L'enrichissement des métadonnées: le cas du secteur commercial du livre

Jacques Hilbey

Master MÉgaDonnées et Analyse Sociale
Mémoire professionnel

le cnam cfa

L'enrichissement des métadonnées

Le cas du secteur commercial du livre

Jacques Hilbey

Entreprise DILICOM
Secteur commercial du livre

Tuteur pédagogique : Gérald Kembellec
Maître d'apprentissage : Vincent Poulvélarie

A Mathilde, pour son soutien sans faille.

Remerciements

Je tiens à remercier tout d'abord mon tuteur pédagogique Gérald Kembellec pour ses conseils précieux, sa confiance, et l'attention qu'il a manifestée en m'accompagnant dans la réalisation de ce mémoire, montrant l'alliance rare des qualités du chercheur et de celles du pédagogue.

Je veux également exprimer toute ma reconnaissance à mon maître d'apprentissage Vincent Poulvélarie pour la patience dont il a maintes fois fait preuve à mon égard, pour son entrain et son infatigable disponibilité pour faire partager sa connaissance du monde du livre.

Table des matières

Liste des tableaux et des figures

Première partie

Deuxième partie

Troisième partie

« Du reste, je déteste tout ce qui ne
fait que m'instruire sans augmenter mon
activité ou l'animer directement ».

Goethe, Lettre à Schiller du 19 décembre 1798

Introduction

L'histoire des métadonnées se confond avec celle du livre. C'est Zénodote, premier bibliothécaire de la Bibliothèque d'Alexandrie, qui aurait eu l'idée d'apposer sur les ouvrages des étiquettes mentionnant le nom de l'auteur et d'autres indications sur la provenance des ouvrages, mais aussi de classer les rouleaux par thèmes et, alphabétiquement, par auteurs. Disposer du nom de l'auteur, du titre d'un ouvrage, ou encore d'une table des matières pour un ouvrage savant, est pour nous tellement évident que nous pourrions être tentés de n'y voir que des parties du document lui-même. Cette histoire nous rappelle que ce sont déjà des informations *à propos du* document, des « métadonnées ». En outre, nous voyons s'esquisser avec cet exemple antique - nous sommes au troisième siècle avant notre ère - différentes questions que nous allons être amenés à aborder au cours de cette étude, puisqu'il s'agit déjà, pour Zénodote, d'être à même de discriminer des auteurs homonymes, de distinguer différentes copies d'un même ouvrage, mais aussi de disposer d'une méthode pour accéder à un exemplaire d'un ouvrage (1, Gartner). Les questionnements qu'a rencontré ce grammairien alexandrin, dans la bibliothèque qui continue de constituer le modèle de toute bibliothèque dans notre mémoire collective, se sont répandus aussi largement que les livres ont été diffusés. Aborder la question de l'enrichissement des métadonnées du livre aujourd'hui, plus de deux millénaires après Zénodote, c'est envisager un registre plus large de métadonnées, de solutions techniques pour les relier aux documents, de collaboration entre différents acteurs, d'intérêts différenciés des lecteurs pour les ouvrages, mais néanmoins s'inscrire dans la même problématique fondamentale du bibliothécaire alexandrin : celle de l'accessibilité.

Paradoxalement, ce qui nous rapproche du problème de Zénodote est aussi ce qui nous éloigne de l'époque où ce problème trouve son origine : si la Bibliothèque d'Alexandrie, avec les cinq cents mille ouvrages qu'on lui prête, est fameuse, c'est parce qu'elle s'inscrit dans une époque où peu de livres sont produits, circulent, sont lus. Une suite d'innovations techniques et industrielles (au premier rang desquelles l'invention des caractères métalliques mobiles par l'imprimeur allemand Johannes Gutenberg, vers 1450), d'évolutions sociales et politiques (notamment la généralisation de l'alphabétisation des populations depuis le XIXè siècle) ont formidablement accru au cours du temps l'édition, le commerce et la disponibilité des livres[1]. Nous connaissons désormais une situation d'abondance de biens culturels, parmi lesquels le livre tient une place éminente. Cette abondance, sans antécédent dans l'histoire de l'humanité, continue sa croissance. La problématique de l'accessibilité s'en trouve renforcée, mais s'y adjoint une autre problématique, si nous envisageons comme nous venons de le faire le marché du livre : celle de la visibilité des ouvrages, évidemment critique pour les éditeurs et les auteurs.

[1] MARTIN Henri-Jean. *Histoire et pouvoir de l'écrit*. Paris : Albin Michel, 1996, 536 p.

Si nous nous plaçons à présent du point de vue des différents usagers du livre, cette offre abondante se traduit par un embarras du choix, et au-delà par une surstimulation de l'attention. C'est cette dynamique entre recherche de visibilité et raréfaction des ressources attentionnelles qui a conduit certains économistes à formuler la notion d'économie de l'attention. Le postulat théorique de départ en est que « dans un monde riche en information, l'abondance d'information entraîne la pénurie d'une autre ressource : la rareté devient ce que consomme l'information. Ce que l'information consomme est assez évident : c'est l'attention de ses receveurs. Donc une abondance d'information crée une rareté de l'attention et le besoin de répartir efficacement cette attention parmi la surabondance des sources d'informations qui peuvent la consommer » (2, Simon, p.40). Les développements actuels de l'économie de l'attention s'attachent plus particulièrement aux médias audiovisuels, au web, aux réseaux sociaux, mais dans le secteur auquel nous nous intéressons, l'essor du livre de poche à partir des années 1930, le développement des bibliothèques publiques dans les années 1970, puis plus récemment l'apparition des livres numériques depuis les années 1990, accréditent la pertinence de cette approche.

L'enrichissement des métadonnées dans le secteur commercial du livre apparaît comme une solution aux problèmes d'accessibilité et de visibilité dont peuvent pâtir les ouvrages. Nous devons néanmoins pointer immédiatement le caractère paradoxal de cette solution, puisqu'en multipliant les métadonnées sur les ouvrages, on ajoute à l'abondance de l'offre une profusion démultipliée d'informations sur celle-ci. Il peut donc sembler que le problème est plutôt dupliqué, reconduit sur un autre niveau : celui des métadonnées, dont l'abondance empêcherait tout autant l'accessibilité et la visibilité que les ouvrages qu'elles qualifient. La question centrale de notre travail sera donc de savoir à quelles conditions l'usage de métadonnées caractérisant les ouvrages peut permettre de surmonter les problèmes d'accessibilité et de visibilité posés par l'abondance de l'offre dans le marché du livre. Afin d'y répondre, nous reviendrons dans une première partie sur des aspects théoriques des métadonnées du livre : quels sont les différents types de métadonnées ? Quelles sont leurs fonctions ? Dans une deuxième partie, nous nous appuierons sur l'expérience de notre alternance au sein de l'entreprise Dilicom pour spécifier les enjeux et les pratiques de l'enrichissement des métadonnées dans le secteur commercial du livre. Enfin, dans une troisième partie, nous envisagerons les perspectives ouvertes par le paradigme des données liées pour l'enrichissement des métadonnées, et nous envisagerons plus concrètement quelles transformations cela supposerait pour les acteurs français du marché du livre.

1. Première partie

Sous le terme générique de « métadonnées », on peut comprendre une importante diversité de données, que ce soit par leur type, par leur rôle, par leur format. Nous devons donc dans ce premier moment de notre étude répondre à la question : de quoi parlons-nous quand nous employons ce terme ?

1.1. Qu'est-ce que les métadonnées ?

1.1.1. Définition des métadonnées

Il n'est pas inutile de partir du mot « métadonnée » lui-même pour commencer à s'interroger sur ce qu'il désigne. Pour le lecteur familier de l'usage académique des préfixes grecs dans la formation de mots savants, le préfixe méta- évoque immédiatement l'idée de réflexivité : un discours ayant pour sujet un ou d'autres discours sera désigné comme un métadiscours, et un langage permettant de décrire d'autres langages comme un métalangage. Cette acception semble éloignée de celles de l'adverbe grec μετα dont il est tiré. Le dictionnaire grec-français de référence pour le grec ancien, celui d'Anatole Bailly, donne pour sens de cet adverbe : ce qui est au milieu de, parmi, mais aussi ce qui vient à la suite de, et ce qui change le lieu ou la condition. La désignation par le titre *Meta ta physica*, qui s'est imposée dès l'antiquité, des livres écrits par Aristote à la suite de sa *Physique* (il faut entendre ici ses livres consacrés à l'étude de la nature, *physis* en grec), n'est sans doute pas étrangère à l'acception académique actuelle. Dans sa métaphysique, Aristote développe en effet une ontologie, soit un discours sur l'être en tant qu'être, une théologie, et d'une manière générale une théorie des causes premières des choses. Écrite *après* sa physique, la métaphysique d'Aristote est également un discours *à propos* de la physique, en ce qu'elle propose des explications théoriques sur le fonctionnement de la nature. Ainsi, ces livres postérieurs, proposant un approfondissement réflexif sur ceux qui venaient avant, pouvaient être considérés comme étant écrits à propos des premiers. Nous pouvons opérer le chemin inverse concernant les métadonnées, et considérer à leur propos cette dimension de consécution temporelle : il peut sembler que des métadonnées doivent nécessairement venir dans un deuxième temps, *après* ce à quoi elles s'appliquent. Est-ce toujours le cas ? Nous serons amenés à y revenir lorsque nous envisagerons la manière dont les métadonnées sont produites.

Mais à quoi s'appliquent-elles ? Muni du sens académique du préfixe méta-, le mot « métadonnées » pourrait nous apparaître désormais évident : il s'agirait de données à propos d'autres données. Nous sommes renvoyés pour étudier cette question à un moment particulier de l'histoire des métadonnées dont nous avons évoqué le commencement dans notre introduction : celui où est formé le mot anglais

metadata, que traduit métadonnées, et qui est contemporain du développement des bases de données. Selon Francis L. Miksa, « le terme métadonnées prend son origine pendant les années 1970, dans la construction de bases de données, où il en est venu à être utilisé comme une manière de différencier entre deux sortes de données trouvées dans les bases de données : les données qui se référaient directement aux objets présents dans la base de données, et les données qui représentaient la catégorie ou le nom du champ (ou du sous-champ) dans lequel la première sorte de données était rangé. Par exemple, dans une base de données d'informations sur les étudiants d'une université, on trouvait les noms des étudiants divisés, disons, en prénom, nom et initiale(s) du milieu. Mais on trouvait également des noms de champs pour ces catégories elles-mêmes, souvent abrégées pour des raisons tenant au codage, par exemple Lastn, Firstn et Midin »[2] (3, Miksa). Pour bien fixer les choses, supposons que l'étudiant John Doe soit recensé dans cette base. « John » est une donnée qui est un attribut de l'étudiant dans la base de données, attribut qui relève de la catégorie « Firstn ». Ce nom de catégorie est lui-même une donnée présente dans la base de données, désignée - pour la distinguer des attributs des étudiants - comme « métadonnée ». Toutefois, l'auteur remarque immédiatement que « cette division des données en deux sortes de couches, données et métadonnées, a bien fonctionné tant que les objets à propos desquels la base de données recensait l'information n'étaient pas des entités informationnelles – en bref, quand elles consistaient en objets comme des personnes, des produits, des processus, et ainsi de suite. Cependant, quand les objets listés dans une base de données devinrent des entités informationnelles telles que des livres, des périodiques, des enregistrements sonores, etc. (un catalogue ou index informatisé est, en fait, un « base de données bibliographique » ou une « base de données d'entités informationnelles ») une nouvelle situation émergea, parce que les entités informationnelles étaient elles-mêmes principalement composées de données. Techniquement, cela produisit pas moins de trois couches de données : les données existantes dans les entités informationnelles elles-mêmes : leurs textes, illustrations graphiques, etc., les données qui existaient dans la base de données informatique et qui consistaient en noms de catégories/champs, et les données proprement dites dans la base de données informatique qui renvoyaient aux entités informationnelles. La solution parmi les organisateurs d'entités informationnelles pour distinguer les différentes sortes de données a été de désigner toutes les données dans la base de données comme métadonnées » (ibid.). Prenons à nouveau le temps de bien comprendre ce que nous dit Francis L. Miska : si *Les misérables* de Victor Hugo se trouve dans une base de données référençant des livres, alors le champ « nom d'auteur » et « Victor Hugo » sont tous deux désignés comme des métadonnées, et seul le livre *Les Misérables* est désigné comme « donnée ».

[2] Nous traduisons.

Ce rappel historique est très précieux pour nous, puisqu'il nous permet de commencer à poser le cadre conceptuel dans lequel nous allons pouvoir inscrire les métadonnées. En premier lieu, il nous rappelle cette vérité d'évidence que quelle que soit l'extension que l'on donne au terme de métadonnées, il s'agit de données. En deuxième lieu, il nous permet de comprendre comment le terme de métadonnées, qui avait bien d'abord cet aspect réflexif de « données sur des données », a pu en venir à désigner un attribut d'un objet qui n'est pas lui-même présent dans la base de données, au titre que cet objet, en tant qu'entité informationnelle, est lui-même composé principalement de données. En troisième lieu, il nous invite à considérer très généralement qu'une donnée porte sur une entité, et que cette entité est susceptible dans certains cas d'être elle-même de nature informationnelle. Rappelons toutefois qu'une métadonnée ne se trouve pas nécessairement dans une base de données, qu'elle n'est pas forcément une donnée numérique. Anticipant sur la troisième partie de notre travail, nous proposons de reprendre un lexique utilisé dans le modèle *Resource Description Framework* et de désigner par le terme « ressource » les entités sur lesquelles portent les métadonnées, indépendamment du mode d'existence de ces entités, qu'il s'agisse de personnes, de documents ou de livres, par le terme « propriété » les noms de catégories (comme dans les exemples ci-dessus « Firstn » ou « nom d'auteur »), et par le terme de « valeur » de la propriété la donnée qui se réfère directement à l'objet (« John » ou « Victor Hugo », toujours dans nos exemples). Parler indifféremment de « ressource » renvoie ici au besoin d'un terme conventionnel et générique pour désigner toute entité à laquelle peut être associée des données. De ce point de vue, le besoin qu'évoquait Francis L. Miksa de distinguer d'un côté personnes, produits, processus et de l'autre entités informationnelles ne se fait plus vraiment ressentir, et la notion de ressource permet même de s'adapter plus souplement à différentes « granularités », les textes et les illustrations que contient une entité informationnelle étant elles-mêmes susceptibles d'être désignées comme telles. Quant à la confusion possible dans le même vocable de « métadonnée » d'une propriété et de la valeur de cette propriété, elle ne nous semble pas plus gênante que celle, omniprésente dans le langage courant, par laquelle on peut dire que « la taille de John est 1,87m » et que « la taille de John est la longueur de son corps ». Nous nous permettrons donc de considérer à la fois que le nom d'auteur est une métadonnée d'un livre et que « Victor Hugo » est une métadonnée du livre *Les Misérables*. Nous disposons toutefois désormais des termes de « propriété » et de « valeur » si nous étions amenés à devoir distinguer explicitement les deux acceptions. Pour répondre à la question que nous nous posions de savoir à quoi s'appliquent les métadonnées, nous pouvons à présent apporter une réponse : une métadonnée est une donnée qui est une valeur d'une propriété d'une ressource, mais par extension et facilité de langage, elle peut désigner également la propriété elle-même.

1.1.2. Métadonnées et modèle Donnée-Information-Connaissance

Même si nous avons choisi de définir les métadonnées dans un modèle qui distingue ressource, propriété de cette ressource et valeur de cette propriété, nous avons admis que les métadonnées étaient des données. Toutefois, l'article de Francis Miksa que nous venons de citer parle parfois des métadonnées comme de données, parfois comme d'informations. Les deux termes peuvent-ils être indifféremment employés ? Il semble au premier abord que ce n'est pas le cas. Si nous ne possédons que la valeur d'une propriété, il ne semble pas qu'elle nous informe beaucoup, tant que nous ne savons pas de quelle propriété elle est la valeur. Mais quand bien même nous saurions la propriété dont il s'agit, encore nous faudrait-il pour que la valeur prenne vraiment sens que nous sachions à quelle ressource l'attribuer. En revanche, un fois que nous savons que « Victor Hugo » est la valeur de la propriété « nom d'auteur » de la ressource « *Les Misérables* », nous pouvons dire que nous savons qui est l'auteur des *Misérables*. Quel est alors le statut, du point de vue de la connaissance, de la seule valeur « Victor Hugo » ? de la seule métadonnée ? Quel est le statut épistémologique de la métadonnée ? Et comment participe-t-elle à augmenter notre savoir sur la ressource ? Que faut-il pour comprendre une métadonnée ? Pour aborder ces questions, nous allons nous intéresser au modèle « données-informations-connaissances », ou « données-informations-connaissances-sagesse » dans une version étendue, modèle issu du champ de la Gestion des Connaissances et qui décrit la création de connaissances.

Frances Wright[3] écrit en 1829, dans son livre *Course of popular lectures*, « La connaissance veut dire des choses connues. Là où aucune chose n'est connue, il ne peut y avoir de connaissance. Nous avons constaté que chaque science, c'est-à-dire chaque branche de la connaissance, est constitué de faits précis, dont nos sensations nous fournissent la preuve. Lorsque cette preuve fait défaut, nous sommes dénués de données ; nous sommes dénués des premières prémisses ; et quand, sans celles-ci, nous tentons de bâtir une science, nous faisons comme ceux qui élèvent des édifices sans fondations. Et que bâtissent de tels constructeurs ? Des châteaux de sable [*castles in the air*] ». Ces prémices d'un modèle reliant donnée et connaissance, anticipent à la fois la représentation de ce modèle qu'en proposent Awad et Ghaziri[4] sous forme de pyramide (solidement établie sur une base de données), mais aussi le

[3] Nous empruntons ces quelques éléments biographiques à la notice Wikipédia de Frances Wright : « Frances Wright (6 septembre 1795 – 13 décembre 1852), également connue sous le nom de Fanny Wright, était une libre-penseuse, féministe et abolitionniste d'origine écossaise qui devint américaine en 1825. Grande admiratrice de La Fayette dont elle devient la maîtresse bien que 38 ans les séparent. Inspirée par les expériences utopiques de Robert Owen, elle fonda en 1825 dans le Tennessee la commune de Nashoba, une communauté multiraciale destinée à démontrer les vertus émancipatrices de l'éducation sur les esclaves ». Disponible sur : <https://fr.wikipedia.org/wiki/Frances_Wright> (consulté le 20 août 2018)

[4] AWAD, Elias M. et GHAZIRI, Hassan M. *Knowledge Management*. Upper Saddle City, NJ : Pearson Educational International, 2004, 480 p.

ton de Russell Ackoff au tout début de son article de 1989 *From Data to Wisdom*, qui est l'un des premiers à lier données et sagesse : « Une once d'information vaut une livre de données. Une once de connaissance vaut une livre d'information. Une once de compréhension vaut une livre de connaissance »[5] (4, Ackoff). Nous voyons qu'au-delà des données, de l'information et de la connaissance, Ackoff introduit la notion de compréhension. Comment définit-il ces notions ? « Les données sont des symboles représentant les propriétés d'objets et d'événements. L'information se compose de données traitées, le traitement visant à accroître leur utilité. Par exemple, les agents recenseurs collectent des données. Le Bureau de Recensement traite les données, ce qui les transforme en informations qui sont présentées dans les nombreux tableaux publiés dans les *Statistical Abstracts*. Comme les données, l'information représente également les propriétés d'objets et d'événements, mais de manière plus compacte et plus utile que les données. La différence entre données et information est fonctionnelle, pas structurelle.

L'information est contenue dans des *descriptions*, des réponses à des questions qui commencent par des mots comme qui, quoi, quand, où, et combien. La *connaissance* est transmise par des enseignements, des réponses aux questions pratiques [*how-to*]. La *compréhension* est transmise par des explications, des réponses aux pourquoi » (4, Ackoff).

Richard Gartner propose dans son livre *Metadata* une telle représentation pyramidale des notions d'Ackoff, en ajoutant un niveau supplémentaire pour la culture. Faisant remonter à Platon et à sa définition, dans le *Théétète*, de la connaissance comme « croyance vraie justifiée », les nombreux débats qui ont émaillé l'histoire intellectuelle sur ce qu'est la connaissance, Gartner explique que « l'idée que la connaissance dérive de l'information s'est bien sûr révélée déterminante dans le domaine des sciences de l'information où de nouveaux modèles ont émergé, qui visent à éviter de l'associer inextricablement à la croyance. L'élément clé ici sont les tentatives d'obtenir quelques lumières sur la relation entre données, information, connaissance, compréhension et même peut-être sagesse. L'une des plus marquantes est connue sous le nom de *Pyramide d'Ackoff*, du nom du théoricien des organisations Russell Ackoff ; il fut celui qui, dans un article grandiosement intitulé *From Data to Wisdom*, a défini une pyramide à partir de ces concepts imbriqués en les empilant les uns sur les autres.

