire par des règles ou des correspondances qui permettent de dire *a priori* ce qu'entraîne au niveau sémantique une transformation physique donnée, ou de prescrire la transformation qu'il faut effectuer pour obtenir un effet sémantique donné. On peut donc toujours *constater* l'influence réciproque entre le support et l'interprétation, mais on ne peut la prédire.

Il reste que parfois on veut résorber cet arbitraire dans une correspondance explicite et *a priori* entre le support et le sens, entre la structure matérielle et l'interprétation. Or, ce qui permet d'établir une correspondance *a priori* entre le signe et le sens est la notion de formalisation. Comme son nom l'indique, ce qui est formel a un sens en fonction de sa forme. En outre, toute modification de la forme entraîne une modification prédictible du sens associé.

C'est pourquoi il faut distinguer deux niveaux dans la structuration physique du support :

• La structuration physique permettant la manipulation, mais le rapport au sens restera arbitraire ;

– Le *format technique* définit les unités de manipulation et ouvre un espace opératoire des possibles techniques accessibles à une inscription ; selon nos définitions, les UTM et les USI sont différentes et les correspondances entre elles arbitraires.

– Dans ce cas, on a affaire à des *informations non structurées*, qu'on appelle aussi des contenus culturels, dans la mesure où il s'agit de la plupart des contenus dans le contexte des médias culturels, d'une part, et dans la mesure où leur interprétation repose sur des conventions culturelles indépendantes du format technique, d'autre part.

• La structuration physique permettant la manipulation et une correspondance réglée entre cette manipulation et ses conséquences sémantiques.

– On parlera à ce niveau de formalisation ou de format logique, qui définit outre des unités de manipulation, une correspondance avec les unités d'interprétation. Dans ce cas, selon nos définitions, les UTM et les USI sont identiques.

– Dans ce cas, on a affaire à des informations structurées, qu'on appellera des contenus logico-formels ou formalisés.

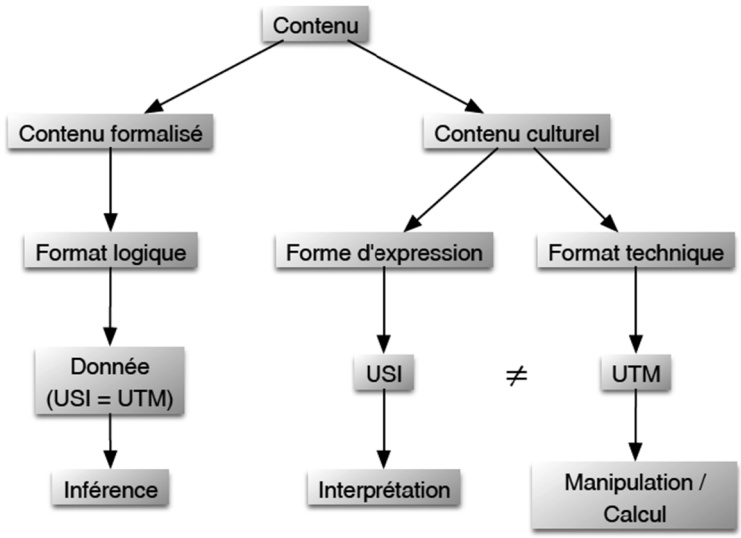
Dans le cas général, le format technique définit des unités matérielles qui offrent une prise à la manipulation, cette dernière étant de type technique et machinale, en tant qu'elle peut être menée sans interprétation. En particulier, une machine est capable de reconnaître les entités définies par le format et de leur appliquer un traitement donné. La reconnaissance des unités prescrites par le format ne requiert aucune compétence interprétative particulière, aucune décision qui ne puisse être traduite en une méthode ou processus appliqué par une machine.

Le format technique suppose donc une syntaxe, une syntaxe d'assemblage, d'une part, qui prescrit l'agencement des unités, et une syntaxe opératoire, d'autre part, qui prescrit les opérations qu'on peut leur appliquer. À ce titre, on peut dire que le format technique constitue une information au sens où l'information formatée implique une signification opérationnelle, à savoir les transformations qu'on peut effectuer.