La pyramide d'Ackoff part à sa base de son plus petit composant, une simple donnée. Elle a très peu de sens en elle-même. Une simple cellule d'un tableur, par exemple, n'est qu'un nombre, anodin, à moins que nous ne sachions quelque chose de son contexte. Nous pouvons glaner ce dernier en observant ses relations à d'autres composants de la même feuille, comme le libellé de la colonne dans laquelle il se trouve. Une fois qu'un peu de sens est inféré de l'observation des relations entre ces éléments, nous commençons à nous élever des données à l'information, le niveau

[5] Nous traduisons.

suivant dans la pyramide. L'information peut être pensée comme des données organisées, arrangées de telle manière qu'elles peuvent répondre aux questions de base sur leur monde.

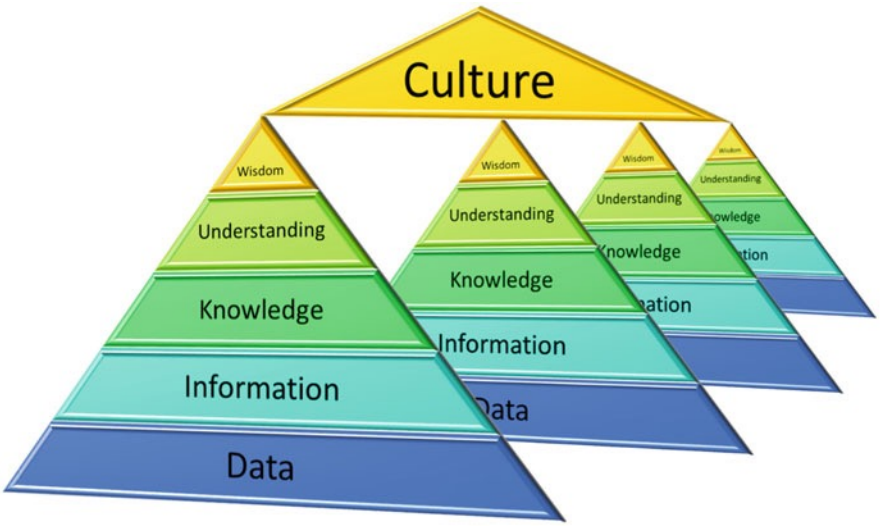

Figure 1 : La pyramide d'Ackoff, surmontée d'une couche pour la culture (R. Gartner, 2016)

Le niveau suivant à partir de celui-là, la connaissance, intervient quand plusieurs informations sont rassemblées et que des liens significatifs sont faits entre leurs composantes ; le tout devient alors plus que la simple somme de ses constituants. Pour que cela puisse advenir, des structures [*patterns*] qui soient elles-mêmes significatives doivent émerger ; la forme de ces structures et leurs relations constituent les fondements de la connaissance. Une autre manière de dépeindre la connaissance est de la voir comme de l'*information en contexte* : une information n'est plus une assertion isolée sur le monde mais acquiert un nouveau sens en interagissant avec ses homologues.

Interpréter et analyser ces structures afin que de nouvelles puissent émerger nous permet de nous élever à un niveau supérieur, de la connaissance à la compréhension. Beaucoup de recherches académiques tentent de faire précisément cela, faire naître de la compréhension en observant des structures dans la connaissance et en inférant à partir de celles-ci. La route vers le niveau suivant, de la compréhension à la sagesse, est plutôt difficile à cerner : c'est là que les notions de bien et de mal, ou de meilleure alternative possible, entrent en jeu.

Sans surprise, tandis que les sciences de l'information sont généralement à l'aise avec les quatre premiers niveaux de la pyramide, essayer de définir, sans parler de modéliser, la sagesse, les rend plus réticentes.

Une autre manière d'envisager cette pyramide est de considérer chaque niveau comme offrant d'éventuelles réponses à un type de question différent. L'information nous permet de répondre aux questions concises comme "qui ?", "quoi ?", "quand ?", "où ?". La connaissance nous permet de répondre à "comment ?", notamment comment telles choses (concrètes ou abstraites) fonctionnent-elles ?". La compréhension nous permet de demander "pourquoi ?", notamment "pourquoi ces

choses sont telles qu'elles sont ?". Les données seules, à la base de la pyramide, ne peuvent répondre par elles-mêmes à aucune de ces questions : c'est seulement en les agrégeant dans les niveaux situés plus haut que cela devient possible »[6] (1, Gartner, pp. 8-9). Il nous semble que l'intérêt de ce modèle réside surtout, à ce point de notre examen, dans ce qu'il permet de fixer un vocabulaire pour décrire différents niveaux de savoir, bien mis en évidence par les types de question auxquels ils répondent, mais ne permet pas de rendre compte concrètement du passage d'un niveau à l'autre de la pyramide, les notions d'inférence ou de contexte restant assez floues.

Toutefois, si nous nous sommes tournés vers l'ouvrage de Richard Gartner pour présenter ce modèle, c'est parce qu'il examine le rôle que les métadonnées y jouent. Il poursuit effectivement en expliquant que « dans ce modèle d'information, connaissance, compréhension et (peut-être) sagesse, les métadonnées ont un rôle central en permettant à l'édifice d'Ackoff d'être bâti à partir des unités infinitésimales de donnée à la base de la pyramide. En fait, s'élever d'un niveau à l'autre repose de façon cruciale sur la constitution de liens entre les composantes de l'un pour créer les agrégats qui deviennent les unités du suivant A chaque étape, les métadonnées ont un rôle vital à jouer en tant que "colle" avec laquelle ces liens sont formés » (ibid.). Les métadonnées apparaissent donc ici comme un élément qui permet de comprendre comment peut s'effectuer le passage d'un niveau à l'autre. Mais ont-elles alors un rôle équivalent entre les différents niveaux ? De quelles métadonnées parle-t-on ? « A la base de la pyramide, transformer les données en information suppose de les mettre en relation avec d'autres données. Cette "relation" est une sorte de métadonnée, quelque nom qu'on lui donne : il s'agit de dire quelque chose de deux données élémentaires, même si ce "quelque chose" n'est pas clair. Dans notre tableur qui transforme des données en information, la localisation de la cellule et sa place par rapport aux autres dans la même feuille nous dit quelque chose à propos de la donnée, ce qui en fait indéniablement "une donnée à propos de donnée". Il y a ici une analogie évidente avec la linguistique : les théories "structuralistes" de Ferdinand de Saussure[7] allèguent que le signe acquiert son sens par sa place dans une structure plus large. Les métadonnées sont ici la structure qui donne à nos signes, les données élémentaires, leur sens premier » (ibid.). Nous pouvons nous demander s'il n'y a pas ici une contradiction entre ce que dit Ackoff lorsqu'il affirme que « la différence entre données et information est fonctionnelle, pas structurelle », et ce que dit Gartner lorsqu'il considère que c'est la place de la donnée dans une structure plus large qui en fait une information. De fait, la valeur qui est inscrite dans la cellule du tableur n'est pas structurellement modifiée par le fait de nommer la colonne dans laquelle elle se trouve. En revanche, elle peut remplir d'autres rôles, ou commencer à remplir un rôle, en termes de savoir. Retenons donc ici que ce sont les métadonnées, par leur rôle

[6] Nous traduisons.

[7] L'auteur fait ici référence au *Cours de linguistique générale*, œuvre posthume de Ferdinand de Saussure publiée en 1916.

structurant, qui permettent le passage des données, qui par elles-mêmes ne sont pas porteuses de sens, à l'information.

L'auteur poursuit en remarquant qu'« en allant de l'information à la connaissance, le rôle des métadonnées est encore plus explicite. Comme nous l'avons vu, il s'agit à cette étape de mettre l'information en contexte et de forger des liens entre ses composantes pour générer de nouvelles significations. Toute métadonnée peut y parvenir, mais c'est particulièrement vrai de celles qui sont descriptives. Attacher une telle métadonnée à un composant d'information établit immédiatement une connexion à d'autres de son genre. Marquer sur un livre le nom de son auteur, par exemple, forge un lien à d'autres livres du même auteur. Ces liens sont *sémantiques* : ils manifestent un certain sens sur les relations entre les informations qu'ils relient. Un réseau de tels liens permet rapidement de commencer à répondre aux "comment ?" qui sont du domaine de la connaissance.

Le passage de la connaissance à la compréhension est un autre niveau d'abstraction de ce réseau de métadonnées : comme nous l'avons vu, comprendre est un processus d'analyse des structures de la connaissance pour en dériver d'autres de plus haut niveau qui répondent à de nouveaux types de question, comme les "pourquoi ?". Nous pourrions, si nous étions particulièrement pédants, les appeler méta-métadonnées, puisqu'il s'agit essentiellement de dire quelque chose des métadonnées du niveau immédiatement inférieur. De la même manière, quand nous allons vers le niveau plus tendancieux de la "sagesse", nous utilisons de nouveaux agrégats de métadonnées, visant à répondre à des questions de bien et de mal, au-dessus de ce niveau de la compréhension » (ibid.). Le rôle structurant des métadonnées que nous notons plus haut doit être donc complété de leur caractère structuré, et c'est précisément à mesure qu'elles le sont de manière de plus en plus complexe qu'elles permettent de progresser dans les niveaux de la pyramide.

Richard Gartner, s'il paraît douter de la légitimité d'introduire dans cette pyramide la notion de sagesse[8], est beaucoup plus convaincu et convaincant sur la continuité avec la culture elle-même. Il introduit à ce sujet l'intéressante notion de *curation*, qu'il serait hâtif de traduire par « conservation ». Avant de proposer une traduction de ce terme, examinons ce qu'il en dit lui-même : « les métadonnées ont clairement un rôle crucial en nous permettant d'agréger les données en connaissance et en compréhension (et peut-être même en sagesse). Mais on peut dire qu'il existe un niveau encore plus élevé au-dessus de ceux de la pyramide d'Ackoff, celui de la culture elle-même. Ici, à nouveau, les métadonnées ont un rôle à jouer.

[8] Nous pouvons partager ses doutes. Il nous semble en effet qu'il existe un hiatus entre les jugements de réalité de la connaissance et les jugements de valeur propres aux considérations sur le bien et le mal. Remarquons toutefois que le modèle tel qu'il est formulé par Ackoff nous semble orienté par des questions d'utilité, de rapport pragmatique au savoir, et que de ce point de vue, une continuité peut être réintroduite : le choix d'une « meilleure option », dans une situation délicate mettant en jeu des questions de valeur (au sens de valeurs morales), peut s'appuyer sur des connaissances concernant la marche du monde. C'est ainsi que dans les « comités d'éthique », les « savants » figurent généralement en bonne place.

Une culture évolue sans arrêt mais cette évolution repose sur la conservation et la transmission et des précédentes manifestations : aucune culture n'apparaît spontanément sans référence à celles-ci, même dans le cas des rébellions contre le passé. Afin que ce soit possible, une culture doit être *curated*, littéralement 'soignée' comme l'étymologie du mot (du latin *curare*) l'indique » (ibid.). A présent que nous voyons mieux ce que l'auteur entend par *curated*, il nous semble que la notion d'*entretien* peut traduire son idée. « L'entretien est souvent confondu avec la conservation, mais c'est beaucoup plus que cela. L'entretien suppose d'identifier les éléments d'une culture qui la définissent particulièrement bien et choisir ceux qui sont importants ; puis de décrire et d'ajouter le contexte de ceux-ci, en faisant des connexions entre eux, afin qu'ils puissent être compris par ceux qui s'y intéressent. Enfin, cela suppose de disséminer une culture, en la rendant accessible. Tout ceci vient en plus du fait de s'assurer que ces éléments vont continuer à exister longtemps dans le futur. Accomplir ces étapes assure par-dessus tout qu'une culture pourra continuer à être comprise en étant transmise à travers les générations. C'est grâce aux efforts d'entretien de nos prédécesseurs que n'importe quelle culture au-delà des plus éphémères peut exister d'une manière ou d'une autre.

Au cœur de l'entretien se trouvent les processus d'organisation et de description ; que nous entretenions un artefact inestimable dans un musée ou un fil Twitter dans une archive numérique, ces deux processus sont essentiels pour mettre la culture en contexte. Le faire nécessite des métadonnées, les mêmes qui sont nécessaires pour condenser l'information en connaissance mais dans ce cas avec un accent particulier sur la classification, la construction d'articulations et l'explication. Ce nouveau rôle des métadonnées peut être vu comme un niveau additionnel en haut de la pyramide d'Ackoff, assemblant de multiples agrégats de connaissance, compréhension voire sagesse en une culture plus large »(ibid.).

Le modèle que nous venons d'exposer, que nous pourrions qualifier d'« humaniste » par l'attention qu'il porte à réinscrire le processus de connaissance dans la culture, n'est pas la seule manière de concevoir les rapports entre données, informations et connaissances. Comme Gartner lui-même le note, de nombreux débats existent dans les sciences de l'information sur ce modèle. Chaïm Zins, dans son article *Conceptual Approaches for Defining Data, Information and Knowledge*, se base sur cent-trente définitions des termes « donnée », « information » et « connaissance » produites par quarante-cinq universitaires dans le champ des Sciences de l'Information pour proposer une sorte de cartographie des différentes approches conceptuelles de ces concepts fondamentaux de la discipline. Comme l'auteur le remarque au début de son article, « évidemment, les trois concepts clés sont interconnectés, mais la nature des relations entre eux est sujette à débat, de même que leur signification »[9] (5, Zins, p. 479). Il part du constat que beaucoup d'universitaires placent ces concepts dans un ordre séquentiel, les données étant le

[9] Nous traduisons.

matériau brut des informations, qui seraient le matériau brut des connaissances. Il se demande alors pourquoi les Sciences de l'Information étudient également l'organisation et la gestion des connaissances, si les connaissances sont un au-delà de l'information : « devrions-nous réfuter l'ordre séquentiel ? Devrions-nous changer le nom du champ disciplinaire, de *Sciences de l'Information* à *Sciences de la Connaissance* ? Ou devrions-nous prendre le parti extrême d'exclure les deux sous-champs de l'Organisation des Connaissances et de la Gestion des Connaissances des Sciences de l'Information ? » (5, Zins, p. 479). Après avoir présenté les différentes définitions qu'il a recueillies, Chaïm Zins pose plusieurs distinctions conceptuelles préparant son analyse, tout d'abord sur différents types de connaissance : « Dans l'épistémologie traditionnelle, il y a trois formes principales de connaissance : la connaissance pratique, la connaissance directe [*knowledge by acquaintance*] et la connaissance propositionnelle [*propositional knowledge*]. La connaissance pratique désigne des savoir-faire. La connaissance directe est la reconnaissance directe immédiate d'objets et d'organismes physiques externes […] ou la reconnaissance directe d'un phénomène interne […] La connaissance propositionnelle […] est le contenu réflexif et/ou exprimé de ce qu'une personne pense qu'il ou elle sait […] La connaissance propositionnelle est divisée entre connaissance inférentielle et connaissance non-inférentielle. La connaissance propositionnelle non-inférentielle désigne la compréhension intuitive directe d'un phénomène […] La connaissance inférentielle est le produit d'inférences, comme l'induction et la déduction » (5, Zins, p. 486). Retenons que la connaissance propositionnelle s'exprime le plus généralement sous la forme « je sais que… ». L'auteur s'attache ensuite à une autre distinction conceptuelle d'importance pour son analyse, entre le « domaine du subjectif » et de « domaine de l'objectif », termes qu'il entend en un sens spécifique : « les phénomènes que sont données (D), informations (I), connaissances (C), ont deux modes d'existence distincts, à savoir l'un dans le domaine du subjectif et l'autre dans le domaine de l'objectif […] *connaissance subjective* est ici équivalent à la connaissance du sujet ou du sachant individuel, et *connaissance objective* est équivalent ici à la connaissance comme objet ou comme chose » (5, Zins, p. 486). Dans le sens où l'entend l'auteur, une connaissance publiée dans un livre est une connaissance objective, non pas parce qu'elle serait vérifiable, indépendante de l'opinion ou que tout un chacun pourrait la reconnaître comme vraie, mais parce qu'elle se trouve formulée sur un livre, qui est un objet. L'auteur adopte d'ailleurs « pour éviter la confusion » le terme de *connaissance universelle* plutôt que celui de connaissance objective. Muni de cette distinction, il envisage successivement la trilogie donnée-information-connaissance dans chacun des deux domaines, subjectif et universel. Dans le domaine subjectif, les données sont des stimuli sensoriels ou leur signification[10] (un bruit de moteur de voiture), l'information est une connaissance

[10] La psychologie cognitive différencie la sensation de la perception, la sensation désignant les seuls stimuli et la perception désignant un premier traitement cognitif qui donne aux stimuli une signification. Typiquement, mon nerf optique reçoit des stimuli visuels, mais je *perçois* qu'un ballon s'approche de

empirique (le moteur tourne et la voiture est en train de partir), la connaissance est la pensée dans l'esprit d'un individu, « caractérisée par la croyance justifiable de l'individu que c'est vrai » (5, Zins, p. 487). Pour l'auteur, l'information n'est pas un stade intermédiaire entre les données et la connaissance, mais déjà une forme de connaissance. Cette connaissance, en tant qu'elle s'appuie sur l'expérience ou l'observation, est dite empirique. Ce que l'auteur appelle « connaissance » renvoie à une croyance étayée par des arguments qu'un certaine pensée est vraie. Dans le domaine objectif, la trilogie donnée-information-connaissance désigne des « artéfacts humains ». L'auteur les définit de la même manière que dans leur versant cognitif, à ceci près qu'il s'agit d'« ensembles de signes » : une donnée est un ensemble de signe représentant un stimuli ou une perception, une information est un ensemble de signes représentant une connaissance empirique, une connaissance est un ensemble de signes représentant le sens ou le contenu de pensées que l'individu croit de manière justifiable être vrai. L'auteur utilise cette notion d'« ensemble de signes », plutôt que de parler de « signification », dans la mesure où nous n'avons pas accès à la *signification* d'une proposition écrite dans un livre, mais à ce qui *représente* cette signification. Une fois développé ce cadre conceptuel, l'auteur peut examiner les différentes définitions qui lui ont été soumises. Aucune ne relève d'une approche métaphysique de la trilogie donnée-information-connaissance — « comme "la connaissance est éternelle", et "la connaissance est une entité/un objet indépendant" » (5, Zins, p. 487). Elles sont presque toutes « exclusivement humaines » [*human-exclusive*] (à part l'un d'entre elles qui fait référence à « un organisme ou un agent intelligent »). Parmi ces approches centrées sur l'être humain, presque toutes sont « basées exclusivement sur le cognitif » (par opposition à des approches intégrant en plus du cognitif le biologique et le physique dans leur conception de la trilogie). Parmi les approches basées sur le cognitif, la plupart sont des approches « exclusivement propositionnelles » — nous sommes renvoyés à la distinction opérée plus haut : « bien que le panel ne se réfère pas spécifiquement aux différents types de connaissance, une distinction doit être opérée entre le fait de se concentrer sur la connaissance propositionnelle et celui d'aborder tous les types de connaissance » (5, Zins, p. 488). Ayant ainsi cerné le courant dominant des conceptions exprimées, Chaïm Zins distingue cinq modèles selon que données, informations ou connaissances sont considérées comme ressortissant du domaine du subjectif ou du domaine de l'universel.

moi. Ou pour reprendre l'exemple de l'auteur, je reçois des stimuli auditifs divers parmi lesquels j'isole ceux qui correspondent au bruit d'un moteur qui tourne.

Model 1		Model 2		Model 3		Model 4		Model 5	
UD	SD	UD	SD	UD	SD	UD	SD	UD	SD
D		D		D		D	D	D	D
I			I	I	I	I	I	I	I
	K		K	K	K		K	K	K

Figure 2 : Cinq modèles pour définir donnée (D)-information(I)-connaissance (C) (C. Zins, 2007)

Nous pouvons remarquer que dans les différents modèles, les données sont toujours considérées comme ressortissant (au moins) du domaine de l'universel (c'est-à-dire considérées comme une représentation d'une perception) tandis que les connaissances sont toujours considérées comme ressortissant (au moins) du domaine du subjectif (c'est-à-dire comme une pensée dans l'esprit d'un individu, qui a des raisons de la croire vraie). Si nous appliquons ce cadre conceptuel aux métadonnées, nous ne pouvons qu'abonder dans le sens d'une conception des métadonnées comme représentation, en l'occurrence représentation d'une ressource. Toutefois, leur statut en « données à propos de données » semble problématique dans le modèle D-I-C, si le lieu de la réflexivité est celui non pas des données mais de la connaissance propositionnelle. Faut-il alors abandonner l'ordre séquentiel, comme l'auteur le suggérait au début de son article, sans y donner suite ?

Le modèle D-I-C tel que nous l'avons vu présenté est fortement marqué par la tradition philosophique de l'empirisme : la connaissance y dérive tout entière de l'expérience sensible. Les données y sont présentées comme des données sensibles. Toutefois, Richard Gartner, pour nous faire comprendre le modèle, prend l'exemple d'une cellule de tableur. Il y a certes une appréhension sensible du contenu de cette cellule quand on la lit, mais cette donnée est elle-même le résultat d'une construction. C'est ce qui fait dire à Bruno Latour, dans son article *Pensée retenue, pensée distribuée* : « Il faudrait remplacer ce terme par celui, beaucoup plus réaliste, d'"obtenues" et parler par conséquent de "bases d'obtenues", de "sublata" plutôt que de "data" pour parler à la fois en latin et en anglais »[11]. La notion de donnée peut désigner à la fois une « donnée immédiate de la conscience », pour reprendre l'expression de Bergson, mais aussi la donnée d'un problème ou la donnée issue d'un capteur. Parler de « donnée », c'est à la fois parler de quelque chose d'élémentaire et dire que nous n'allons pas interroger cette chose élémentaire mais construire à partir d'elle. Il peut néanmoins y avoir de bonnes raisons d'interroger des données, par exemple dans une démarche de qualité des données, pour comprendre comment la « donnée » a été obtenue, ou même — dans le cas des données sensibles — parce que l'expérience nous apprend que notre perception elle-même peut être modifiée par de nouvelles connaissances. En ce sens, c'est la linéarité du modèle D-I-C qui nous paraît devoir être remise en cause. Il est tout à fait légitime de hiérarchiser différents niveaux de savoir entre données élémentaires, repérage des similitudes entre

[11] LATOUR, Bruno. *Pensée retenue, pensée distribuée* In JACOB, Christian (dir.), *Lieux de savoir*, Tome I. Paris : Albin Michel, p. 609.

certaines de ces données, articulation en structures plus complexes, compréhension des relations entre ces structures, voire modélisation de corrélation entre ces relations. En revanche, il nous semble qu'à chacune de ces étapes, les mouvements ascendants ou descendants sont possibles et légitimes selon les circonstances. La construction d'un modèle peut nous amener à détecter une donnée aberrante, à interroger la manière dont cette donnée a été obtenue, à constater qu'elle est juste (à partir d'autres données et d'autres connaissances, donc) et à devoir amender voire reconstruire l'édifice de connaissance que nous avions bâti. Pour conserver les notions de données, d'information et de connaissance, nous serions tentés de considérer, plutôt qu'une pyramide, des boucles d'acquisition de données donnant lieu à des boucles d'acquisition d'information, elles-mêmes pouvant donner lieu à des boucles d'acquisition de connaissance, chaque boucle pouvant donner passage vers le niveau de boucle supérieur ou vers le niveau de boucle inférieur.

Dans cette perspective, les métadonnées peuvent être pensées de la même manière que les données en général, alors qu'elles se plient mal au modèle D-I-C. Richard Gartner proposait la solution ingénieuse de voir, dans les métadonnées, la « colle » qui permettait de passer des données aux informations et des informations aux connaissances, mais il nous semble que seul le premier passage, des données aux informations, est vraiment compréhensible, et qu'il est lié à la confusion que nous avons notée en 1.1.1 entre la métadonnée comme propriété et la métadonnée comme valeur : pour reprendre son exemple du tableur, on peut considérer que la connaissance du nom de la colonne du tableur permet de faire, des valeurs contenues dans les cellules, des informations. Mais comme nous l'avons vu, le terme de métadonnée désigne à la fois les valeurs des cellules et l'intitulé de la colonne. En revanche, la vision que propose Richard Gartner des métadonnées permettant l'entretien, le soin, la *curation* de la culture, nous semble tout à fait fondamentale et précieuse. Si l'auteur se place sur le long terme, cette conception a également tout son sens dans le cadre de la réutilisabilité des jeux de données grâce aux métadonnées.

1.1.3. Extension des métadonnées : où trouve-t-on des métadonnées ?

A présent que nous avons posé quelques jalons concernant la signification, l'*intension* - pour employer le vocabulaire de la logique théorique - du concept de métadonnée, nous souhaitons en montrer l'*extension*. La définition que nous avons retenue des métadonnées comme propriété (ou valeur de cette propriété) caractérisant une ressource ouvre en droit un champ d'application infini. En pratique, il est effectivement extrêmement étendu et nous n'en donnerons ici que quelques exemples qui permettront de mesurer cette étendue et de poser quelques distinctions utiles.

Le site de DoRANum[12] propose une infographie intitulée « Les métadonnées sont partout ! »[13] qui commence par rappeler que « les métadonnées peuvent être présentes de manière embarquée, par exemple dans un fichier informatique (photo, logiciel, document, …) [ou] de manière externe, par exemple dans un catalogue d'accompagnement d'un jeu de données ou dans un annuaire d'entrepôts », pour préciser ensuite qu'il sera seulement question des métadonnées embarquées. Le document vise à sensibiliser précisément sur cette existence, invisible pour l'utilisateur, des métadonnées embarquées. Celles-ci sont « créées automatiquement dans nos activités numériques quotidiennes : dans les documents texte ; […] dans les mails ; […] dans les photos ». Une photographie numérique peut ainsi être accompagnée, nous le savons, des coordonnées GPS de l'endroit où elle a été prise, mais également de la date, de la marque de l'appareil, de son numéro de série, et dans le cas d'une photographie retouchée, d'une miniature de la photographie originale. Dans le cas d'un mail, les métadonnées embarquées ne sont pas invisibles, par exemple « objet, expéditeur, destinataire, date d'envoi, personnes en copie ». En revanche, l'utilisateur ne sait pas, et la différence est ici plus subtile, que ces éléments du document existent sous forme de métadonnées embarquées dans le document, tout en les utilisant lorsqu'il effectue une recherche pour retrouver un mail dans sa messagerie. Dans un traitement de texte comme *Word* ou *OpenOffice Writer*, l'onglet Fichier > Propriétés permet d'accéder et de modifier des métadonnées d'un document avant de les partager. Le message principal de l'infographie de DoRANum est en effet d'avertir les utilisateurs qu'en transmettant leurs fichiers, ils transmettent - sans toujours le savoir - les métadonnées embarquées, éventuellement à leurs dépens, et de leur fournir quelques éléments pour nettoyer, vérifier, modifier ces métadonnées.

Le web est également un espace dans lequel les métadonnées sont omniprésentes. Comme le rappelle Gérard Puimatto, dans son article *Les métadonnées : pourquoi et pour quoi faire ?*[14], « dès l'émergence du html 1.0 (1990), les balises <meta…> sont disponibles pour être utilisées dans une zone d'entête (<head>), non affichée par les navigateurs. A cette période, chaque webmestre utilise les métadonnées pour conserver avant tout les informations qui sont utiles au fonctionnement de son site, comme les liens sur des composantes d'interactivité, des images de fond, etc. Si quelques termes viennent progressivement jalonner ce nouvel espace (*author*, *creator*, *keywords*, etc.), aucun standard ne permet d'homogénéiser les traitements » (6, Puimatto, p. 5). Dans un premier temps, donc, les métadonnées ont un usage, sinon privé, du moins limité à leurs créateurs, dans un domaine où pourtant la nécessité de standards concernant l'adressage physique s'est d'emblée

[12] Pour « Données de la Recherche : Apprentissage NUMérique à la gestion et au partage » ; il s'agit d'un projet associant la Bibliothèque Scientifique Numérique, l'Inist-CNRS et le réseau des Urfist.
[13] DoRANUM. *Les métadonnées sont partout*. [En ligne]. <https://doranum.fr/les-metadonnees-sont-partout/> (consulté le 20 août 2018).
[14] https://www.reseau-canope.fr/savoirscdi/fileadmin/fichiers_auteurs/Societe_de_l_information/Tic_et_documentation/Les_metadonnees_Puimatto.pdf

imposée comme une évidence. Comme le remarque l'auteur, « rapidement cependant, la réflexion s'organise autour de la nécessité de donner une approche sémantique à l'organisation du web, […] et donc d'établir des standards permettant de s'organiser à l'échelle du web. […] Après divers travaux relevant davantage de l'investigation et de structures exploitées par exemple au sein d'une université, un groupe de travail réuni à Dublin (Ohio) définit en 1995 un jeu de métadonnées destiné à la description des documents officiels du gouvernement fédéral des États-Unis. L'initiative *Dublin Core* est née. C'est le point de départ d'une déjà longue aventure, avec une gamme de structures généralistes ou spécialisées qui se diffusent largement sur le web » (6, Puimatto, p. 5). Cet épisode de l'histoire du web permet de mettre en lumière comment se fait jour le besoin de standards pour les métadonnées - question sur laquelle nous reviendrons plus loin - à l'intérieur d'une communauté. Il présente également le schéma de métadonnées *Dublin Core*, utilisé pour décrire des documents. Enfin, pour prolonger ce que dit l'auteur, cet épisode ouvre sur la question de l'usage des métadonnées d'une page web dans une optique de référencement par les moteurs de recherche.

Outre les exemples de la vie quotidienne ou du web, les métadonnées sont susceptibles de se trouver dans tous les systèmes d'information spécialisés. Nous ne pourrions en faire une recension exhaustive, mais nous pouvons prendre l'exemple des systèmes d'information géographiques (SIG), domaine dans lequel il est particulièrement crucial de disposer de métadonnées pour que les données puissent avoir du sens pour tout autre utilisateur que leur créateur. Ainsi, une carte doit être accompagnée de métadonnées précisant l'étendue couverte, l'échelle, le type de projection utilisée ou encore la symbologie, pour ne citer que les métadonnées les plus spécifiques à ce type d'objet. Dans le cas des disciplines spécifiques qui s'entendent sur un ensemble de métadonnées pour décrire les ressources qui les intéressent, ces métadonnées sont généralement structurées selon un schéma qui les désignent comme un ensemble. Ainsi, on parlera du schéma de description *Learning Object Metadata* pour les ressources d'enseignement et d'apprentissage, qui comporte bien entendu des métadonnées caractéristiques de l'aspect pédagogique d'une ressource ; on parlera aussi du format de description *Text Encoding Initiative*, qui recourt à un langage de balisage à la fois formel et sémantique des textes.

Nous nous devons également de citer les métadonnées relatives au « patrimoine culturel », c'est-à-dire les métadonnées des musées, des bibliothèques et des archives. Les communautés de professionnels gérant les objets informationnels et les collections physiques abritées par ces institutions ont en effet développé à la fois des schémas de métadonnées structurés, des vocabulaires contrôlés et des thésaurus, des règles de catalogage, des standards d'échange de métadonnées, permettant de collecter, de traiter, de spécifier et de diffuser les données de manière très riche et très précise.

À la question de savoir où l'on trouve des métadonnées, nous voyons que la réponse doit être déclinée selon plusieurs dimensions, selon qu'on s'intéresse au lien

entre la ressource et les métadonnées qui la spécifient, ou aux cas d'usage de métadonnées. Dans la première dimension, nous avons vu que les métadonnées peuvent être soit encapsulées dans le document qu'elles décrivent, soit externes, et dans ce dernier cas soit fournies parallèlement à la ressource, soit présentes dans une base de données (il faudrait ajouter le cas particulier des métadonnées « englobantes », qui contiennent la ressource elle-même, comme dans le cas du standard d'archives *Encoded Archival Description* – EAD). Dans la deuxième dimension, nous nous sommes limités à quelques exemples, mais le principe est simple : partout où il y a des ressources à décrire, il est susceptible d'y avoir des métadonnées. C'est bien entendu cette deuxième dimension qui était d'abord visée par notre interrogation et nous espérons avoir montré sinon toute l'étendue des usages, du moins une variété de ceux-ci, permettant de donner une idée de leur étendue. Ajoutons enfin une autre dimension, qui relève plutôt de la sociologie des usages : les métadonnées peuvent être partagées au sein d'une communauté (nous avons envisagé le cas de professionnels ayant une communauté de pratiques, mais les *tags* apposés par les contributeurs d'une folksonomie[15] sont également des métadonnées) ou permettre l'usage d'une ressource en dehors de la communauté qui y a associé lesdites métadonnées.

[15] Pour désigner cette « classification de contenus de l'internet par l'attribution de mots-clés librement choisis par un utilisateur », la Commission d'enrichissement de la langue française préfère parler d'« indexation personnelle » (JORF n°0300 du 27 décembre 2009 page 22539 - texte n° 71). Disponible sur : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000021530619> (consulté le 20 août 2018)

1.2. Quels usages pour quelles métadonnées ?

En nous intéressant aux lieux où se trouvent des métadonnées, nous avons été amenés à mettre en lumière la multiplicité des cas d'usages, mais également à envisager la multiplicité des métadonnées elles-mêmes. C'est l'évidence de la présence, voire de l'omniprésence, des métadonnées, sur laquelle nous souhaitons ici revenir. A quoi servent les métadonnées ? A quelles finalités répondent-elles ? Peut-on dégager des familles, des types de métadonnées ? Répondre à ces questions nous permettra de tracer plus distinctement les contours de ce que peut désigner l'enrichissement des métadonnées.

1.2.1. Finalités des métadonnées

Se poser la question de la finalité des métadonnées, c'est se demander quel besoin il y a de décrire des ressources. Une ressource qui n'est décrite d'aucune manière est une entité qui existe par elle-même, mais qui est d'une certaine manière close, refermée sur elle-même. Imaginons un fichier de tableurs ne contenant que des chiffres, un empilement de dossiers à la couverture vierge contenant des feuillets manuscrits, une collection de simples images radiographiques. A moins de réussir à se souvenir avec la plus grande précision du contenu de ces documents, nous pouvons immédiatement imaginer la difficulté à retrouver une image radiographique en particulier dans la collection, la difficulté à utiliser les données contenues dans un tableur par tout autre que son créateur, l'incertitude qu'un dossier, après de multiples déménagements, contiennent les mêmes feuillets qu'il contenait auparavant. Par ces exemples, nous voulons mettre en évidence les trois principales fonctions des métadonnées : l'accessibilité, l'interopérabilité et la conservation.

Jenn Riley, de la *National Information Standards Organization* (NISO)[16], propose dans le document *Understanding metadata* différents cas d'usage des métadonnées dans les systèmes d'information : « La **découverte** est peut-être l'usage le plus commun, des métadonnées structurées permettant aux utilisateurs de rechercher ou d'explorer afin de trouver des ressources ou une information d'intérêt. Plusieurs métadonnées sont utiles à l'**affichage** aux utilisateurs pour les aider à identifier ou à comprendre une ressource. L'**interopérabilité**, l'échange efficace de contenu entre systèmes, repose sur des métadonnées décrivant ce contenu de manière telle que les systèmes impliqués puissent efficacement profiler les documents

[16] Selon leur site : « NISO, la National Information Standards Organization, une association à but non-lucratif accréditée par le American National Standards Institute (ANSI), identifie, développe, entretient et publie des standards techniques pour gérer l'information dans l'environnement numérique actuel, en perpétuelle mutation. Les standards NISO s'appliquent à la fois aux technologies traditionnelles et nouvelles, et à l'information à travers tout son cycle de vie, de la création à la documentation, l'usage, la réaffectation, le stockage, les métadonnées et la conservation ». Disponible sur : <https://www.niso.org/what-we-do> (consulté le 20 août 2018)

reçus et les harmoniser à leurs structures internes. Les métadonnées permettent la **gestion d'objets digitaux** en fournissant l'information nécessaire pour restituer correctement un contenu digital ou fournir la version adéquate au besoin de l'utilisateur. La **conservation** est obtenue en créant des métadonnées qui permettent la vérification de l'intégrité du contenu après transfert et d'autres points caractéristiques, et en signalant le moment où des actions de conservation comme une migration de format ou un contrôle d'intégrité devraient être entrepris. Enfin, les métadonnées permettent la **navigation** entre les parties des items, par exemple d'une page ou d'une section à la suivante, et entre différentes versions d'objets, comme des résolutions différentes d'images photographiques »[17] (7, Riley, pp. 6-7). Ce que Jenn Riley entend par découverte est connexe de ce que nous appelons accessibilité. Il nous semble toutefois que nous pouvons effectuer une distinction entre accéder et découvrir, dans le rapport de l'utilisateur au document, suivant qu'il sait d'avance ce qu'il cherche ou pas. Nous serons amenés à revenir sur cette distinction. L'affichage (*display*) est également une question connexe de l'accessibilité, nous semble-t-il. Nous pouvons penser, dans le domaine des livres, à une première de couverture qui s'affiche sur l'écran d'un libraire et permet au client d'identifier qu'il s'agit bien du livre qu'il cherche. L'interopérabilité renvoie à une problématique de transfert de la ressource : c'est ce qui permet que la ressource puisse avoir une pertinence hors de la situation initiale où elle se trouve. La gestion d'objets digitaux relève certainement d'une pratique spécifique, mais du point de vue qui est le nôtre ici des finalités, il nous paraît qu'elle relève à certains égards des questions de conservation et à d'autres des questions juridiques relatives à la propriété intellectuelle des données et au droit à leur réutilisation[18]. La question de la conservation concerne quant à elle la pérennité de l'archivage des données, et se révèle capitale notamment dans un contexte d'évolution technologique rapide des supports physiques de l'information. Enfin, la navigation présente elle aussi une problématique connexe de celle de l'accessibilité, à un niveau de granularité différent puisqu'il s'agit d'accéder soit à des parties spécifiques d'un document, soit à des versions différentes mais réunies en une entité d'un même document. Remarquons une dimension de l'accessibilité qui n'est pas présente ici, sauf à la ranger dans la catégorie « affichage » : l'accessibilité aux handicapés. Dans le cas du livre, en France, le Syndicat National de l'Édition a missionné un groupe de travail qui rassemble des membres de l'interprofession, « Normes & standards », pour réfléchir sur les moyens d'améliorer l'accessibilité des livres pour les malvoyants. Un exemple simple peut être l'ajout de métadonnées de description des images. Il nous semble toutefois que notre tripartition initiale entre préservation, interopérabilité et accessibilité n'est pas fondamentalement remise en cause, même si la notion

[17] Nous traduisons et nous soulignons.

[18] Précisons que ces aspects juridiques concernant la propriété intellectuelle, les termes de licence, sont à distinguer du *Digital Rights Management* (DRM) ou gestion des droits numériques, qui concernent les mesures techniques de protection prises pour contrôler l'utilisation d'œuvres numériques. Dans le cas du livre numérique, les DRM sont également mentionnés dans les métadonnées.

d'accessibilité doit être déclinée selon au moins quatre dimensions : la capacité à retrouver ou à découvrir, la modalité d'affichage (dans laquelle nous inclurons donc l'accessibilité aux handicapés), la possibilité de naviguer à l'intérieur d'un document et la gestion des aspects juridiques, des droits d'accès à une ressource

Dans l'ouvrage *Introduction to metadata*, dirigé par Murtha Baca, du *Getty Research Institute*[19], le premier chapitre, d'Anne Gilliland, se propose également de passer en revue les rôles des métadonnées. Pointons tout de suite deux différences qui nous semblent intéressantes par rapport au document de Jenn Riley : d'abord le rôle des utilisateurs en tant qu'acteurs de l'organisation et de la description des ressources, et ensuite la conception des métadonnées comme s'inscrivant dans le cycle de vie considéré comme circulaire d'un objet informationnel et passant par les étapes décrites ci-après :

« *Création, versions multiples, réutilisation et recontextualisation d'objets informationnels :*

> Les objets entrent dans un système d'information numérique en étant créés numériquement ou en étant convertis dans un format numérique. De multiples versions du même objet peuvent être créées à des fins de conservation, de recherche, d'exposition, de diffusion, ou même à des fins de développement de produits. Certaines métadonnées administratives et descriptives peuvent et même devraient être incluses par le créateur ou celui qui numérise, particulièrement si la réutilisation est envisagée, comme dans un système de gestion des ressources numériques [*Digital Asset Management System*].

Organisation et description :

> Une fonction primordiale des métadonnées est la description et l'ordonnancement des objets originaux ou des items dans un répertoire ou une collection, ainsi que les objets informationnels relatifs à ces originaux. Les objets informationnels sont organisés automatiquement ou manuellement selon la structure du système d'information et peuvent inclure des descriptions produites par le créateur originel. Des métadonnées additionnelles peuvent être créées par des professionnels de l'information par des processus d'enregistrement, de catalogage et d'indexation, ou par d'autres via des folksonomies et d'autres formes de métadonnées d'utilisateurs.

Validation :

> Les utilisateurs examinent les métadonnées et d'autres aspects des informations qu'ils récupèrent afin de vérifier l'autorité et la fiabilité de ces ressources.

[19] Le *Getty Research Institute* est un institut de recherche émanant du *J. Paul Getty Trust*, institution culturelle privée basée à Los Angeles regroupant musée, bibliothèque spécialisée, collections d'art, finançant des recherches et publiant des ouvrages. Le *J. Paul Getty Trust* est l'institution culturelle privée la plus richement dotée au monde.

Recherche et extraction :

> De bonnes métadonnées descriptives sont essentielles pour que l'utilisateur puisse trouver et extraire des métadonnées et des objets informationnels pertinents. Les objets informationnels stockés localement ou dispensées virtuellement sont sujets à recherche et à extraction par les utilisateurs, et les systèmes d'information créent et entretiennent des métadonnées qui suivent les algorithmes d'extraction, les opérations des utilisateurs, et l'efficacité du système dans le stockage et l'extraction.