Le format logique constitue une abstraction supplémentaire, car il ajoute au format technique une liaison nécessaire à une sémantique interprétative et pas seulement opérationnelle. À ce stade, le format logique permet de définir ce qu'on appelle habituellement des informations.

La notion de donnée

Dès lors qu'on a la notion d'information formatée logiquement, on peut définir celle de donnée. La donnée sera la plus petite partie d'information pourvue d'une



Format logique, format d'expression et format technique

signification. La donnée est donc un cas particulier d'UTM : de manière générale, une UTM n'a pas obligatoirement de signification. La donnée est une UTM qui a un sens. C'est l'enregistrement élémentaire de la base de données, ou l'information de base d'un système de gestion.

Cela implique qu'il n'est pas toujours possible de déterminer un « format de donnée ». En effet, quand on a affaire à un contenu quelconque, il existe un lien arbitraire entre la manipulation technique du support matériel et les conséquences sémiotiques de cette manipulation. Autrement dit, en modifiant la configuration matérielle d'un contenu, je sais bien que je modifie le sens, mais je ne sais pas dire en quoi, dans quelle mesure, dans quelle proportion. Par conséquent, de manière générale, si je peux définir *a priori* des UTM, c'est-à-dire les plus petites unités de manipulation, je ne sais pas définir *a priori* ce que seront les données et leur format, car je ne sais pas dire comment s'articuleront le sens et le support.

Pour cela, il faut recourir à une configuration particulière du format matériel, et le concevoir de manière telle qu'il possède *a priori* une corrélation avec la signification qu'il véhicule. Dans ce cas, si on modifie l'information, on sait dire en quoi la signification a changé.

Par exemple, si on considère des textes en langue naturelle, les UTM sont les caractères. Or, si on manipule les caractères, on ne sait pas prédire *a priori* les conséquences sémiotiques. C'est d'ailleurs ce que les artistes ont bien compris, jouant sur le fait de créer des configurations matérielles signifiantes inédites (!) et insoupçonnées à partir d'une libre manipulation des formats d'inscription. Dans la langue naturelle, on n'a pas de donnée, le format d'inscription n'est pas un format de donnée.

Si on considère les langages formels de la logique ou de l'informatique, en particulier les langages permettant l'expression des schémas de base de données, on

mobilise alors des langages conçus pour établir un lien explicite entre la syntaxe et la sémantique, entre le formatage physique et la signification sémiotique. Dans ces derniers cas, toute modification matérielle a une conséquence prédictible et réglée (les règles de la sémantique formelle) sur le sens. On a donc ici des formats de données.

Les initiatives récentes du web sémantique adoptent la notion de format logique en proposant des formalismes permettant la déduction et l'inférence. L'objectif étant de confier à la machine des tâches requérant une compétence sémantique, l'approche est de représenter cette sémantique par une syntaxe qui, manipulée par la machine, produit les effets de sens escomptés. Mais pour cela, il faut un parallélisme strict entre syntaxe et sémantique. Ces langages[10], RDF, OWL par exemple, contribuent à transformer les contenus du web en données, d'où les appellations récentes de web de données qui désormais remplacent la terminologie de « web sémantique ».

La recherche de l'information et ses paradigmes

Il appert à présent qu'il faut distinguer plusieurs choses :

• L'information structurée, comme contenu formaté logiquement dont la structure formelle commande l'interprétation et la signification, et l'information non structurée ;

• Pour une information, le format qui la structure techniquement de la forme sémiotique qui commande son interprétation ; la forme se constitue autour de signes dégagés *a posteriori*, le format définit *a priori* les structures composant le contenu. Il faut donc distinguer l'in*formati*on de l'in*form*ation.

Comme on l'a dit plus haut, l'enjeu de l'indexation est la recherche d'information, à savoir l'information, peu importe sa forme documentaire, nécessaire pour un usager donné. Comme nos considérations sur la notion d'information peuvent le laisser supposer, il faudra distinguer les différents types de recherche d'information selon la manière dont l'information recherchée se caractérise. Schématiquement, on retiendra trois grands types de recherche d'information :

• L'information est structurée : la ressource d'information est elle-même structurée. La recherche d'information reposera sur cette structure : c'est le paradigme des bases de données.