Utilisation et conservation :

> Dans le domaine numérique, les objets informationnels peuvent être sujets à différents usages au cours de leurs vies, processus au cours desquels ils peuvent également être reproduits ou modifiés. Des métadonnées liées aux annotations des utilisateurs, au suivi des droits et au contrôle de version peuvent être créés. Les objets numériques, particulièrement ceux qui sont nativement numériques, doivent également être les objets d'un régime de conservation continu et subir des processus tels que des migrations d'actualisation et des contrôles d'intégrité pour s'assurer de leur disponibilité permanente et pour documenter tout changement de l'objet informationnel qui aurait pu advenir durant le processus de conservation.

Disposition :

> Les métadonnées sont un composant clé dans la documentation de la disposition (par exemple l'enregistrement, la cession) d'objets originaux et d'items dans un répertoire, ainsi que de celle d'objets informationnels relatifs à ces originaux. Les objets informationnels qui sont inactifs ou ne sont plus nécessaires peuvent être éliminés »[20] (8, Gilliland, pp. 13-14).

Ce dernier usage des métadonnées peut sembler assez spécifique du domaine des collections d'art (le fait qu'un musée dispose ou non d'un objet). Nous verrons toutefois que nous pourrons lui trouver un équivalent dans le secteur du livre avec la notion de disponibilité à la vente (à paraître ou paru, paru et disponible, ou paru mais temporairement indisponible, etc.), et que nous pourrions lui trouver également des équivalents dans le cas d'un livre en bibliothèque (disponibilité au prêt ou à la consultation seulement, par exemple). Cette notion de disponibilité est à mettre en rapport avec la notion des droits d'accès à une ressource. Les droits d'accès renseignent (évidemment) sur *le droit*, quand la disponibilité renseigne sur *le fait*. Les deux répondent à la question : peut-on accéder à la ressource ? En ce sens, nous considérons également la disponibilité comme une dimension de l'accessibilité.

[20] Nous traduisons.

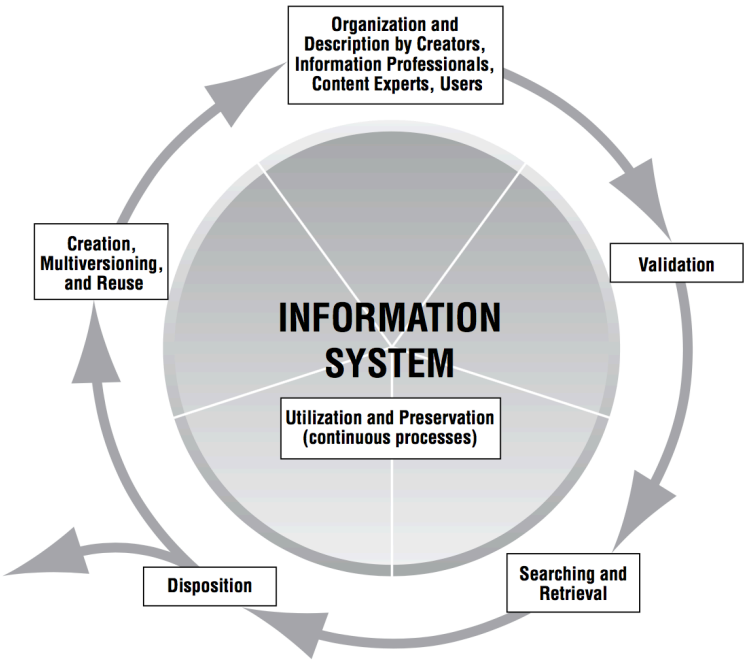

Figure 3 : Le cycle de vie d'un objet informationnel (M. Baca éd., 2008)

Comme nous le remarquions, l'accent est mis sur les utilisateurs, avec trois conséquences majeures : le rôle qu'ils jouent dans l'ajout de métadonnées descriptives, leur rôle dans la vérification de l'autorité et de la fiabilité d'une ressource, et leur statut de réutilisateurs des données, qui pousse à orienter le système d'information vers la réutilisation des métadonnées. Dans notre recensement des finalités des métadonnées, devons-nous ajouter cette idée de métadonnées de validation par les utilisateurs de la ressource ? Nous laissons pour l'instant cette question en suspens, qui ouvre sur un paradigme très différent de conception des métadonnées, mais nous serons amenés à y revenir quand nous étudierons l'enrichissement des métadonnées. Pour clore cette sous-partie sur les raisons d'apposer des métadonnées, nous voudrions revenir sur la question de la mémoire que nous évoquions, et citer l'ouvrage de Richard Gartner *Metadata* : « Les métadonnées existent pour une raison et cette raison tient fondamentalement aux limites du cerveau humain. Aussi étonnant qu'il soit de bien des manières, le cerveau est limité dans la quantité d'information qu'il peut stocker et retrouver avec précision. Il y a des manières de contourner cela, bien sûr. Un truc que beaucoup apprennent est d'utiliser des moyens mnémotechniques simples pour mémoriser des concepts abstraits qui seraient sinon difficiles à retenir […] De tels moyens mnémotechniques marchent en tirant des éléments clés des données qu'ils sont faits pour conserver dans notre mémoire et en les retaillant sous une forme plus gérable cognitivement. Ils emploient un principe-clé des métadonnées : abstraire des structures des données pour les rendre plus facilement retrouvables. Dès que nous sortons des limites d'un cerveau unique, le besoin de métadonnées devient rapidement plus pressant. L'un

des bouleversements les plus significatifs des manières de conserver nos souvenirs hors du crâne humain a été l'invention de l'écriture. L'anthropologiste Jack Goody a écrit avec perspicacité dès 1963 sur l'impact de ce développement, en remarquant que cela a permis pour la première fois à la culture d'être transmise relativement intacte entre générations. Écrire, dit-il, évite que la culture transmise oralement, en passant par une chaîne de conversations entremêlées, ne soit automatiquement modifiée en raison des pressions et des impératifs de la vie sociale du moment[21]. […] Aussi profond qu'ait été l'impact de l'écriture, ces effets auraient été sévèrement circonscrits sans quelque manière d'organiser la trace écrite au-delà du niveau d'un seul et même texte. Il ne suffit pas d'écrire quelque chose, il faut encore pouvoir le retrouver. Une fois que notre écrit s'étend au-delà de quelques bribes, nous devons lui donner une forme logique si nous voulons un jour en faire quelque chose. Des systèmes de classement d'une forme quelconque sont rapidement devenus une nécessité et pour les faire fonctionner, il nous faut des métadonnées ; tout au moins il nous faut attacher des étiquettes à ce que nous avons écrit, plus petites que les textes eux-mêmes, afin qu'elles puissent être mises dans un ordre cohérent. Les métadonnées, qui sont déjà d'une grande aide pour retrouver les informations contenues dans un cerveau humain, deviennent vite essentielles quand on sort de ses limites. Mais les métadonnées ont un rôle qui est plus important que celui de simplement permettre qu'un fragment d'information soit emmagasiné puis retrouvé ensuite. Elles rendent possible que ceux-ci soient liés ensemble pour former une connaissance et à cette connaissance d'être consolidée en ce que nous entendons par culture »[22] (1, Gartner, pp. 8-9).

Nous souhaitions suivre cet aparté anthropologique de l'auteur pour mettre en évidence le lien essentiel entre métadonnées et culture. La capacité à retrouver et à transmettre un savoir, capacité qui se trouve à la base de la culture, est rendue possible par les métadonnées, comme le montre Gartner.

1.2.2. Typologie des métadonnées

Dans le texte d'Anne Gilliland, nous avons vu apparaître les termes de métadonnées administratives et métadonnées descriptives. Force est de constater que dans de nombreux documents proposant une vue d'ensemble sur les métadonnées, la question de savoir *comment* les métadonnées décrivent les ressources intervient avant celle de savoir *pourquoi* elles les décrivent. Nous avons pris le parti inverse, qui peut permettre de mieux aborder la proposition d'une typologie des métadonnées.

[21] Goody, J., & Watt, I. (1963). The consequences of literacy. Comparative Studies in *Society and History*, 5, 304–345.
[22] Nous traduisons.

Comme nous l'avons vue dans la première partie, nous pouvons considérer que les métadonnées sont des réponses à des questions qui pourraient être posées à propos de la ressource, questions en termes de qui ? quoi ? où ? quand ? comment ? combien ? pourquoi ? Ces questions sont évidemment susceptibles d'être spécifiées selon différentes dimensions. Le NISO propose une typologie largement partagée qui distingue métadonnées descriptives, administratives et structurelles. Dans le document déjà cité de Jenn Riley, celle-ci explique ces différentes catégories : « l'information à propos du contenu d'une ressource qui aide à la trouver ou à la comprendre est désignée comme métadonnée descriptive. […] Métadonnée administrative est un terme générique désignant l'information nécessaire à la gestion de la ressource ou qui se rapporte à sa création. Dans le champ des métadonnées administratives se trouvent les métadonnées techniques, l'information sur les fichiers numériques nécessaire pour les décoder et les convertir, comme le type de fichier ; les métadonnées de conservation subvenant à la gestion de long terme et à la migration ou à l'émulation de fichiers numériques, par exemple une somme de contrôle ou une fonction de hachage ; et des métadonnées juridiques, telles qu'une licence *Creative Commons*, qui détaille les droits de propriété intellectuelle attachés au document. Les métadonnées descriptives et administratives sont considérées comme distinctes des métadonnées structurelles, qui décrivent les relations les unes avec les autres de parties de la ressource ; on peut citer en exemple des pages dans une séquence, une table des matières avec des liens vers les débuts de sections, et la connexion de différentes résolutions ou finesses de représentation en couleur d'un même contenu » (8, Gilliland, p. 10). Pour résumer, nous distinguons trois types de métadonnées :

- Descriptives, permettant de trouver ou comprendre une ressource
- Administratives :
  - Techniques, permettant de savoir manipuler correctement le fichier
  - De conservation, permettant la gestion au long terme
  - Juridiques, renseignant sur les droits de propriété intellectuelle
- Structurelles, renseignant sur les relations entre les parties de la ressource

Jenn Riley ajoute ensuite à ces trois catégories une quatrième : « Une dernière catégorie de métadonnées sont *les langages de balisage*. Ces langages mélangent métadonnées et contenu, pratique utilisée parfois seulement avec d'autres formes de métadonnées. Des balises insérées dans le contenu indiquent des caractéristiques notables. Pour une ressource textuelle, cela peut être d'indiquer des éléments structurels comme des paragraphes ; baliser des mots avec de l'information sémantique – le fait que le mot est un nom de lieu ou une certaine partie du discours, par exemple ; ou fournir une information sur la typographie, comme des italiques » (ibid.). Il est effectivement notable que l'usage d'un langage de balisage puisse permettre de produire des documents dans lesquels contenu et métadonnées soient intimement liés. Toutefois, il nous semble que cette caractéristique ressortit plutôt de

la manière dont les métadonnées sont associées aux données, du lieu où se trouvent les métadonnées, comme nous l'avons étudié en 1.1.3. En l'occurrence, nous pourrions dire que les métadonnées que constitue un langage de balisage sont encapsulées au cœur même du document. Mais fondamentalement, nous pouvons les considérer comme des métadonnées soit descriptives (notamment dans le cas de l'annotation sémantique), soit structurelles, en tant qu'elles permettent de distinguer et de spécifier des parties d'un document.

Nous reprenons ci-dessous le tableau que propose Jenn Riley, qui permet de mettre en regard la typologie des métadonnées avec leurs usages, et qui donne quelques exemples de propriétés pouvant caractériser un livre numérique, par exemple.

Type de métadonnées	Exemples de propriétés	Usages primordiaux
Métadonnées descriptives	Titre Auteur Sujet Genre Date de Publication	Découverte Affichage Interopérabilité
Métadonnées techniques	Type de fichier Taille du fichier Date de création Schéma de compression	Interopérabilité Gestion d'objets digitaux Conservation
Métadonnées de conservation	Somme de contrôle (*checksum*) Événement relatif à la conservation	Interopérabilité Gestion d'objets digitaux Conservation
Métadonnées juridiques	Droits d'auteur Termes du contrat de licence Détention des droits	Interopérabilité Gestion d'objets digitaux
Métadonnées structurelles	Séquence Place dans la hiérarchie	Navigation
Langages de balisage	Paragraphe En-tête Liste Nom Date	Navigation Interopérabilité

Tableau 1 : Les types de métadonnées et leurs usages (J. Riley, 2017)

Nous remarquons qu'il n'y a pas, au premier abord, de correspondance directe entre les types de métadonnées et les usages des métadonnées dans le tableau de Jenn Riley. À partir des remarques que nous avons faites à la fois sur les usages et sur les types de données, nous pouvons proposer une mise en corrélation des types et des usages en spécifiant ces derniers.

Métadonnées descriptives	Accessibilité – Retrouver ou découvrir
Métadonnées administratives techniques	Interopérabilité
Métadonnées administratives de conservation	Conservation
Métadonnées administratives juridiques	Accessibilité – Droits d'accès
Métadonnées structurelles	Accessibilité – Intra-documentaire

Tableau 2 : Corrélation entre types et usages des métadonnées

1.2.3. Enrichir les métadonnées

Nous avons vu les multiples finalités des métadonnées et les types de métadonnées qui permettaient d'atteindre ces fins. Dans le cadre de ce mémoire, ce sont les métadonnées descriptives qui retiennent particulièrement notre attention, dans la mesure où ce sont celles qui permettent de favoriser l'accessibilité dans ses différentes dimensions. Pour améliorer cette accessibilité, il s'agirait alors d'*enrichir les métadonnées*. Nous devons donc à présent nous interroger sur cette expression pour étudier la réalité qu'elle recouvre. L'idée sous-jacente ne semble pas au premier abord sujette à questionnement : disposer de peu de métadonnées donne peu de prises sur la ressource ; avoir beaucoup de métadonnées multiplie les voies pour la retrouver, la découvrir, la faire apparaître selon différentes modalités. Il en va toutefois de cette métaphore de l'enrichissement comme de beaucoup de métaphores de la vie quotidienne[23], qu'il est bon d'interroger. Elle est porteuse d'une valorisation implicite : le riche est mieux que le pauvre, des données riches sont plus intéressantes que des données pauvres. Cela peut se comprendre, mais jusqu'à quel point ? Y a-t-il, ou pas, une limite souhaitable à l'enrichissement des métadonnées, au-delà de laquelle la quantité de métadonnées deviendrait un embarras ? Outre cette valorisation implicite, la métaphore de l'enrichissement porte également – nous semble-t-il - un modèle structurel implicite, qui est celui de l'accumulation : parler sans plus de précision d'enrichissement des métadonnées pourrait laisser penser qu'il s'agit d'accumuler, à propos d'une ressource, des métadonnées comme des écus dans un coffre - dont chacun aurait sa valeur indépendamment des autres et ajouterait sa valeur entière à celle de l'ensemble. Est-ce vraiment le cas ? Enrichir les métadonnées, est-ce seulement ajouter des métadonnées qu'on juxtapose aux précédentes ?

Nous devons d'abord signaler le rôle fondamental de l'« inventeur » de la ressource. Nous exploitons à dessein l'ambiguïté de ce mot pour désigner indifféremment celui qui crée la ressource ou qui découvre une ressource en tant que

[23] Voir à ce sujet : LAKOFF, George et JOHNSON, Mark. *Les métaphores dans la vie quotidienne*. Éditions de Minuit, 1986, 256 p.

ressource (c'est-à-dire qui décide qu'une entité peut constituer une ressource), dans la mesure où les ressources, nous l'avons signalé, peuvent être de natures très diverses. D'une certaine manière, c'est l'ajout des premières métadonnées, celles qui permettent d'identifier une entité, qui signe définitivement l'accession de celle-ci au statut de ressource. C'est parce qu'elle est identifiée qu'elle va pouvoir être caractérisée d'autres manières, ce que peut faire l'inventeur de la ressource en multipliant les descriptions ou les « étiquettes ». C'est la première manière d'enrichir les métadonnées. Si deux personnes s'intéressent à la même ressource mais produisent des métadonnées différentes, répondant à des intérêts initialement différents, ces deux personnes peuvent souhaiter enrichir leurs propres métadonnées de celles de l'autre. Dans la mesure où elles sont capables de s'assurer qu'elles parlent bien de la même ressource, elles peuvent croiser leurs métadonnées. C'est une deuxième manière d'enrichir les métadonnées : ajouter, aux métadonnées qu'on a soi-même produites sur une ressource, les métadonnées qu'un autre a produites sur la même ressource. Si la manière de décrire la ressource est suffisamment précise et partagée, plus précisément si l'on dispose d'un identifiant unique et pérenne et d'une manière commune (ou susceptible d'une traduction exacte) de décrire les objets, il devient possible de croiser les métadonnées avec celles de toute personne qui aura également produit des métadonnées utilisant un vocabulaire équivalent à propos d'une ressource ayant le même identifiant. Cette troisième manière d'enrichir les métadonnées est celle des « données liées », sur lesquelles nous reviendrons en détail dans la troisième partie de cette étude. Ces trois manières d'enrichir les métadonnées, notons-le, peuvent être utilisées par des professionnels de la ressource envisagée ou par des professionnels des métadonnées. Toutefois, comme nous l'avons déjà signalé à propos des folksonomies et déjà remarqué avec le texte d'Anne Gilliland, les métadonnées peuvent dépasser largement ce cadre et être utilisées hors de celui-ci. Lorsque c'est le cas, des utilisateurs non-professionnels sont eux-mêmes susceptibles de produire des métadonnées à propos d'une ressource. Nous voyons là une quatrième manière d'enrichir les métadonnées, que nous allons détailler : les « métadonnées sociales ».

Dans son article *Indexation collaborative : traces de lectures et constitution de communautés*, Evelyne Broudoux « se propose de s'intéresser plutôt aux usages qu'aux dispositifs du "boomarking social" en analysant deux types de pratiques », à savoir « les traces de lecture d'articles », ce qui « permet, d'une part, de dégager les tendances (domaines de recherche) des projets en cours et, d'autre part, de disposer d'un complément qualitatif aux mesures de citation effectuées par les éditeurs de revues scientifiques », et « la participation des usagers aux dispositifs collaboratifs initiés par les professionnels des bibliothèques, des centres d'archives, de documentation ou de musées » afin de savoir « quelle appropriation est faite de ces services 2.0 » ou d'en étudier l'« apport aux dispositifs classiques de repérage de l'information » (9, Broudoux, p. 126). Nous voyons donc se dégager deux directions assez différentes lorsqu'on parle de métadonnées issues de l'action des usagers,

selon que celui qui donne accès à la ressource produise des métadonnées à partir des transactions des utilisateurs dans l'interface d'accès, ou qu'il introduise dans les métadonnées de ses ressources des métadonnées produites par les usagers. C'est ce second aspect qui nous intéressera plus particulièrement. Si l'auteur considère qu'avec l'indexation collaborative, « un horizon s'est ouvert de connexion possible entre ces listes [de mots-clés générés par les usagers] et les vocabulaires contrôlés, thésaurus, ontologies ou autres formes d'organisation hiérarchisée des connaissances » (ibid.), le constat final reste en demi-teinte. Nous voudrions toutefois souligner deux initiatives, et d'abord, l'intéressante enquête du Online Computer Library Center (OCLC)[24] dont « l'objectif était d'établir un état de l'art des initiatives qui valorisent les collections en s'appuyant sur l'expertise de leur public pour enrichir les métadonnées, ce dernier terme étant d'ailleurs à entendre dans un sens très large puisqu'intégrant tagging mais aussi commentaires et recommandations, notations, etc. » (ibid., p.130). Comme le précise ensuite Evelyne Broudoux : « Deux objectifs principaux à la production de "métadonnées sociales" sont évalués en particulier : améliorer les métadonnées générées par les bibliothèques, archives, musées, de manière à accroître la qualité et la pertinence des résultats de recherche dans les catalogues et sur les moteurs de recherche ; contextualiser les contenus, de manière à faciliter leur compréhension et leur utilisation » (ibid.). Cette notion de contextualisation des contenus, que nous avons déjà vu apparaître dans le texte d'Anne Gilliland, est particulièrement importante si l'on s'attache à la réutilisation possible des métadonnées. En effet, les différentes contextualisations et recontextualisations possibles d'une ressource font sens au sein de communautés d'utilisateurs et sont difficilement envisageables de façon exhaustive par l'inventeur des métadonnées. Si l'on choisit d'ouvrir les métadonnées sur des usages larges, il est donc utile d'ouvrir également à la possibilité de l'ajout de métadonnée par les utilisateurs. Pour donner l'éventail possible des métadonnées produites par les usagers, Evelyne Broudoux rappelle que dans le cas de l'enquête du OCLC, « les métadonnées produites par les usagers ont été rangées en sept catégories : métadonnées pour la description, métadonnées pour l'accès, marquage (tagging), construction des collections et des contenus, notations et appréciations, partage et facilitation de la recherche, travail et construction de communautés, promotion des activités hors site » (ibid.).

Afin de comprendre quelle place peuvent avoir les métadonnées sociales dans la description d'une ressource, nous pouvons nous appuyer sur l'exemple, cité par Evelyne Broudoux dans son article, du projet *PhotosNormandie* de Patrick Peccatte et Michel Le Querrec. Patrick Peccatte le présente ainsi dans son allocution *Une plate-*

[24] Le OCLC se présente comme « une coopérative de bibliothèques mondiale, [qui] aide des milliers de bibliothèques à rendre l'information plus accessible et plus utile pour d'innombrables utilisateurs à travers la planète. Nous offrons des services informatiques partagés, des recherches originales et des programmes communautaires pour aider les bibliothèques à répondre aux besoins changeants de leurs utilisateurs, de leurs communautés et de leurs affiliés ». https://www.oclc.org/fr/home.html

forme sociale pour la redocumentarisation d'un fonds iconographique : « le projet *PhotosNormandie* a pour but d'améliorer l'indexation d'un fonds de photos historiques sur la Bataille de Normandie qui s'est déroulée du 6 juin à fin août 1944 durant la Seconde Guerre mondiale. Il s'agit d'un travail collaboratif à finalité patrimoniale » (10, Peccatte). En l'occurrence, il s'agit comme le précise l'auteur, de « photos du site *Archives Normandie 1939-1945* qui proviennent des Archives Nationales des États-Unis et du Canada » et qui « sont déclarées *libres de droits* » (ibid.). L'auteur rappelle ce que signifie cette mention et les conditions d'utilisation qui y sont attachées, qui rendent juridiquement viable le projet. Il explique ensuite la motivation de ce projet : « il est apparu […] que si la qualité des numérisations proposées sur le site Archives Normandie 1939-1945 est correcte, il n'en est pas de même des légendes qui comportent de nombreuses inexactitudes et incohérences. Certaines erreurs sont très importantes du point de vue historique ou même tout simplement descriptif et diminuent grandement l'intérêt documentaire de cette collection accessible au grand public » (ibid.). D'où le choix qu'opèrent l'auteur et Michel Le Querrec, d'utiliser un site de partage de photographies offrant des « possibilités de discussion sur une photo et [une] capacité à exploiter un sous ensemble significatif du standard de métadonnées IPTC/IIM » : *Flickr*, sur lequel « chaque photo peut être décrite par un titre, une description, des tags, des commentaires, des notes associées à des zones choisies de la photo. Le titre et la description sont modifiables uniquement par le gestionnaire du compte ayant téléchargé la photo tandis que les tags, les notes et les commentaires peuvent être ajoutés aux photos par tout visiteur si le gestionnaire du compte l'a autorisé. La recherche de photos s'effectue sur les titres, les descriptions et les tags, mais pas sur les commentaires ou les notes. Il est aussi possible de géotagger les photos, c'est-à-dire de leur attribuer des coordonnées géographiques, et d'effectuer des recherches cartographiques sur ces photos. L'utilisateur peut organiser ses photos par albums et classeurs (ensembles d'albums). Il est aussi possible de créer ou de participer à des groupes d'utilisateurs réunis sur des thématiques très variées et d'ajouter des photos à ces groupes. Flickr maintient aussi des flux RSS et Atom divers permettant de surveiller l'activité d'un utilisateur, d'un album, d'un groupe, etc., et propose une API permettant aux programmeurs de développer des services. On peut enfin choisir des contacts parmi les autres utilisateurs de Flickr et ajouter des photos quelconques à ses favoris.

Cet ensemble de possibilités que l'on vient de résumer rapidement permet de constituer un réseau de relations multiples entre les utilisateurs et leurs photos et justifie que Flickr soit souvent considéré comme l'un des sites exemplaires du Web social » (ibid.). Nous voyons donc ici la diversité des métadonnées qui peuvent être ajoutées, mais aussi comment s'organise le rapport entre le gestionnaire du compte et les visiteurs/contributeurs, et enfin les possibilités secondaires ouvertes en termes de création de réseaux entre les visiteurs. Patrick Peccatte détaille d'ailleurs « le processus documentaire et rédactionnel » : « Tout visiteur peut rechercher, afficher, télécharger les photos en haute définition. Pour commenter les photos et participer

ainsi à l'amélioration de leurs descriptions, l'utilisateur doit ouvrir un compte gratuit sur *Flickr* et propose alors ses corrections dans le champ *Ajoutez votre commentaire*. Une discussion peut s'établir entre les divers participants du projet et se termine par la validation éditoriale des modifications proposées.
Pour faciliter le suivi des discussions, nous avons créé un groupe *Discussions sur PhotosNormandie* qui donne une vue d'ensemble sur les photos où une discussion est en cours. Ce groupe est alimenté automatiquement par un programme développé à l'aide de l'API *Flickr* à partir du flux RSS permettant de surveiller les commentaires postés par les utilisateurs ; la consultation d'un groupe *Flickr* est en effet plus simple pour les utilisateurs qui ne maîtrisent pas la technique des flux RSS. Quand la discussion est terminée, l'administrateur rédige une nouvelle description dans les métadonnées IPTC de la photo, sur sa base locale de photos. Puis un autre programme développé à l'aide de l'API efface l'ancienne photo sur *Flickr* en se basant sur la référencé et envoie la même photo contenant la nouvelle description IPTC. La photo mise à jour avec sa nouvelle description apparaît alors automatiquement au début de l'ensemble *PhotosNormandie*.
Les photos sont classées dans différents albums : *Manche*, *Calvados*, *Orne*. Celles qui ne se rapportent pas directement à la Bataille de Normandie sont regroupées dans d'autres albums » (ibid.). L'auteur dresse enfin le « bilan documentaire » de son projet en établissant une « typologie des améliorations » : « identification des localisations, […] identification des personnages, […] identifications d'unités militaires, […] précisions de dates, […] précisions descriptives sur l'image, […] références : renvois à des livres, à des sites, […] identification des photos censurées, des photos en couleur, des photos en double et des séries, […] contextualisations historiques : précisions sur un mouvement d'unité, une action, etc. en rapport avec l'image, […] contextualisations iconographiques à l'aide d'autres sources » (ibid.). Nous pouvons faire deux remarques à propos de ce bilan : d'abord qu'il montre l'apport — qu'il serait difficile d'obtenir autrement — de compétences techniques, de connaissance des personnes figurant sur les photographies et des contextes de celles-ci, et d'autre part qu'il existe un lien fort entre enrichissement et qualité des métadonnées. Comme l'auteur le souligne : « l'activité régulière du projet durant plus d'une année a permis d'améliorer de façon substantielle la qualité des descriptions, et même, pensons-nous, de les enrichir d'informations probablement inédites » (ibid.). Nous irions plus loin en affirmant qu'améliorer la qualité des métadonnées est en soi une modalité de leur enrichissement, d'abord parce que l'enrichissement peut apparaître comme vain si les métadonnées ajoutées sont inexactes, mais aussi parce que des métadonnées exactes permettent de futurs enrichissements. Revenons enfin sur le titre-même de l'allocution de Patrick Peccatte, plus précisément sur le terme de « redocumentarisation ». Il est ici employé au sens où le définit Manuel Zacklad : « Redocumentariser, c'est documentariser à nouveau un document ou une collection en permettant à un bénéficiaire de réarticuler les contenus sémiotiques selon son interprétation et ses usages à la fois selon la dimension interne (extraction de

morceaux musicaux pour les réagencer avec d'autres, ou annotations en marge d'un livre suggérant des parcours de lecture différents…) ou externe (organisation d'une collection, d'une archive, d'un catalogue privé croisant les ressources de différents éditeurs selon une nouvelle logique d'association). Dans ce contexte, la numérisation offre des opportunités inédites pour la réappropriation des documents et des dossiers en vue de satisfaire les intérêts de nouveaux bénéficiaires » (11, Zacklad, p. 282). Notons la dimension collective incluse dans la notion de redocumentarisation, qui nous permet de désigner précisément ce que nous visions dans cette partie, à savoir l'enrichissement des métadonnées relatives à une ressource par la contribution de métadonnées produites par les utilisateurs.

Nous pouvons à présent tenter de répondre aux questions que nous nous posions concernant l'enrichissement des métadonnées. S'agit-il d'une pure et simple accumulation ? Cette dimension n'est pas à négliger : la réunion, la collecte de métadonnées diverses, qui envisagent la ressource sous différents aspects, constitue indubitablement un enrichissement des métadonnées. Toutefois, les pratiques de croisement des métadonnées, de classification des ressources à partir des métadonnées, indiquent que l'enrichissement des métadonnées ne va pas sans une certaine structuration de celles-ci. En outre, comme nous l'avons remarqué, l'enrichissement passe aussi par une meilleure qualité des données : l'enrichissement n'est pas selon nous uniquement quantitatif, mais aussi qualitatif. Y a-t-il une limite souhaitable à l'enrichissement des métadonnées ? Dans le cadre conceptuel de la redocumentarisation, de l'usage de métadonnées sociales, de la prise en compte de la diversité de réutilisations possibles, il ne semble pas qu'on puisse assigner une telle limite. Nous voyons toutefois les problèmes que peut poser un ajout sans limite de métadonnées pour décrire une ressource. C'est tout l'intérêt de l'ouvrage de Getaneh Alemu et Brett Stevens, qui se présente comme « Une théorie naissante des métadonnées des bibliothèques numériques – Enrichir puis filtrer » (12, Alemu et Stevens, p. 3)[25]. Nous reviendrons plus en détail dans la troisième partie de cette étude sur la théorie proposée, dont l'objectif est justement d'intégrer une approche des métadonnées basée sur les utilisateurs et d'en finir avec un modèle uniquement basé sur les métadonnées standardisées définies par des experts. Nous voulons seulement pour le moment noter d'une part qu'il existe une solution aux problèmes d'un enrichissement non limité des métadonnées, et d'autre part que le plus souvent, actuellement, certains considèrent qu'il y a bien une limite assignable à l'enrichissement des métadonnées. Alemu et Stevens indiquent que « selon l'*International Federation of Library* Associations (IFLA)[26] et Svenonius[27], on trouve

[25] Nous traduisons.

[26] IFLA. *Functional requirements for bibliographic records: Final report*. International Federation of Library Associations and Institutions. 2009.

[27] SVENONIUS, Elaine. *The intellectual foundation of information organization*. Cambridge, MA/London : MIT Press, 2000, 264 p.

parmi les plus importants principes fondateurs guidant les approches des métadonnées basées sur des standards : le principe de nécessité et suffisance ; le principe de commodité pour l'utilisateur ; le principe de représentation ; le principe de standardisation » (ibid., p. 11). Selon le principe de nécessité et suffisance, nous disent Alemu et Stevens qui citent l'IFLA, « seuls les éléments de données dans les descriptions et les formes de noms contrôlées pour l'accès, qui sont requis pour suffire aux tâches des utilisateurs et sont nécessaires pour identifier de façon univoque une entité, doivent être inclus » (ibid., p.12). Les enjeux relèvent à la fois de l'économie de l'attention, les métadonnées superflues pouvant induire la confusion parmi les utilisateurs, et aussi de l'économie tout court, comme le notent les auteurs : « dans la mesure où la création de métadonnées est un effort coûteux, le principe de suffisance et nécessité est considéré comme permettant l'efficacité, et par conséquent comme réduisant les coûts. […] Ainsi que le remarque Svenonius, les bibliothécaires appliquent la règle du rasoir d'Occam, selon laquelle les métadonnées tenues pour superflues sont éliminées. Toutefois, une telle conception de la simplicité peut affecter significativement les besoins des utilisateurs » (ibid.). N'est-ce pas pris en compte par le « principe de commodité pour les utilisateurs » ? Selon les auteurs, les tentatives de placer l'utilisateur au centre des services offerts par la bibliothèque se heurte à différents problèmes : à « des limitations inhérentes aux bibliothèques physiques et aux coûts associés à la création et à l'entretien de métadonnées » (ibid., p. 13), mais plus fondamentalement à l'utilisation de standards, de hiérarchies, de catégories, de données issues de sources externes difficiles à modifier, et plus radicalement encore du postulat que le bibliothécaire peut déterminer quels sont les besoins des utilisateurs, considérés comme « des consommateurs passifs de métadonnées ». Le principe de représentation, qui « stipule que les métadonnées devraient représenter objectivement et précisément un objet informationnel et que de telles représentations évitent les descriptions idiosyncratiques » (c'est-à-dire propres à un individu particulier), peut faire l'objet d'une critique similaire, puisqu'il « présuppose que les experts des métadonnées sont les seuls et uniques créateurs de métadonnées » et, en outre, « interdit l'inclusion de deux descriptions contradictoires d'un objet informationnel, excluant ainsi la coexistence d'interprétations diverses, potentiellement conflictuelles » (ibid., p.14). Quant au principe de standardisation préconisé par l'IFLA et la NISO, qui s'applique aux descriptions et aux points d'accès au catalogue (il faut ici entendre un accès par créateur, par titre, par sujet ou d'autres critères), les auteurs reconnaissent son utilité pour assurer la cohérence et l'interopérabilité des métadonnées, mais constatent qu'en pratique, les différentes institutions recourent souvent à des mélanges de standards locaux, nationaux et internationaux, ce qui selon eux déplace le problème vers la question de savoir « comment rendre interopérables et relier entre eux différents standards » (ibid., p.15).

1.3. L'interopérabilité des métadonnées

Il est une dimension des métadonnées dont nous n'avons pas encore explicitement traité, bien que nous nous ayons été amenés, à de nombreuses reprises lors de nos développements précédents, à l'aborder en partie : celle qui tient à la façon dont les métadonnées sont stockées et diffusées. Nous avons croisé les notions de format, de standard, de classification, de plateforme, sans définir de quoi il s'agissait. Nous avons également vu l'importance pour l'enrichissement des métadonnées de pouvoir utiliser des métadonnées produites dans d'autres contextes et d'autres systèmes d'information, et donc employant possiblement des manières différentes de stocker et de diffuser leurs métadonnées. C'est pourquoi il nous semble que la manière la plus pertinente de faire le point sur ces questions est de les aborder sous l'angle de l'interopérabilité, enjeu majeur pour les métadonnées.

Le *Référentiel Général d'Interopérabilité* (RGI) produit par la Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État (DINSIC) fait référence à la définition de l'interopérabilité de l'Association Francophone des Utilisateurs de Logiciels Libres (AFUL) et de Wikipedia : « L'interopérabilité est la capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre » (13, DINSIC, p. 7). A travers les exigences de connaissance intégrale des interfaces et d'absence de restriction d'accès ou de mise en œuvre, nous pouvons constater que cette notion correspond à un objectif élevé de possibilité de coopération entre systèmes, qui dépasse la simple compatibilité mais reconnaît la diversité possible des systèmes d'information (sans quoi l'objectif deviendrait l'intégration). Le RGI retient en revanche la définition de la Commission Européenne du cadre d'interopérabilité, qui concerne donc les États mais présente des éléments intéressants : « Un cadre d'interopérabilité est une approche concertée de l'interopérabilité pour les organisations qui souhaitent travailler ensemble à la délivrance conjointe de services publics. Au sein de son champ d'application, il spécifie un ensemble d'éléments communs tels que le vocabulaire, les concepts, les principes, les politiques, directives, recommandations, normes, spécifications et pratiques » (ibid., pp. 7-8). Nous suivons le RGI lorsqu'il souligne que « plusieurs éléments importants sont à retenir dans ces définitions : l'approche concertée entre les parties ; le fait que les interfaces des systèmes par lesquelles les échanges sont réalisés soient intégralement connues et donc décrites d'un point de vue technique, sémantique, fonctionnel et opérationnel ; la capacité à fonctionner avec d'autres systèmes sans restriction ; le fait que l'interopérabilité ne soit pas qu'une question technique, mais touche également aux questions de vocabulaire, de concepts métiers, de principes d'architecture et d'organisation, de réglementation, de droit, de politiques » (ibid., p. 8). En effet, l'interopérabilité peut être considérée sous plusieurs aspects : l'interopérabilité technique, l'interopérabilité syntaxique, l'interopérabilité sémantique, voire pour

certains – c'est le cas pour le RGI – l'interopérabilité organisationnelle. C'est à l'étude de ces différents aspects que nous allons nous attacher à présent.

1.3.1. L'interopérabilité technique

L'interopérabilité technique concerne la manière dont vont être échangées les métadonnées. D'une certaine manière, nous pourrions considérer que nous ne sommes pas encore ici au niveau des métadonnées, et qu'un protocole d'échange de données entre systèmes est de nature à permettre la circulation de toutes sortes de données, et pas seulement des métadonnées. Remarquons toutefois que ce premier aspect est absolument nécessaire aux suivants, et saisissons d'autre part cette occasion pour présenter certains protocoles apparus dans le monde du livre et qui nous paraissent dignes d'intérêt.

La Bibliothèque Nationale de France (BnF) consacre une page de son site aux « Protocoles d'échange de données »[28], dans laquelle il est expliqué que « tout échange d'information repose nécessairement sur un ensemble de conventions partagées entre l'émetteur et le destinataire d'un message : il faut que l'un et l'autre sachent notamment à quel moment elle se termine. De tels ensembles de conventions sont appelés des protocoles. Il existe plusieurs protocoles qui permettent à des machines d'échanger des données bibliographiques ou des métadonnées ». Trois de ces protocoles sont ensuite présentés : Z39.50, OAI-PMH et SRU. L'intérêt de Z39.50 est son caractère « historique », puisqu'il est apparu dès les années 1980. C'est selon la BnF « le plus ancien protocole d'échange de données, encore largement utilisé dans les bibliothèques », qui le décrit ainsi : « La norme Z39.50 définit un client/serveur basé sur un service et un protocole pour la recherche et le transfert d'informations. Ces informations peuvent être en formats MARC[29] structurés pour l'échange selon la *norme ISO 2709 : Information et documentation- Format pour l'échange de l'information*. Le protocole Z39.50 spécifie les procédures et les formats pour permettre à un client de requêter une base de données proposée par un serveur, d'identifier les informations correspondant aux critères de la recherche, et de récupérer les informations identifiées. Z39.50 fonctionne en mode synchrone, à la différence d'un moissonneur. Les requêtes émanant du client sont exécutées en direct sur le serveur distant et les résultats sont rapatriés instantanément en retour ». NISO, qui en a fait un standard en 1988, rappelle que « Z39.50 est né du Linked Systems Project (LSP), une initiative datant des années 1980 pour standardiser la recherche dans les principales bases de données bibliographiques de l'OCLC, la Librairie du Congrès, le Washington (Western) Library Network (WLN), et le Research Libraries Information

[28] BnF. *Protocoles d'échanges de données*. [En ligne]. Disponible sur : <http://www.bnf.fr/fr/professionnels/protocoles_echange_donnees.html> (consulté le 20 août 2018)
[29] Le MARC, pour *MAchine Readable Cataloging*, est un format de diffusion et d'échange de notices bibliographiques utilisé par les bibliothèques et apparu à la fin des années 1960.

Network of the Reasearch Libraries Group (RLG). En parallèle de l'initiative du LSP, des efforts de standardisation étaient accomplis autour d'un protocole de récupération de l'information pour des applications en bibliothèque sous les auspices de la NISO. Le protocole développé par le projet LSP fut transféré à la NISO et développé plus avant en standard de récupération d'information Z39.50, puis approuvé comme un standard NISO en 1988 »[30]. Nous laisserons de côté le SRU, que la BnF définit ainsi : « le protocole d'échange de métadonnées SRU (Search/Retrieval via URL) est l'équivalent fonctionnel du protocole Z39.50 en s'adaptant aux standards technologiques du web (protocole http et format XML) », pour nous intéresser plus particulièrement à l'OAI-PMH, en nous tournant à nouveau vers la définition de la BnF : « OAI-PMH est le sigle de l'Open Archives Initiative – Protocol for Metadata Harvesting, ce qui signifie "protocole pour la collecte de métadonnées de l'Initiative pour les Archives Ouvertes" ». Sa description est la suivante : « Le protocole OAI-PMH est un moyen d'échanger sur Internet des métadonnées entre plusieurs institutions, afin de multiplier les accès aux documents numériques. Il permet :

- d'accroître la visibilité des collections numériques sur Internet,
- de reconstituer virtuellement des corpus à partir de ressources accessibles sur différents sites,
- d'alimenter des portails thématiques.

Son utilisation est libre, tout comme ses spécifications, disponibles sur le site www.openarchives.org.
L'OAI-PMH définit deux types d'acteurs :

- les fournisseurs de données, qui déposent leurs métadonnées sur un serveur web appelé "entrepôt",
- et les fournisseurs de service qui collectent (on dit aussi "moissonnent") ces données pour les intégrer à l'index de leurs propres bibliothèques numériques.

Un même établissement peut jouer les deux rôles : diffuser ses métadonnées et collecter celles des autres.
Le fonctionnement de base du protocole OAI-PMH repose sur une communication de client à serveur. Le client envoie des requêtes au serveur en http, le serveur répond par un flux de données en XML »[31].