• L'information est non structurée : la recherche d'information devra introduire une structure rendant manipulable et récupérable le contenu ; c'est le paradigme de l'indexation. Dans ce contexte on verra se dégager l'approche « documents indexés », où l'on indexe des documents comme tels, et l'approche « ressource annotée », où les documents ne sont plus que des agrégats de ressources assemblées et annotés.

• L'information est non structurée et on recherche la structure à même la ressource : c'est le paradigme de l'exploration des données (*big data* et *machine learning*).

---

10. Fabien Gandon, Catherine Faron-Zucker et Olivier Corby, *Le Web sémantique : comment lier les données et les schémas sur le web ?*, Paris, Dunod, 2012 (InfoPro).

L'information structurée

Selon cette première approche, il est possible de prédéfinir ce que l'on recherche et de structurer l'information dont on dispose. En particulier, l'information peut être formatée selon une structure formelle dont on maîtrise la logique. Cela signifie que l'information à explorer pour retrouver ce que l'on recherche est agencée selon une syntaxe formelle générative, et interprétée selon une sémantique compositionnelle.

La notion de syntaxe générative correspond au fait que tous les énoncés ou informations que l'on a à considérer renvoient à une même structure logique fondée sur des règles de génération, spécifiant comment agencer et combiner les primitives formelles pour en faire des objets complexes. Par exemple, si on sait que *p* et *q* sont des formules, alors une règle précise que l'on peut construire une autre formule, *p ou q* à partir de ces dernières. La sémantique compositionnelle précise comment donner une signification d'une formule complexe à partir des formules élémentaires la composant. Ainsi, *p ou q* sera fausse si et seulement si *p* et *q* sont fausses, et vraie sinon.

On comprend que, dans ce cas, la syntaxe formelle spécifie les UTM et leur manipulation, et la sémantique compositionnelle les USI et leur interprétation. Par construction, puisque la sémantique est compositionnelle, USI et UTM sont alignées : tout ce qui est manipulable est interprétable, toute signification visée est constructible à partir des UTM.

C'est l'approche adoptée dans les bases de données où le schéma conceptuel détermine la structure des enregistrements et optimise les conditions de la recherche puisqu'il est possible d'organiser le stockage interne des informations en fonction de leur structure connue et du type des requêtes prévues.

Cette approche repose cependant sur une hypothèse forte, qui consiste dans la possibilité de rapporter l'information disponible à une structure formelle et organisée. Elle entraîne en outre un coût – il faut traduire cette information disponible en sa contrepartie formalisée, mais entraîne un gain considérable : d'une part, la recherche est optimisée ; d'autre part, tout enregistrement ou toute combinaison d'enregistrements renvoient à une structure formelle donc interprétable de l'information. En structurant *a priori* l'information, on définit une organisation où tout ce qui est manipulable est porteur de sens, et ce qui est porteur de sens est manipulable.

L'information non structurée

Cette approche fait merveille dans le domaine de l'information structurée, notamment ce qui relève de la gestion et de l'administration. Mais quand l'information est trop complexe pour être rapportée à une structure formelle, ou non disponible *a priori*, il faut alors envisager d'apporter de manière exogène et *a posteriori* la structure dont on a besoin pour manipuler l'information et la retrouver. On a désormais une ressource, ou un document, auquel on adjoint *a posteriori* une structure d'index, éléments connus et manipulables, qui confèrent à la ressource à la fois un découpage et une signifiance des éléments ou fragments ainsi repérés. Il n'y a donc pas de contrainte sur la création de la ressource, elle est ce qu'elle est, mais sur le traitement *a posteriori* de cette dernière pour apporter l'intelligibilité qui nous intéresse et la manipulabilité pour la retrouver. Toute ressource est éligible, contrairement à

l'approche précédente, mais l'accès aux ressources doit en passer par l'indexation explicite ou manuelle, démarche coûteuse mais précise, ou une indexation automatique, peu coûteuse mais engendrant bruit et silence.

Le principe de base est donc que l'index est ajouté à la ressource, et toute la structure que l'on confère à la ressource provient de l'index. C'est pourquoi l'index repose sur plusieurs étapes complémentaires[11] :

• L'index prescrit une *localisation* ; autrement dit, il délimite une partie du contenu, qu'il rend saisissable par son truchement.