Nous pouvons compléter cette rapide description avec ce que dit en 2009 Gérard Puimatto dans un paragraphe — intitulé « Le moissonnage, complément naturel des métadonnées » — de son article, afin de mieux comprendre quelle est la pratique effective de l'OAI-PMH : « les formats utilisés comportent toujours le noyau commun du *Dublin Core non qualifié*, comme base commune d'interopérabilité. En sus, chaque acteur, ou catégorie d'acteurs, peut associer ses propres métadonnées

[30] NISO. *Z39.50 – A Primer on the Protocol*. [En ligne] 2002. Disponible sur : <https://www.niso.org/sites/default/files/2017-08/Z3950_primer.pdf> (consulté le 20 août 2018)
[31] BnF. *Protocole OAI-PMH*. [En ligne]. Disponible sur : < http://www.bnf.fr/fr/professionnels/protocoles_echange_donnees/a.proto_oai.html > (consulté le 20 août 2018)

selon ses schémas propres : schémas de description des thèses, notices MARCXML, métadonnées LOM ou SCORM, schémas de l'inventaire en matière de patrimoine, etc. Ce modèle institue ainsi un réseau d'entrepôts OAI que l'on peut interroger de façon très simple au moyen de quelques requêtes. Pour autant, ces interrogations sont rarement le fait d'usagers directs : ce sont le plus souvent des moissonneurs OAI qui dialoguent avec les entrepôts en vue de collecter de façon sélective les métadonnées répondant à des critères de recherche définis. On constitue ainsi des "silos" de notices dans lesquels on peut organiser des recherches, en s'appuyant sur l'homogénéité des descriptions initiales.

Via OAI-PMH, les métadonnées deviennent un système d'information documentaire structuré et déployé au plan mondial, avec ses modes propres de fonctionnement. Base de solutions documentaires spécialisées, OAI-PMH est aussi utilisé par les grands moteurs de recherche pour fournir des services de recherche plus efficaces, la structuration permettant de réduire le bruit documentaire » (6, Puimatto, p. 13).

Si le moissonnage apparaît, au moment où écrit l'auteur, comme un « complément naturel des métadonnées », cette vision n'est plus autant partagée quelques années plus tard. Jenn Riley nous donne ainsi une profondeur de regard historique sur cette pratique : « Les institutions du patrimoine culturel ont une longue histoire de partage de métadonnées, remontant à la distribution par la Bibliothèque du Congrès des Etats-Unis de fiches de catalogage (en premier lieu pour des livres) aux bibliothèques locales. Au début des années 2000, la communauté du patrimoine culturel est entrée dans une nouvelle phase de coopération en commençant à partager des métadonnées basées sur le XML à travers le *Open Archives Initiative Protocol for Metadata Harvesting*[32] (OAI-PMH). Avec OAI-PMH, les types de document faisant habituellement l'objet de partage de métadonnées entre bibliothèques, archives, musées se sont largement diversifiés pour inclure des documents tels que des collections de photographies et des préimpressions d'articles de recherche produits dans les universités. Pendant un temps, Google a soutenu et nourri l'utilisation d'OAI-PMH comme une partie de son protocole Sitemaps, mais ce soutien s'est arrêté en 2008. Bien qu'OAI-PMH soit encore utilisé par les communautés de dépôt institutionnel[33] [*institutional repository*] et de collections numériques en raison de sa présence dans des packs logiciels répandus, comme DSpace, ses limites sont bien connues. ResourceSync, un protocole lui succédant qui opère selon les spécifications XML de Sitemaps, a gagné du terrain, bien que plus récemment, les données liées semblent plus prometteuses pour le futur de l'échange des métadonnées » (7, Riley, p. 9).

Nous verrons dans la troisième partie de notre travail quelles sont les perspectives ouvertes par le paradigme des données liées en termes d'interopérabilité.

[32] Initiative d'Archives Ouvertes – Protocole de Moissonnage de Métadonnées

[33] Ce terme désigne, dans le cas des universités, un ensemble de services permettant la gestion et la diffusion de documents numériques créés par l'institution et ses membres.

1.3.2. L'interopérabilité syntaxique

L'interopérabilité syntaxique des métadonnées concerne les formats qui permettent de les stocker et éventuellement de les transmettre, indépendamment de leur contenu (les questions d'interopérabilité relatives au contenu ressortissent de l'interopérabilité sémantique, que nous traitons dans le point suivant). Comme le note Jenn Riley, « les métadonnées peuvent se trouver sous différentes formes et différents encodages. Dans le modèle traditionnel des systèmes d'information, elles peuvent être stockées en tant que champs dans les tables d'une base de données relationnelle. Un ensemble de métadonnées, dans ce contexte, est appelé un *enregistrement*. Dans ce modèle, une conception efficace est basée sur une normalisation adéquate des tables de la base de données conciliant la maximisation de l'efficacité du stockage et l'optimisation de la performance des requêtes. Dans ce scénario, typiquement, les métadonnées seraient chargées par lot selon des processus spécifiques ou entrées à la main à travers des interfaces utilisateurs à cet effet, toutes contrôlées par une programmation spécifique ». Nous retrouvons avec cette description le monde que nous présentait Francis Miksa, celui où les métadonnées sont les noms des champs d'une base de données. Jenn Riley continue cependant en remarquant qu'« aujourd'hui, les systèmes logiciels qui utilisent ce modèle de métadonnées et souhaitent partager leurs métadonnées avec d'autres le font habituellement au travers d'*Application Programming Interfaces* (APIs), en publiant des documents techniques que les développeurs de logiciels externes peuvent utiliser pour construire des outils qui requêtent le système et récupèrent des métadonnées d'intérêt » (7, Riley, p. 8). Ce modèle traditionnel, même s'il permet d'une certaine manière de rendre accessibles ses métadonnées au travers d'APIs, n'est pas celui de l'interopérabilité syntaxique. Pour cela, il faut un format qui soit connu.

Nos exemples dans cette partie seront évidemment orientés, puisque nous traitons du secteur du livre, et que les formats sont tellement nombreux que nous ne pouvons tous les envisager. Si nous nous en tenons donc au domaine du livre, le format qui s'est imposé est le format XML. C'est ce que décrit Jenn Riley dans la suite de son ouvrage : « Dans les années 2000, le XML (eXtensible Markup Language) a surgi comme un dispositif d'encodage, de transfert, voire de système de stockage interne, communément utilisé pour les métadonnées. Les métadonnées en XML existent en tant qu'ensembles de fichiers, appelés document XML. XML définit des éléments, des balises qui signalent que les valeurs qu'elles enclosent ont une certaine signification. Les éléments peuvent également enclore d'autres éléments, et c'est de cette caractéristique que les documents XML tirent leur structure. Un document XML est une arborescence qui commence par un unique élément racine. Les autres éléments et valeurs sortent comme des branches de cette racine originelle, construisant une structure emboîtée qui contribuent au sens des valeurs des métadonnées dans le document. Les éléments XML peuvent recevoir des attributs, qui ont généralement leurs propres valeurs. Un attribut XML et sa valeur raffinent le

sens de l'élément dans lequel ils apparaissent. XML permet le multilinguisme des métadonnées en fournissant un attribut prédéfini pour indiquer la langue dans laquelle est exprimée la valeur d'un élément. Comme pour les bases de données relationnelles, un document XML décrivant un certain objet est appelé un enregistrement de métadonnées » (ibid., p. 12). Si nous nous plaçons maintenant du point de vue de celui qui reçoit des métadonnées exprimées au format XML, quelles sont les possibilités offertes pour manipuler et intégrer les données ? Comme le précise Jenn Riley, « des langages spécialisés (notamment XPath, XSLT, XQuery) existent pour transformer et requêter des documents XML, ainsi que des boîtes à outils de traitement pour les principaux langages de programmation. Une conception efficace en XML se focalise sur des choix équilibrés dans l'utilisation d'éléments et d'attributs, et une attention à la taille du document permettant une bonne performance dans les requêtes. Les métadonnées stockées en XML dans un système sont souvent chargées depuis des sources externes, ou peuvent dans d'autres cas être générées à travers des interfaces logicielles pour les utilisateurs, ou mappées en masse d'autres sources de données. Dans de nombreux cas, les données XML sont ingérées dans un système qui les restitue sous d'autres formes pour le stockage et l'indexation, bien qu'il existe des bases de données nativement en XML. L'utilisation du XML n'est pas limitée aux métadonnées descriptives ; plusieurs types de métadonnées peuvent être enregistrées dans des documents en XML » (ibid., p. 13).

A présent que nous voyons plus précisément ce que signifie l'interopérabilité syntaxique, il nous semble nécessaire de faire un point sur différents termes que nous avons rencontrés mais dont le sens mérite nous semble-t-il d'être précisé. « Format », « norme », « standard », « schéma » de métadonnées : à quoi ces termes renvoient-ils ? Bien évidemment, les questions de normalisation et de standardisation sont d'un usage générique et dépassent largement la question des formats de métadonnées. Quand on parle de norme ou de standard de métadonnées, il faut donc comprendre format normalisé ou format standardisé de métadonnées. Mais quelle est la différence entre une norme et un standard ? En anglais, les deux sont désignés par le mot de *standard* et le RGI précise que « même si la langue française distingue les deux termes "standard" et "norme", le terme "standard" est utilisé par défaut dans l'ensemble du document ». Toutefois, il est de coutume de les distinguer par le type d'organisme qui promeut l'utilisation d'un certain format. Ainsi, l'Organisation Internationale de Normalisation (ISO) définit la norme comme un « document établi par consensus et approuvé par un organisme reconnu, qui fournit, pour des usages communs et répétés, des règles, des lignes directrices ou des caractéristiques, pour des activités ou leurs résultats garantissant un niveau d'ordre optimal dans un contexte donné »[34]. Cette définition, tout utile qu'elle soit, ne va pas sans soulever des questions, qui rejoignent d'ailleurs celles que se posent Alemu et Stevens à propos des principes de l'IFLA (cf 1.2.3), et nous pouvons reprendre à notre compte les

[34] Guide ISO/CEI 2

questions et réflexions de Bernard Blandin : « Qui établit le consensus, et en vertu de quelle légitimité ? », « A qui s'impose le "niveau d'ordre optimal" garanti ? A ceux qui ont établi le consensus, ou à d'autres ? », « Si le "niveau d'ordre optimal" garanti s'impose à d'autres que ceux qui l'ont approuvé, est-il acceptable par ceux à qui il s'impose ? », « Tant que la norme concerne des objets techniques ou des procédés industriels, le processus de normalisation peut se satisfaire d'un consensus entre experts représentant les concepteurs de l'objet ou du procédé, garantissant un résultat optimal obtenu avec l'objet ou par les procédés aussi bien aux concepteurs qu'à ceux qui vont les utiliser ou les mettre en œuvre. Avec la normalisation des technologies de l'information, on ne normalise plus un objet ou un procédé, mais la représentation numérique de cet objet ou de ce procédé et les traitements que l'on peut opérer sur ces représentations » (14, Blandin). Les standards peuvent échapper en partie à ces critiques, lorsqu'ils émanent d'une communauté d'utilisateurs. Mais c'est pour tomber de Charybde en Scylla, comme le montre la carte des standards de métadonnées que propose Jenn Riley sur son site[35] et que nous reproduisons ci-dessous :

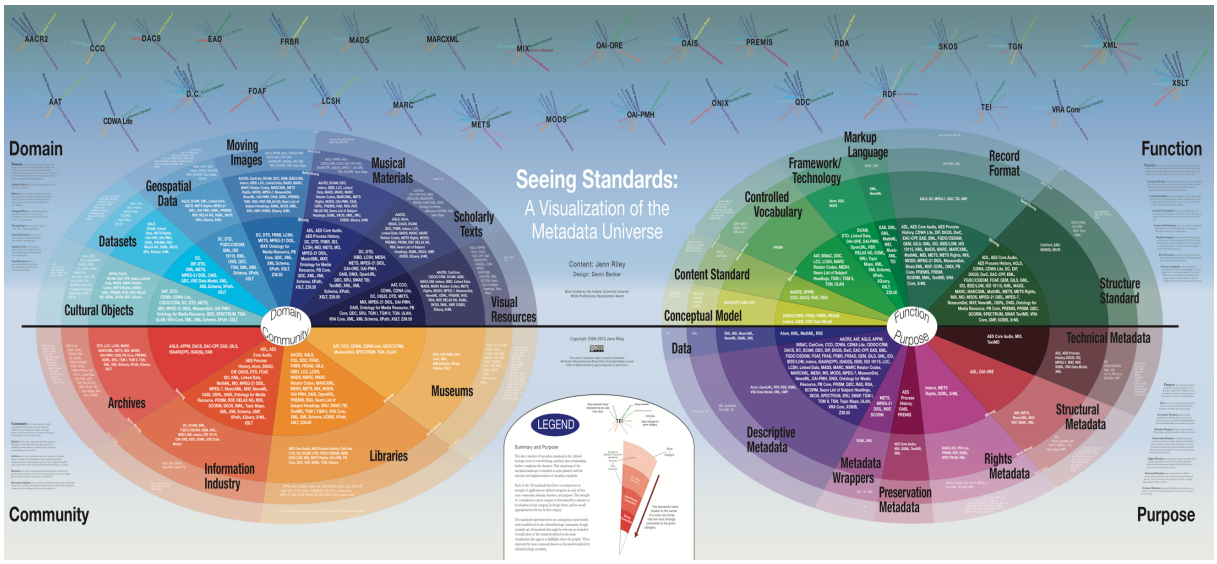

Figure 4 : Carte des standards de métadonnées (J. Riley, 2009-2010)

Comme le remarque *cum grano salis* Richard Gartner en commentant cette carte : « pour ceux qui ont prédit que l'ère informatique verrait la disparition des métadonnées, un rappel salutaire du caractère malvenu de cette idée se trouve dans le diagramme […] de la célèbre bibliothécaire numérique Jenn Riley. Intitulé *Seeing Standards*, il offre une carte du monde des standards de métadonnées au moment de sa conception, en 2010. Dans les ovales sont contenus les acronymes de plus d'une centaine d'entre eux : pour chacun, les sommités de leurs communautés respectives ont délibéré en détail pour distiller ce qu'ils regardent comme l'essence de leur savoir

[35] RILEY, Jenn. *Seeing Standards : A Visualization of the Metadata Universe*. [En ligne]. Disponible sur : <http://jennriley.com/metadatamap/seeingstandards.pdf> (consulté le 20 août 2018)

collectif dans un ensemble finement ciselé de règles et d'instructions pour les métadonnées.

Le diagramme de Riley montre, comme nous pouvions nous y attendre, qu'il y a différents types de standards pour différents types de données (appelés *Domain* dans sa représentation). Il y a aussi différents standards pour les différentes fonctions et finalités que les métadonnées peuvent avoir, ce qui n'est pas non plus une surprise. Ce qui est intéressant, c'est à quel point les standards sont séparés par communauté (ce qui est montré en bas à gauche du diagramme) : les musées, les bibliothèques, l'industrie de l'information et les archives ont tous leurs standards propres, et peu de recoupements entre eux sont permis ou même désirés. Loin d'une technologie qui rassemblerait les métadonnées pour accomplir les visions d'une conversion d'Internet en une seule et vaste bibliothèque d'information et de connaissances, il apparaît que les métadonnées se fragmentent en factions, aussi résolument qu'aux autres moments de leur histoire.

Les standards sont des choses complexes : la spécification d'un seul est susceptible de prendre des centaines de pages et des mois à apprendre. Il n'est pas surprenant qu'une fois l'effort fait de se familiariser avec le standard prédominant dans une certaine communauté, peu souhaitent franchir la frontière vers le territoire des autres, qu'ils ne connaissent pas. Il n'y a rien d'étrange dans le fait que les pratiques relatives aux métadonnées soient devenues à présent aussi fortement retranchées qu'elles l'ont toujours été à travers l'Histoire.

Ce n'est pas pour rien que le célèbre informaticien Andrew Tanenbaum blaguait en 2003 en disant "ce qui est bien avec les standards, c'est qu'on en ait tant parmi lesquels choisir"[36]. Il n'y a certainement aucun sentiment d'un point final dans cette enquête historique sur les métadonnées, aucun sentiment que nous aurions avancé vers une approche unique, universelle, d'organisation des données, de l'information, de la connaissance, de la compréhension, de la sagesse. Nous nageons dans un torrent de métadonnées aussi résolument qu'à tout autre moment de l'Histoire : les médias qui les contiennent ont changé du tout au tout, pas la nature humaine » (1, Gartner, pp. 37-38).

Pour conclure notre point lexical, il nous faut encore évoquer les schémas de métadonnées. Si les normes comme les standards visent à fixer quelles métadonnées doivent être spécifiées dans un certain domaine ou pour une certaine finalité, il peut arriver que la simple spécification d'un périmètre de métadonnées ne suffise pas, mais qu'il faille proposer une structuration plus poussée des métadonnées entre elles. Comme le définit DoRANum dans une infographie sur « Les schémas de métadonnées »[37], « un schéma de métadonnées est […] une liste structurée composée d'éléments descriptifs reliés entre eux. Pour chaque élément, le schéma définit sa signification (par exemple ici se trouve le titre, ici l'auteur, et là la date de

[36] TANENBAUM, Andrew. *Computer networks*. Upper Saddle River: Prentice Hall PTR. 2003.

[37] DoRANum. *Les schémas de métadonnées*. [En ligne]. Disponible sur : <https://doranum.fr/wp-content/uploads/script_schema_metadonnees.pdf> (Consulté le 20 août 2018)

publication), le type de contenu attendu (comme du texte ou des nombres), sa formulation (ça peut être du texte libre, un format précis, ou encore une norme à respecter) et enfin les valeurs qu'il est possible d'attribuer (comme un terme issu d'un thésaurus, ou encore un choix à faire dans une liste fermée) […] Le schéma définit aussi ce qu'il est possible ou non de faire avec les éléments. On peut distinguer : le niveau d'obligation (quels éléments sont obligatoires, conseillés, ou simplement facultatifs ?), la possibilité de rajouter ou non des éléments, et enfin des règles plus spécifiques (par exemple si tel champ est renseigné alors celui d'après doit l'être aussi) ».

Nous voyons donc que l'interopérabilité syntaxique comporte en réalité deux niveaux qu'il convient de distinguer. Le schéma de métadonnées relève plus d'une modélisation de la structure des métadonnées. Il n'est pas en droit dépendant d'un format particulier. Il peut d'autre part être tout à fait simple et correspondre à une liste de champs d'une table dans une base de données, ce qui pourrait estomper la distinction que nous cherchons à mettre en évidence. Cette distinction, dans le cas d'un document en XML, est celle entre un document *bien formé* et un document *valide*. Le World Wild Web Consortium (W3C) définit ainsi cette distinction : « Un document XML "bien formé" n'est pas la même chose qu'un document XML "valide". Un document XML "valide" doit être bien formé. En outre, il doit se conformer à une définition de type de document [Document Type Definition - DTD]. Deux définitions de type de document peuvent être utilisées avec XML : DTD – la définition de type de document originelle ; XML Schema – une alternative à la DTD basée sur XML. Une définition de type de document définit les règles et les éléments et attributs licites pour un document XML »[38]. Nous pourrions dire que la correction syntaxique d'un document XML, qui va permettre son interopérabilité, joue donc à deux niveaux, l'un intrinsèque au format (la « bienformation » du document) et l'autre extrinsèque au format dans lequel sont transmises les métadonnées (la validité du document), qui tient au schéma de celles-ci. Pour que l'interopérabilité soit effective, il faut un accord entre ceux qui s'échangent des fichiers, cet accord tenant à l'établissement d'un standard ou d'une norme.

1.3.3. L'interopérabilité sémantique

L'interopérabilité sémantique, comme nous le remarquions, se situe au niveau des valeurs prises par les métadonnées. L'enjeu est donc, selon une définition couramment reprise de Sandra Heiler, « de s'assurer que les échanges qui s'effectuent conservent leur sens, c'est-à-dire que les parties communicantes ont une compréhension commune de la signification des données qu'elles s'échangent »[39]. Il

[38] W3C. XML validator. [En ligne]. Disponible sur : <https://www.w3schools.com/xml/xml_validator.asp> (consulté le 20 août 2018). Nous traduisons
[39] HEILER, Sandra. *Semantic interoperability*. ACM Computing Survey, vol. 27 (2), 1995, pp. 271-273.

y a de nombreuses raisons pour lesquelles cette compréhension commune peut ne pas avoir lieu et engendrer des conflits sémantiques : différentes manières d'exprimer une date, donner une valeur sans préciser l'unité de mesure ou l'échelle, employer un mot qui a un homonyme de même graphie sans spécifier ce qu'il désigne, désigner une même chose par des synonymes – dans le cas d'acteurs différents, classer un objet dans des catégories différentes, pour ne prendre que ces exemples. Le langage naturel, souvent polysémique, favorise ainsi les confusions : « requête », par exemple, est un mot qui n'a pas le même sens pour un juriste et pour un informaticien. Les enjeux de l'interopérabilité sémantique sont donc d'une part de préciser ce que l'on entend par un terme que l'on emploie et d'autre part de préciser les relations hiérarchiques entre les termes que l'on emploie.

Le premier enjeu que nous évoquons est dédoublé lorsqu'on parle de métadonnées, et nous allons devoir ici recourir à la distinction entre propriété et valeur d'une propriété. En effet, préciser le sens d'un terme s'applique aussi bien aux propriétés qu'aux valeurs. Concernant les propriétés, les standards de métadonnées définissent généralement un « vocabulaire de métadonnées », qui est un ensemble fini de propriétés (parfois appelé « schéma », là aussi). L'utilisation des mêmes termes par plusieurs standards peut prêter à confusion. Pour l'illustrer, nous pouvons nous appuyer sur un exemple très parlant que présente Richard Gartner, à partir des différents sens de la propriété « titre » dans différents standards de métadonnées, que nous présentons dans le tableau ci-dessous.

Standard de métadonnées	Définition de « titre »
Dublin Core	Le nom donné à une ressource. Typiquement, un titre sera le nom sous lequel la ressource est officiellement connue
Règles de catalogage anglo-américain (AACR2)	Le nom premier [*chief name*] d'un item, en incluant tout titre alternatif mais en excluant les titres parallèles et autres informations sur le titre
Encoded Archival Description	Le nom, soit officiel soit donné, du document décrit
VRA Core (objets visuels)	Le titre ou une formule d'identification donnée à un travail ou une image
PBCore (Radiodiffusion publique)	Un nom ou un libellé pertinent pour la ressource

Tableau 1 : Définition de "titre" dans cinq standards de métadonnées (R. Gartner, 2016)

Comme le pointe Richard Gartner, « ces définitions varient notablement dans ce qu'elles considèrent comme un titre : elles y voient toutes une sorte de "nom" pour

une ressource mais diffèrent sur le nom en particulier qui devrait avoir ce statut. AACR2 se concentre spécifiquement sur le “nom premier”, en excluant par exemple ses équivalents dans une autre langue (appelé titre parallèle). VRA Core, un standard utilisé pour cataloguer des images, définit de manière indirecte et assez peu utile le titre comme un “titre”. PBCore, un standard clé dans l’industrie de la radiodiffusion, le considère de manière plus vague encore comme un nom “pertinent” pour la ressource, quoi que “pertinent” veuille dire. Donc même en connaissant le standard dans lequel un titre est défini, nous pouvons néanmoins être un peu troublé sur ce qu’il signifie. C’est peut-être là que la métaphore linguistique devrait abandonner “sémantique” pour “pragmatique”, l’étude de la manière dont la signification s’acquiert par le contexte du langage utilisé ». Il y a toutefois pour Richard Gartner une manière de sortir de cette difficulté : « parce qu’il est si important de connaître la provenance du nom d’un champ afin d’être au clair sur sa sémantique, le monde des métadonnées a trouvé une manière plus précise d’identifier celles-ci qu’un libellé lisible par les humains […] en utilisant une chaîne de lettres, de nombres et de signes de ponctuation connue sous le nom d’Identifiant de Ressource Uniforme [*Uniform Resource Identifier*] (habituellement abrégé en URI) ». Cet identifiant est une brique de base du paradigme des données liées, nous y reviendrons dans notre troisième partie.

Concernant les valeurs que peuvent prendre les propriétés, préciser les termes est l’objectif des « vocabulaires contrôlés ». Les vocabulaires contrôlés se présentent donc comme des ensembles finis de valeurs possibles pour une propriété. Dans la mesure où ces vocabulaires doivent permettre d’éviter les ambiguïtés lexicales, les termes employés ne doivent pas présenter de recouvrement sémantique (une même réalité qui serait représentée par plusieurs termes) ni de polysémie (un même terme qui représenterait plusieurs réalités). Les vocabulaires contrôlés peuvent se présenter sous forme de simple liste, mais présentent le plus souvent une structure. Plus cette structure est complexe, mieux elle est à même de représenter la complexité de la réalité représentée. Ainsi, une taxonomie introduit une hiérarchie entre les termes d’un vocabulaire contrôlé. Cette hiérarchie peut obéir à certaines contraintes, mais elle est globalement arborescente, selon un unique axe allant du moins générique au plus générique. Un thésaurus est plus complexe, permet de décrire plus précisément les termes en recourant aux notions de terme relié, de terme préféré, de terme plus large ou plus étroit, et en proposant une explication du terme. Une ontologie est plus ambitieuse puisqu’elle propose une description du monde, plutôt qu’une simple hiérarchisation de termes, en définissant des types d’objets, les propriétés qui les caractérisent, et les relations qu’entretiennent à la fois les objets entre eux et les propriétés entre elles. L’intérêt particulier des ontologies est de permettre des inférences. Pour prendre un exemple simple, supposons que nous définissions la classe « livre » et la sous-classe « roman », et que nous attribuions à la classe « livre » la nécessité d’avoir une propriété « auteur », nous n’avons pas besoin de spécifier qu’un roman doit avoir un auteur : c’est inféré de son appartenance à la classe des livres. Si un livre a également une propriété « langue d’écriture », on pourra définir des

sous-classes de « roman », par exemple « roman francophone » ou « roman anglophone », selon la valeur de cette propriété, et on pourra éventuellement définir une propriété « langue d'écriture » de l'auteur dont les valeurs seront induites des langues d'écritures des romans dont il est l'auteur.

La façon la plus simple d'assurer l'interopérabilité sémantique est bien entendu d'adopter entre différents acteurs un système de classification commun. Lorsqu'il s'agit d'assurer l'interopérabilité entre plusieurs systèmes de classification, la complexité de la tâche croît avec la complexité de la hiérarchie. Lois Mai Chan et Marcia Lei Zeng, dans leur communication *La réalisation de l'interopérabilité entre vocabulaires d'accès matière et système d'organisation de la connaissance : une analyse méthodologique*, font le point sur les méthodes pour rendre interopérables des vocabulaires :

- « Dérivation/Modélisation - Un vocabulaire spécialisé ou plus simple est développé à partir d'un vocabulaire existant plus complet comme point de départ ou modèle.
- Traduction/Adaptation - Un vocabulaire contrôlé est développé à partir des termes traduits d'un vocabulaire dans une langue différente, avec ou sans modification.
- Équivalences (Mapping intellectuel) - Un système de mapping est développé qui consiste fondamentalement à établir des équivalents entre les termes de différents vocabulaires contrôlés ou entre des termes et des indices de classification. Un tel mapping requiert généralement beaucoup d'effort intellectuel.
- Équivalences (Mapping assisté par ordinateur) - Un système de mapping est développé qui se fonde en partie ou fortement sur l'informatique.
- Maillage (Linking) - Une liste de termes est développée en reliant ces termes avec d'autres termes qui ne sont pas des équivalents conceptuels mais sont étroitement liés linguistiquement. De tels liens se sont avérés aptes à augmenter les résultats de la recherche.

Commutation (Switching) - Un langage ou un système de commutation est développé pour servir d'intermédiaire et se déplacer parmi des termes équivalents dans différents vocabulaires » (15, Chan et Zeng).

> Quant au « mapping de systèmes ayant des structures différentes », les auteurs constatent que « l'établissement de concordance ou de traduction entre un thésaurus et une classification ou parmi divers systèmes devient parfois impossible ou extrêmement difficile. C'est particulièrement vrai quand le système-cible à un niveau plus élevé de spécificité que le système-source ou d'autres systèmes en cause », ce qui les amène à trois options méthodologiques : « traduction ; fusion ; ou création à partir de zéro ». Nous verrons que les ontologies, par la liberté qu'elles laissent dans le langage de description mais aussi par la possibilité de les combiner ou d'établir des passerelles entre elles, permettent une plus grande souplesse, même si les problèmes de fond rencontrés ici subsistent.

Comme nous le disions, l'interopérabilité présente plusieurs aspects. Toutefois, ces aspects ne sont pas indépendants ; ils sont même, comme nous l'avons vu, fortement reliés. Nous pourrions parler de *degrés* d'interopérabilité, tant ils apparaissent comme des marches ou des étapes dont chacune doit être franchie pour accéder à la suivante, sans qu'il soit possible de modifier cette consécution. Sans interopérabilité technique, les métadonnées ne peuvent même pas être échangées. Sans interopérabilité syntaxique, nous savons que le document reçu contient les métadonnées que nous souhaitons, mais nous ne savons pas exactement où les trouver. Sans interopérabilité sémantique, nous savons où trouver les métadonnées mais nous ne sommes pas assurés du sens que nous devons leur donner. Comme nous le signalions, au-delà de ces trois degrés d'interopérabilité, certains ajoutent un quatrième degré qui prend en compte le fait que les métadonnées transférées parviennent éventuellement non pas à un individu, mais à une organisation dans laquelle chacun a son rôle. Le RGI la définit ainsi : « L'interopérabilité organisationnelle est liée aux organisations et aux processus notamment mis en œuvre pour favoriser et opérer les échanges. Elle concerne aussi les compétences et les connaissances associées au fonctionnement de ces organisations. En termes d'organisation, il s'agit par exemple de définir les rôles et les responsabilités des personnes qui prennent part à l'échange au sein de leur entité. En termes de processus il s'agit de définir qui envoie la donnée, à quel moment, suite à quel événement... mais aussi comment sont partagés les rôles et les responsabilités entre les différentes parties prenantes » (13, DINSIC, p. 12). Ce degré d'interopérabilité joue certainement un rôle essentiel dans le processus complet d'une transmission de métadonnées, mais il présente une différence notable qui nous a poussé à ne pas le traiter de façon plus approfondie : la manière dont une organisation définit les rôles et les responsabilités de chacun peut être modifiée sans que cela affecte les autres parties prenantes de l'échange de métadonnées. Remarquons enfin, pour y revenir dans notre troisième partie, que le paradigme des données liées apparaît comme un horizon souhaitable à tous les niveaux de l'interopérabilité.

Conclusion

Nous avons au cours de cette première partie tenté de préciser ce que sont les métadonnées, ce qui nous a d'emblée placé au cœur des enjeux qui les concernent. L'études des métadonnées d'un point de vue théorique nous a convaincu de l'intérêt de les penser dans le cadre d'un modèle distinguant données, informations et connaissances, mais en réaménageant la construction traditionnellement linéaire de ce modèle. Notre postulat de la prééminence de la question de l'accessibilité s'est trouvé confirmé par l'examen des finalités des métadonnées. Cette notion d'accessibilité doit toutefois être déclinée, au regard de nos développements, en deux dimensions : l'accès aux ressources comme fonction essentielle des métadonnées, qui était l'angle sous lequel nous avions d'abord engagé notre questionnement, mais également l'accès aux métadonnées elles-mêmes. Cette deuxième dimension des métadonnées se traduit souvent en termes d'interopérabilité, et ouvre sur la question de la réutilisation des métadonnées, ce qui fait apparaître la figure, ou plutôt les figures de tous les usagers possibles des métadonnées. C'est à cette aune que doit être repensé l'enrichissement des métadonnées, celui-ci étant nécessairement limité quand l'accès aux métadonnées est fermé à la réutilisation. L'élargissement des éventualités de réutilisation à travers la société mais aussi dans le temps long assigne aux métadonnées un rôle dans la culture collective. Pour ces trois ordres de raison, considérations épistémologiques, apparition de la figure de l'usager et rôle culturel, le cycle de vie des métadonnées doit — nous semble-t-il — être repensé en prenant en compte les retours, les modifications, les repentirs, les validations, tout en replaçant les métadonnées dans un espace commun (plus ou moins étendu et ouvert selon les circonstances, bien entendu, mais nous nous intéressons ici en priorité aux métadonnées du livre), où chacun a potentiellement un rôle à jouer dans leur création et leur entretien.

2. Deuxième partie

Après nous être intéressés aux métadonnées d'une manière générale afin d'en poser un cadre de compréhension et de préciser les dimensions qu'il convenait d'aborder à ce sujet, nous souhaiterions à présent préciser le champ de notre analyse à un domaine particulier : celui des métadonnées commerciales dans le domaine du livre. En effet, il nous a été donné l'opportunité, au cours de ce Master, d'effectuer notre contrat d'alternance au sein de l'entreprise Dilicom, dont le rôle principal est l'échange de données informatisées entre les distributeurs et les libraires. En nous focalisant à présent sur l'expérience professionnelle acquise au cours de ces deux années, nous allons être à même de préciser les enjeux de l'enrichissement des métadonnées d'une manière plus concrète, mais aussi d'aborder en détail ce qui signifie en pratique cet enrichissement.

Pourquoi et comment enrichir les métadonnées ? Nous avons déjà vu dans notre première partie plusieurs réponses à cette question. Pourquoi et comment enrichir les métadonnées des livres ? Cette question a été abondamment traitée dans le cas des bibliothèques et des métadonnées bibliographiques. Pourquoi et comment enrichir les métadonnées du livre dans le secteur commercial ? C'est la question à laquelle nous allons maintenant de tenter de répondre. Pour ce faire, nous devrons dans un premier temps présenter ce qu'il est de coutume d'appeler la « chaîne du livre », ses différents acteurs et leurs enjeux respectifs. Nous aurons l'occasion d'y présenter l'entreprise Dilicom. Nous avons eu la chance d'y vivre un moment exceptionnel dans la « vie » des métadonnées du livre matériel dans le secteur commercial : le changement du format dans lequel les métadonnées sont transmises par les distributeurs et diffusées aux revendeurs. En effet, alors que les métadonnées étaient historiquement diffusées dans différents formats dits « plats » ou « positionnels », nous avons pu participer à l'accompagnement des distributeurs et des revendeurs dans l'adoption du format standard ONIX (acronyme pour *Online Information Exchange*), dont le support est un fichier exprimé dans le langage de balisage XML. Nous consacrerons le deuxième temps de cette partie à étudier ce format, présenter les possibilités qu'il offre d'enrichir les données, ainsi que les changements auxquels il oblige les différents acteurs qui se mettent à l'employer. Puis dans le troisième et dernier temps de cette partie, nous nous attacherons aux aspects les plus pratiques de la mise en place d'un nouveau format en termes de gestion de projet et de conduite du changement. Nous serons ainsi à même d'étudier de manière très concrète ce qui se joue lors de la collecte, du traitement et de la diffusion des métadonnées.

2.1. La chaîne du livre : ses acteurs et leurs enjeux

2.1.1. Des acteurs différenciés

L'expression de « chaîne du livre » fait image : elle laisse entendre l'existence de différents maillons et la séquentialité du processus qui les lie. La réalité peut se révéler plus complexe, comme nous allons le voir, mais l'expression n'est pas fondamentalement trompeuse. Pour reprendre sa caractérisation par l'École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), la chaîne du livre désigne « le processus de production et de commercialisation du livre, considéré dans sa stricte dimension économique. […] dans sa configuration la plus élaborée, elle fait intervenir tour à tour l'auteur, l'agent, l'éditeur, l'imprimeur, le diffuseur, le distributeur et le point de vente. Chacune de ces entités perçoit une partie du prix du livre, et ce partage s'effectue sous des contraintes légales spécifiques au secteur » (16, ENSSIB). Vient ensuite la présentation des différents « maillons » de cette chaîne :

- l'auteur, qui « est le créateur de l'œuvre et en conserve […] la propriété intellectuelle, même si l'initiative du projet peut parfois revenir à l'éditeur (livres de commandes) » ;
- l'agent, « intermédiaire à qui l'auteur délègue la négociation des contrats », qui reste rare en France ;
- l'éditeur, qui « est, au sens économique, le vrai producteur du livre, même si une large partie du processus peut être sous-traité, en particulier auprès de l'imprimeur (maquettiste, infographiste, relieur, etc.), qui est le plus souvent un prestataire » ;
- le diffuseur, qui « prend en charge le démarchage des points de vente et la mise en place (nombre d'exemplaires commandés par ces derniers avant publication) du livre. L'éditeur lui délègue également la négociation des conditions commerciales (la remise, la faculté de retour et l'échéance du paiement) accordées aux points de vente » ;
- le distributeur, qui « gère les stocks de livres, répond aux commandes des points de vente (mise en place et réassort), les achemine vers la plateforme (points de vente en province) ou les confie au coursier (Île-de-France). Il se charge également du recouvrement des factures : c'est le créancier du point de vente » ;
- le point de vente (librairie, hypermarché, grande surface culturelle, maison de la presse, site Internet,etc.) qui « accueille les acheteurs et propose le livre dans les conditions prévues par la loi » ;
- l'acheteur « qui peut être un particulier ou une collectivité (bibliothèque, …) » (ibid.).

Notons d'abord que contrairement à d'autres secteurs marchands où la notion de distributeur renvoie au rôle de détaillant ou de revendeur, elle renvoie ici à l'aspect logistique de la chaîne d'approvisionnement ; ensuite que le rôle de diffusion commerciale, celui du diffuseur, est parfois assuré par le même acteur qui assure la distribution, qui est dans quelques cas l'éditeur également ; enfin qu'il est possible pour un éditeur d'être multi-distribué, pour un ouvrage d'être coédité ou coécrit.

Afin de compléter ces éléments, nous pouvons proposer, pour représenter la chaîne du livre, le schéma suivant, tiré du site du Ministère de la Culture[40], qui met en évidence l'existence d'autres acteurs (clubs, grossistes…) dont la présence est plus anecdotique mais complexifie la « chaîne » de chemins parallèles :

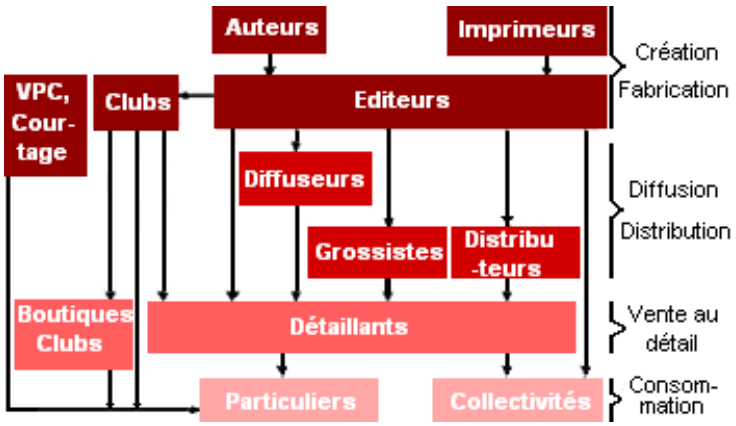

Figure 5 : Marché du livre (Ministère de la Culture)

Nous avons pris le parti de citer longuement cette fiche de l'ENSSIB qui permet déjà d'anticiper les enjeux qui vont être propres à chacun de ces acteurs. Il nous faut également, comme il est rappelé dans cette fiche, évoquer le cadre législatif spécifique au prix du livre en France : celui de la loi du 10 août 1981, restée dans le langage courant comme « loi Lang », qui « précise le principe du prix unique (rabais de 5% maximum pour les particuliers, ristourne de 9% maximum pour les collectivités) ; […] impose la supériorité de la remise qualitative sur la remise quantitative ; […] encadre les transactions (obligation de proposer le service de commande à l'unité pour les points de vente) », toutes dispositions qui visent à « à maintenir un réseau de librairies indépendantes contre le développement des grandes surfaces du livre ». Il est à noter que cette loi ne définit pas ce qu'est un livre. Nous ne pourrions pas lui reprocher de n'avoir pas anticipé le développement de l'édition numérique, mais comme le montre une question au gouvernement du député Jacques Pélissard en 2012[41], cette absence de délimitation juridique précise de l'objet « livre » peut poser certains problèmes et

[40] Ministère de la Culture. *Marché du livre*. [En ligne]. Disponible sur : <http://www.culture.gouv.fr/Thematiques/Livre-et-Lecture/Economie-du-livre/Marche-du-livre > (consulté le 20 août 2018)
[41] Disponible sur : <http://questions.assemblee-nationale.fr/q13/13-91286QE.htm> (consulté le 20 août 2018)

certains « franchissements de frontière » au cours du temps, en l'occurrence concernant les partitions de musique. Lors de notre alternance chez Dilicom, nous avons pu rencontrer à deux reprises des questionnements connexes, concernant le livre scolaire, qui est lui soumis à un régime dérogatoire du prix unique du livre dans le cas de la vente à des collectivités et dont le décret du 8 août 1985 modifié le 31 août 2004 (article D314-128 du Code de l'éducation) prétend préciser la définition : un manuel scolaire publié sous forme numérique bénéficie-t-il du même régime dérogatoire ? un manuel de catéchisme entre-t-il dans la définition d'un livre scolaire ?

Nous ne pouvons pas parler néanmoins de vide juridique total, puisque le livre fait l'objet d'une définition fiscale (Direction générale des impôts, 30 décembre 1971, 3C-14-71) : « Un livre est un ensemble imprimé, illustré ou non, publié sous un titre ayant pour objet la reproduction d'une œuvre de l'esprit d'un ou plusieurs auteurs en vue de l'enseignement, de la diffusion de la pensée et de la culture. Cet ensemble peut être présenté sous la forme d'éléments imprimés, assemblés ou réunis par tout procédé, sous réserve que ces éléments aient le même objet et que leur réunion soit nécessaire à l'unité de l'œuvre. Ils ne peuvent faire l'objet d'une vente séparée que s'ils sont destinés à former un ensemble ou s'ils en constituent la mise à jour. Cet ensemble conserve la nature de livre lorsque la surface cumulée des espaces consacrés à la publicité et des blancs intégrés au texte en vue de l'utilisation par le lecteur est au plus égale au tiers de la surface totale de l'ensemble, abstraction faite de la reliure ou de tout autre procédé équivalent ». Cette définition, nous l'avons vu, se révèle néanmoins insuffisante.