– C'est à ce niveau que l'indexation recouvre la ressource d'une structure d'UTM nouvelles, les parties repérables et saisissables par leur truchement. Ce sera par exemple le repérage dans un texte, compris numériquement comme un flux de caractères, de telle position dans le flux à telle autre position. Les contenus, étant formatés, ont déjà une structure d'UTM (les caractères du texte), mais cette structure est trop fine et arbitraire vis-à-vis du sens ; on a besoin d'en délimiter d'autres, plus pertinentes vis-à-vis du sens et qui pourront être qualifiées comme étant signifiantes : c'est l'objet de l'étape suivante.

• L'index prescrit une *qualification* : il associe à chaque partie délimitée une qualification qui lui donne un sens, l'aspect sous lequel on veut considérer cette partie ; par exemple, considérer que la partie délimitée dans le texte est une introduction ;

– L'index donne alors un statut d'USI à la partie délimitée. Cela permet d'aligner ce qui a du sens à ce qui est manipulable. L'indexation établit une correspondance entre UTM et USI, correspondance qui n'existe pas par définition dans les contenus culturels (informations non structurées).

• L'index s'inscrit enfin dans une *structuration* : les index s'agencent entre eux pour constituer une représentation structurée de la ressource. Cette structuration est méréologique (agencement des tous et des parties ou relations de composition de type *est-partie-de*), par exemple en considérant qu'un journal se compose d'articles, ou ontologique (classification en genres et espèces ou relations de spécialisation du type *est_un*), par exemple en considérant qu'un article de journal est un genre de production littéraire ayant un auteur (contrairement aux documents administratifs).

Une manière particulièrement pratique d'apposer des index et de les manipuler est de les insérer sous la forme de balises XML : la structure canonique <balise> partie de ressource </balise> spécifie à la fois la partie délimitée (ce qui est entre la balise ouvrante <balise> et la balise fermante </balise> et la signification conférée à cette partie via le sens du terme utilisé pour la balise, par exemple <introduction>… </introduction> qui permet de repérer la partie du texte correspondant à l'introduction. Enfin, la DTD ou schéma XML déterminant quelles balises on peut utiliser prescrit également la manière de les agencer. Ainsi, l'introduction sera une partie du document qu'on ne pourra trouver qu'une seule fois, au début, suivie d'un corps principal, puis d'une conclusion.

Quand on a affaire à des contenus non textuels, la situation est moins simple, car on ne peut plus utiliser le fait que le texte est un simple flux de caractères. Dans ce cas,

---

11. Bruno Bachimont, *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*, Paris, Hermès-Lavoisier, 2007 (Science informatique et SHS).

en effet, la position où est insérée la balise ouvrante détermine la partie sélectionnée du texte : la localisation de la balise dans un texte localise en même temps la partie concernée de la ressource. Pour un contenu sonore, image ou audiovisuel, il faudra avoir une technique particulière de localisation, ce qu'intègre XML notamment grâce à l'héritage de normes comme HyTime.

L'indexation est un type de grammatisation. En effet, en apposant une structure de manipulation et de signification, l'indexation permet de manipuler un contenu selon le sens qu'on lui associe : je peux retrouver les introductions de plusieurs documents. En ce sens, l'indexation est une opération particulièrement puissante, qui évolue constamment pour affiner la grammatisation qu'elle permet. On peut citer plusieurs spécifications de ce schéma global.

La première, et la plus connue, est sans nul doute l'indexation en texte intégral, c'est-à-dire le fait que toute chaîne de caractères séparée par deux caractères blancs est potentiellement un index : cette chaîne est une partie délimitée du contenu, et elle a pour sens celui qu'elle possède dans la langue d'expression à laquelle elle appartient. Ainsi, le texte est auto-indexé, c'est-à-dire que l'on peut plaquer directement sur le texte une structure d'indexation. Cela ne signifie pas que le texte soit une information structurée, mais qu'on peut lui appliquer par défaut une structure d'indexation.

La seconde est le fait de décomposer les ressources en blocs élémentaires pourvus d'un schéma d'indexation spécifiant leur nature et signification. Ce sont notamment les ressources annotées dans le format RDF agençant des triplets « structure – attribut – valeur ». Dans ce cadre, la grammatisation franchit une étape décisive en renversant le point de vue de l'indexation : il ne s'agit pas tant de prendre une ressource pour la décomposer et la manipuler grâce à une structure d'indexation, que de considérer que toute ressource est l'agencement de ressources élémentaires annotées, les annotations donnant les moyens de les combiner de manière cohérente.

Cette dernière vision est celle que nous connaissons dans le contexte du web de données. Elle donne lieu à une déconstruction assez radicale de l'unité éditoriale qu'est le document pour lui substituer une conception des contenus fondée sur les ressources annotées. Les contenus ne sont plus des documents pensés dans l'unité auctoriale ou éditoriale, mais des agrégats dynamiques rassemblant des ressources élémentaires. Conformément à la vision propre de l'indexation, la ressource ne possède comme intelligibilité pour les outils de manipulation que celle conférée par les annotations. La ressource reste en quelque sorte anonyme, transférée et manipulée par les systèmes de traitement comme une boîte noire, et ne sera interprétée comme signifiante que par l'utilisateur final.

L'information extraite

La grammatisation des contenus ne s'arrête pas à la décomposition en fragments annotés. En effet, la structure manquante des contenus culturels (informations non structurées) peut ne pas être ajoutée de manière exogène, via l'indexation, mais être dégagée de manière endogène, à même les contenus, via des techniques d'extraction de connaissances. C'est le paradigme de l'exploration des données, qu'il faut bien distinguer de celui de l'annotation des données – les données n'ayant pas dans ce contexte le même sens.

Selon ce paradigme, les données ne sont pas des informations comme telles, qu'il suffirait d'annoter pour expliciter leur sémantique ; les données sont en quelque sorte infra-informationnelles, car elles ne prennent sens que par leur redondance et régularité. C'est donc par l'apprentissage que les informations sont dégagées et structurées pour permettre l'usage visé, en général une décision ou classification.

Mais cette approche où la ressource est distinguée de sa structuration – cette dernière étant rapportée *a posteriori* par l'indexation – est en passe d'être contestée par les approches centrées sur les données, dans la mesure où ces dernières, par leur masse, sont leurs propres métadonnées ou index. Retrouver une information dans une masse de ressources consiste alors à l'interroger selon les principes du *machine learning* ou du *big data*.

Outre le déplacement méthodologique qu'implique ce paradigme, il s'agit d'une modification de l'objectif. En effet, il ne s'agit pas tant de retrouver l'information que l'on sait être présente dans les contenus, mais de dégager l'information qui y *serait* présente sous une forme conforme à l'usage visé. L'enjeu n'est donc pas, à partir de contenus qualifiés ou certifiés, d'être capable de les retrouver, mais au contraire de trouver à partir de contenus rassemblés de manière arbitraire, selon la disponibilité de masses d'information rassemblées, quelles informations en dégager.

C'est la raison pour laquelle ce déplacement reconfigure l'apport de l'indexation sans l'annuler ou le supprimer. En effet, quand il ne s'agit pas d'avoir un système sachant décider à notre place car il permet d'accéder à des connaissances ou informations inaccessibles pour nous sinon, l'indexation permet de retrouver la source d'information, authentifiant et légitimant l'information recherchée. L'information n'est pas qualifiée seulement par son utilisation, mais aussi par son origine, ce que l'exploration des données perd par construction, puisque son intérêt est précisément de faire l'économie de l'origine des données pour les intégrer dans un même apprentissage[12].

Des approches complémentaires

La recherche d'information s'articule donc en trois approches complémentaires et distinctes, selon le lien entre les ressources ou contenus et la structure qu'on leur prête. Quand la ressource se fait structure, on formalise et organise une base de données ; quand la ressource est distinguée de sa structure, c'est l'approche par l'indexation ; enfin quand la structure est immanente à la ressource et qu'il faut l'en extraire, on recourt aux approches d'exploration des données. Ces approches renvoient à des paradigmes distincts de l'information.

La première concerne la connaissance formalisée, quand sa réduction formelle clarifie sa signification et optimise sa manipulation. Dans ce contexte, il est toujours utile d'avoir un référentiel des informations, un schéma conceptuel et logique des

12. Bruno Bachimont, « Traces, calcul et interprétation : de la mesure à la donnée », *Azimuth : Philosophical Coordinates in Modern and Contemporary Ages*, vol. 7, IV (2016), p. 13-36.

données, permettant à la fois de lever les ambiguïtés, facilitant ainsi l'exploitation des informations, et d'offrir des services de consultation efficaces.

La deuxième approche concerne l'information non structurée, que nous avons également qualifiée de contenus culturels. Ces ressources sont les formes culturelles sous lesquelles nous rencontrons les sources d'information. Traditionnellement, ce sont les documents que l'indexation permet de rendre manipulables de manière signifiante. Mais la manipulabilité introduite par l'indexation conduit à étendre ce processus à sa limite, en ne voyant plus dans les contenus que des ressources élémentaires, dont l'information se réduit à l'annotation formelle et structurée associée. De cette manière, l'indexation rejoint la première approche concernant l'information structurée, puisque ce sont les annotations formelles qui sont au final gérées et exploitées, les ressources étant accédées et consultées en dernière intention.

Enfin, l'exploration des données n'est plus une approche de recherche d'information visant à retrouver une information que l'on sait être présente, mais concerne davantage une extraction des connaissances qui pourraient y être contenues. La difficulté est que l'information extraite n'est pas intelligible comme telle sinon par l'application qui en est faite, l'explication permettant d'articuler les sources d'information et l'usage des connaissances extraites restant impossible à expliciter.

Ces trois paradigmes ne s'annulent pas mais se complètent, chacun permettant de traiter des situations particulières. Leur apparition successive permet cependant de mieux comprendre l'apport réel de chacun, de les spécialiser et de les adapter selon leur mérite propre et leur efficacité. Le récent engouement qu'on connaît actuellement pour l'exploration des données, où certains y voient le paradigme annulant et remplaçant les autres, ne doit pas faire oublier et ignorer ce constat que les usages sauront assez facilement nous rappeler.

BRUNO BACHIMONT

BIBLIOGRAPHIE

Auroux (Sylvain), *La Révolution technologique de la grammatisation : introduction à l'histoire des sciences du langage*, Liège, Mardaga, 1994 (Philosophie et langage).

Bachimont (Bruno), *Patrimoine et numérique : technique et politique de la mémoire*, Bry-sur-Marne, INA, 2017 (Médias et humanités).

Bachimont (Bruno), « Traces, calcul et interprétation : de la mesure à la donnée », *Azimuth : Philosophical Coordinates in Modern and Contemporary Ages*, vol. 7, IV, 2016, p. 13-36.

Bachimont (Bruno), *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*, Paris, Hermès-Lavoisier, 2007 (Science informatique et SHS).

Briet (Suzanne), *Qu'est-ce que la documentation ?*, Paris, Éditions documentaires, industrielles et techniques, 1951 (Collection de documentologie).

Gandon (Fabien), Faron-Zucker (Catherine) et Corby (Olivier), *Le Web sémantique : comment lier les données et les schémas sur le web ?*, Paris, Dunod, 2012 (InfoPro).

Goody (Jack), Bazin (Jean), Bensa (Alban), *La Raison graphique : la domestication de la pensée sauvage*, Paris, Les Éditions de Minuit, 1978 (Le sens commun).

Lalement (René), *Logique, réduction, résolution*, Paris-Milan-Barcelone, Masson, 1990 (Études et recherches en informatique).

Pratt (Vernon), *Machines à penser : une histoire de l'intelligence artificielle*, traduit par Christian Puech, Paris, Presses universitaires de France, 1995 (Sciences, modernités, philosophies).

Stiegler (Bernard), *La Technique et le temps. Tome I : La faute d'Épiméthée*, Paris, Galilée-Cité des sciences et de l'industrie, 1994 (La technique et le temps).

Turing (Alan), « Théorie des nombres calculables, suivie d'une application au problème de la décision », dans Jean-Yves Girard (éd.), *La Machine de Turing*, Paris, Le Seuil, 1995 (Sources du savoir), p. 49-104.

Wagner (Pierre), *La Machine en logique*, Paris, Presses universitaires de France, 1998 (Science, histoire et société).