Le décret n° 2011-1499 du 10 novembre 2011, décret d'application de la loi relative au prix du livre numérique, donne comme pour la loi Lang le pouvoir à l'éditeur de fixer un prix unique pour les différents revendeurs. Toutefois, comme l'illustre le schéma ci-dessous, de vraies différences existent entre la chaîne du livre matériel et la chaîne du livre numérique, qui ne sont pas sans conséquence sur les questions de savoir qui transmet des métadonnées, qui en assure la qualité, qui en évalue la pertinence.

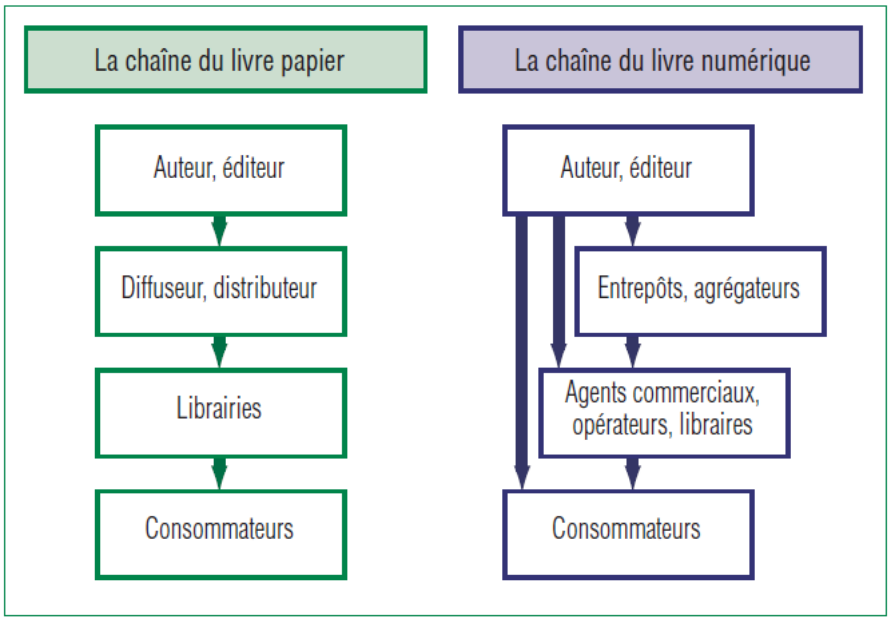

Figure 6 : Chaîne du livre numérique (Ministère de la culture et de la communication, 2010)

Notre devons ajouter à ce recensement des acteurs qui paraissent échapper au processus de production des livres mais sont pourtant producteurs de métadonnées : nous voulons parler des sites et applications de catalogage social (on parle également de *bookmarking* social), qui permettent aux usagers le partage de catalogues d'objets. En France, les plus connus dans le domaine du livre sont Babelio, Booknode, ou encore Sens Critique.

A l'issue de ce panorama de la chaîne du livre, nous voyons se dégager le contexte dans lequel nous allons pouvoir étudier les enjeux économiques qui structurent le marché du livre : des métiers variés, un cadre législatif favorable aux librairies indépendantes, et rappelons-le un secteur extrêmement concentré en ce qui concerne l'édition, la distribution et la diffusion. Le graphique ci-dessous, volontairement anonymisé, permet de le mettre en évidence pour le livre matériel : les lignes blanches les plus larges séparent les distributeurs, et à l'intérieur des rectangles ainsi constitués, les lignes blanches les plus fines séparent les marques éditoriales (pour rappel, un même éditeur peut éditer différentes « marques ») distribuées par un distributeur, la taille des pavés indiquant le nombre de livres édités sous cette marque éditoriale. Ce qui apparaît au premier abord est la similarité entre la distribution (au sens statistique) des distributeurs (de quelques très gros distributeurs à de nombreux petits distributeurs, selon une distribution de Pareto[42]) et pour la plupart des distributeurs la

[42] Le sociologue et économiste italien Vilfredo Pareto (1848-1923) est connu notamment pour son observation de la répartition de la richesse en Italie, 20% de la population possédant 80% des richesses et les 80% les moins fortunés se partageant les 20% restants. Cette « règle des 80-20 » a ensuite été reprise dans de nombreux domaines, notamment par l'un des fondateurs de la démarche qualité, Joseph Juran, mais également théorisée mathématiquement. Un de ses applications fameuses est la loi de Zipf, qui met en relation le nombre d'occurrences d'un mot dans un texte et son rang d'apparition (le n^{ième} mot apparaissant K/n fois, avec K une constante à déterminer). Une autre application fameuse est la théorie de la « longue traîne », concernant la distribution de ventes de produits : il y a distribution

distribution (toujours au sens statistique) des marques éditoriales (distribution de Pareto toujours), même si on remarque lorsque les distributeurs baissent en taille des cas plus fréquents pour lesquels un distributeur de taille moyenne ne distribue qu'une marque éditoriale ou presque. Pour fixer les proportions, le plus gros rectangle, en haut à gauche, représente plus de 140.000 titres édités par une marque éditoriale.

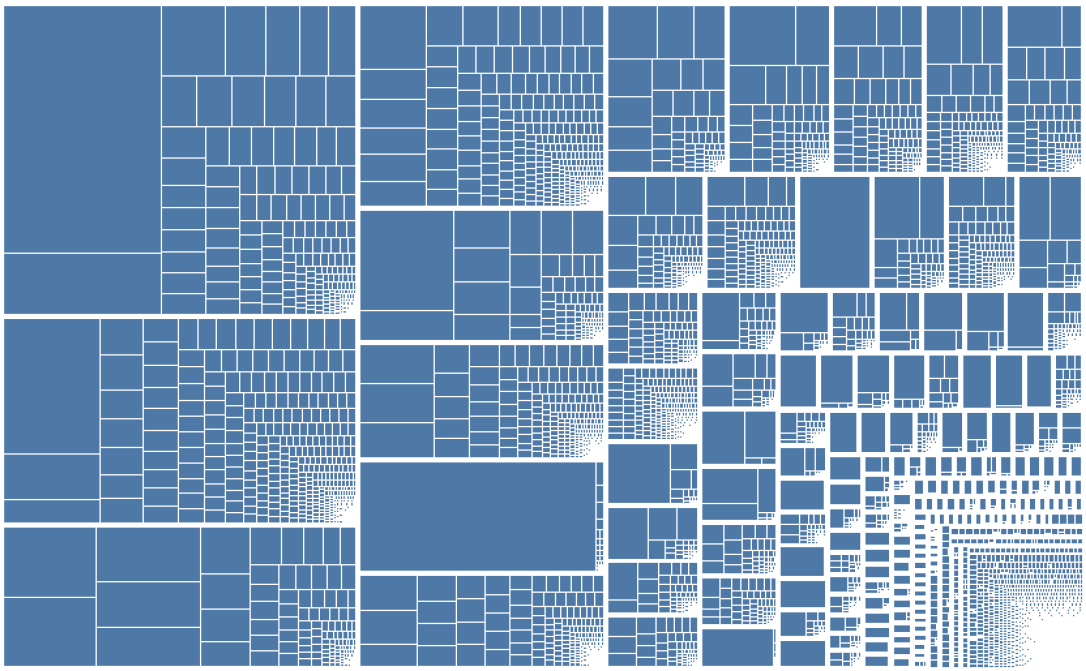

Figure 7 : Répartition des éditeurs par distributeurs du livre matériel

2.1.2. Les enjeux économiques

Avant de nous intéresser aux enjeux proprement commerciaux des différents acteurs, nous souhaitons reprendre le fil, ouvert dans notre introduction, de l'économie de l'attention. Georg Franck, architecte, économiste et philosophe allemand, a publié dès 1993 le texte «Ökonomie der Aufmerksamkeit », dans lequel il formalise un cadre conceptuel pour penser l'économie de l'attention, et dont la traduction constitue de

---

à « longue traîne » lorsque de très nombreux produits se vendent en petite quantité, constituant une masse plus importante que les produits se vendant pourtant en bien plus grande quantité. Cette théorie s'applique notamment aux produits culturels, les *best sellers* ou les *blockbusters* atteignant des scores de vente incomparables, mais représentant finalement moins que la somme de tous les autres produits, qui se vendent pourtant beaucoup moins individuellement.

La distribution que nous considérons ici est encore plus « écrasée » qu'une distribution obéissant à la loi de Zipf. Si nous ne prenons que les 109 distributeurs distribuant plus de 1000 ouvrages, la règle de Pareto semble respectée : les 21 premiers représentent 80% de la distribution de livres des 109. Si nous considérons l'ensemble des distributeurs, nous voyons que nous ne pouvons pas parler de distribution à longue traîne : les 90% des distributeurs qui distribuent le moins d'ouvrages ne représentent en volume que 8,9% de l'ensemble. La règle qui s'applique dans le domaine de la distribution serait plutôt le 90-10.

chapitre 2 de l'ouvrage collectif dirigé par Yves Citton « Économie de l'attention » (17, FRANCK). Georg Franck commence par y rappeler que « dans la société d'abondance, il est de plus en plus commun de classer le revenu en attention plus haut que le revenu en argent » et que lorsque « de plus en plus de personnes peuvent prétendre à la richesse matérielle, alors le désir de distinction doit tendre vers d'autres attributs, plus sélectifs que le revenu monétaire » (ibid., p. 55). Désignant du terme de « prééminence » le fait de disposer de revenus d'attention élevés, il envisage une objection qu'on pourrait lui opposer : « la prééminence est essentiellement une distinction qualitative. À la différence de la richesse, elle ne peut devenir un phénomène de masse » (ibid., p. 56). Il écarte toutefois cette objection en arguant que « l'individu prééminent n'est plus seulement celui qui est en route vers l'apogée de la gloire ou du pouvoir » (ibid., p. 56), mais qu'une carrière standardisée permettant la présence médiatique suffit. Cet argument était nécessaire pour asseoir la possibilité d'une économie de l'attention à l'échelle d'une société entière, puisqu' « il faut des individus prééminents en masse, si on veut exploiter l'attractivité de l'attention comme un commerce de masse » (ibid., p. 58). L'intérêt heuristique de cette approche est de faire apparaître le succès financier d'un médium comme enté sur sa capacité à attirer l'attention. C'est cette dernière qui permet de valoriser l'offre de surfaces publicitaires. Ce qui fait dire à l'auteur que « les revenus d'attention priment sur le succès financier, pour le medium lui-même » (ibid., p. 57). Il produit ensuite ce que nous pouvons considérer comme une définition des médias du point de vue de l'économie de l'attention, en affirmant que « Les médias ne sont nullement de simples lieux de transit de l'information. Ils sont des canaux qui capturent les émotions et les sensations en les approvisionnant en information, afin d'y puiser de l'attention » (ibid., p. 58).

Il en vient dans la suite de l'article à une analyse parfaitement topique pour notre sujet, en étudiant la puissance des médias pour promouvoir des individus prééminents. Il remarque que « cette puissance, c'est seulement peu à peu que les médias l'ont accumulée. La reproduction mécanique de l'écrit, du son et de l'image en marque le point de départ technique » (ibid., p. 58). En effet, cette reproduction mécanique permet la démultiplication d'une présence individuelle auprès d'un grand nombre d'individus susceptibles d'y accorder une part de leur attention disponible. Il poursuit donc en remarquant que « ce n'est pas la demande d'information en tant que telle qui a rendu les médias puissants. Ce qui les a rendus puissants et ce qui continue d'assurer leur croissance, c'est l'idée ingénieuse d'offrir au public de l'information pour obtenir son attention » (ibid., p. 58). Il prend alors comme exemple le secteur de l'édition : « Sans le revenu d'attention que la diffusion publique garantit, le secteur de l'édition n'aurait pu se développer de manière significative. Si seul ce qui promettait un succès commercial avait été publié dans les livres et les revues, le paysage littéraire serait très différent de ce qu'il est aujourd'hui. C'est seulement parce que les auteurs espèrent se faire payer en revenus d'attention qu'on peut expliquer leur acceptation de salaires de misère pour se tourmenter l'esprit dans la recherche de mots justes. L'ingéniosité du commerce éditorial tient à la division des recettes en revenu monétaire

et en revenu d'attention. La condition de production de notre culture littéraire consiste, pour schématiser, en ce que l'éditeur reçoit l'argent et l'auteur l'attention » (ibid., pp. 58-59). A ce point de son développement, il présente donc des intérêts très différenciés pour l'éditeur et l'auteur, le premier qui chercherait uniquement des revenus monétaires, quand le second viserait à obtenir des revenus d'attention. Sans cette dernière condition, autrement dit si un auteur cherchait lui-même des revenus monétaires en se faisant éditer, Georg Franck semble penser que le secteur éditorial ne fonctionnerait pas. Il tempère toutefois immédiatement cette analyse contrastée en ajoutant : « Si, en outre, l'éditeur gagne en réputation et l'auteur en richesse, alors cela entraîne, économiquement parlant, un profit supplémentaire, qui n'est toutefois pas indispensable à la bonne marche du commerce » (ibid., p. 59). Le commerce du livre est sans doute possible sans cela, mais c'est bien cette condition, celle d'un mélange de revenus d'attention et de richesse, pour les auteurs aussi bien que pour les éditeurs, qui lui permet d'être florissant, comme il le note ensuite : « C'est précisément ce calcul mixte qui est à l'origine de la transformation de l'organe de publication en medium de masse. Un medium de masse ne peut pas être exigeant dans le choix des moyens par lesquels il capte l'attention. L'auteur, en revanche, a besoin que l'attention obtenue pour son œuvre ou sa personne soit en adéquation avec son être : c'est pourquoi le désir d'attention est étroitement lié à celui de la réalisation de soi. Or, on ne mobilise guère les masses en travaillant à la réalisation de soi. On les mobilise uniquement si on considère avec la plus grande précision ce que le public veut lire, entendre et voir. […] La production pour ce goût médiatisé nécessite bel et bien des esprits créatifs, parmi lesquels il faut en trouver qui soient prêts à servir une cause qui n'est pas la leur. Et c'est leur disponibilité qui doit être assurée par l'argent » (ibid., p. 59).

Nous pouvons tenter de transposer le « calcul mixte » que décrit Georg Franck dans le champ des métadonnées, en nous appuyant sur une distinction, au sein des métadonnées descriptives, entre les métadonnées « distinctives », qui permettent de distinguer une ressource des autres et sont de nature à singulariser un produit, et les métadonnées « de regroupement », qui constituent des points communs entre des ressources et sont au contraire de nature à permettre d'opérer des regroupements entre des produits[43]. Ainsi, le titre, les dimensions, la date de publication, sont des métadonnées distinctives, et seraient donc dans la logique de la recherche prioritaire d'un revenu d'attention favorisées par l'auteur. A l'autre pôle, celui de la recherche de revenu monétaire à travers une captation de l'attention plus « généraliste » visée prioritairement par l'éditeur, on trouverait plutôt les métadonnées permettant des regroupements, comme un thème, des mots-clés, la mention d'une collection. Comme nous le voyons, ces métadonnées spécifient également la ressource, mais permettent aussi de découvrir cette ressource à partir d'autres ressources. La frontière n'est pas forcément toujours évidente à tracer entre ces deux types de métadonnées mais pour

[43] Nous pourrions également, si nous considérions qu'il y a ici une lutte pour l'attention entre individus, parler en termes plus psychologiques de métadonnées « égoïstes » et de métadonnées « altruistes ».

prendre un exemple simple, que deux livres fassent le même poids n'est pas une raison de s'intéresser à l'autre si l'un des deux nous a intéressé. En revanche, que deux livres fassent partie de la même collection éditoriale « humanités numériques » peut être une raison d'aller consulter la notice de l'autre si l'un des deux a retenu notre attention. Si nous cherchons à étendre aux autres acteurs de la chaîne du livre cette problématique, il nous semble que les diffuseurs, les distributeurs et les revendeurs partagent essentiellement les mêmes intérêts que les éditeurs. Les auteurs, selon Georg Franck, recherchent d'abord un revenu d'attention mais sont prêts contre un revenu monétaire à transiger sur le désir d'attention pour eux-mêmes. Les acheteurs, quant à eux, sont ceux qui accordent ou pas une partie de leur attention, et de leurs revenus, aux produits. Tout au moins tant qu'ils restent acheteurs. En tant qu'ils participent au *bookmarking* social sur des sites web, nous pouvons considérer qu'ils cherchent, dans la posture critique qui leur est proposée, d'affirmer un « goût », et qu'ils visent donc un revenu d'attention (ils ne peuvent d'ailleurs espérer en obtenir d'autre) pour l'éminence de ce goût. S'il s'agit bien, dans un catalogue social, de regrouper des œuvres, ce regroupement lui-même doit être singulier, et s'appuiera donc de façon préférentielle sur la singularité des ouvrages regroupés, plutôt que sur une manière de les regrouper qui serait de nature à effacer la spécificité de ce goût (une simple thématique, par exemple). Le site lui-même, en revanche, doit pouvoir couvrir le spectre le plus large possible et donc proposer des évaluations, des catalogages, concernant le plus d'ouvrages possibles. Ils sont dans une posture plus proche de celle de l'éditeur.

Les hypothèses que nous avons produites à partir du socle théorique fourni par l'économie de l'attention peuvent être confrontées à une certaine réalité du marché telle qu'elle apparaît dans les études consacrées à la consommation. Un des acteurs majeurs de ce secteur, Nielsen, a produit depuis 2012 des « Livres Blancs » sur l'importance des métadonnées dans la facilité d'accès (*discoverability*) et dans les ventes des livres, aux États-Unis et en Grande-Bretagne (18, Walter). C'est à ce dernier, plus détaillé, que nous ferons référence ici. Dès le premier Livre Blanc, paru en 2012, Nielsen est parvenu à mettre en évidence « un lien fort entre la complétude des données pertinentes et les ventes obtenues », ce qui leur permet d'affirmer dès l'introduction que « fournir des métadonnées complètes et pertinentes aide la commercialité [*tradability*] et la facilité d'accès [*discoverability*] des titres » (ibid., p.3). « Commercialité » indique ici « la facilité avec laquelle les produits peuvent être identifiés et commercialisés, et parcourir la chaîne d'approvisionnement du livre » (ibid., p. 3). Le Livre Blanc rappelle quelques complexités spécifiques du marché du livre : des millions de produits vendables séparément émanant de plusieurs milliers d'éditeurs différents, des librairies qui peuvent proposer des dizaines de milliers de titres présents sur leurs rayons en seulement un ou quelques exemplaires. La conséquence en termes de commandes et de reconstitution des stocks est qu'elles sont de type « peu et souvent ». A cela s'ajoute la complexité des « retours », qui permet à un libraire de retourner - dans un certain délai - au distributeur les livres qu'il

n'a pas vendus. « Ces facteurs signifient que la création d'une chaîne d'approvisionnement durable pour le commerce du livre requiert de l'attention, de la planification et de la coopération entre les parties prenantes » (ibid., p. 3). L'importance du numéro international standardisé du livre (*International Standard Book Number*, ou ISBN), comme clé fondatrice pour identifier de façon univoque un ouvrage et lui associer des attributs, est rappelée. Pour le revendeur, des données exactes concernant la date de publication, le prix, le fournisseur, les dimensions, le poids, permettent de gérer son stock en planifiant ses futures commandes ou en organisant ses rayons, et de s'assurer que les frais d'expédition sont les moins élevés possibles. Faute de ces données, le revendeur perd en efficacité ou peut hésiter à commander un produit. La facilité d'accès, la facilité avec laquelle un produit en particulier peut être trouvé, peut concerner aussi bien les partenaires commerciaux engagés dans le commerce du livre que les consommateurs finaux acquérant un titre. Nielsen remarque, dans la logique de l'opposition que nous tracions entre métadonnées distinctives et métadonnées de regroupement, que la recherche par métadonnées peut être reliée à un titre spécifique ou s'appuyer sur un critère plus général qui débouche sur l'identification d'un titre correspondant au goût ou au besoin du client. Les deux qualités dégagées, facilité d'accès et facilité à commercialiser un livre, « reposent fortement sur des métadonnées pertinentes, exactes et fraîches » (ibid., p. 4). Nous retrouvons les principales dimensions de base de la qualité des données en général[44].

Si l'on se place à présent du côté des consommateurs : comment trouvent-ils les livres ? Selon Nielsen, environ 30% des achats interviennent après avoir exploré une bibliothèque, une librairie, un catalogue, un site web…, environ 20% en connaissant déjà l'auteur ou la série, et presque 20% à nouveau après une recommandation (notamment du bouche-à-oreille dans deux cas sur cinq) ou une critique (de journal ou de magazine dans un cas sur cinq). D'un point de vue méthodologique, c'est le lien entre métadonnées et ventes (ou emprunt, car l'étude porte également sur les bibliothèques) pour un produit identifié par un ISBN, qui est la base de l'étude. Le contexte britannique est à prendre en compte, dans la mesure où l'industrie du livre, par l'intermédiaire de l'organisation Book Industry Communication (BIC) a édicté un standard sur un périmètre de données essentielles pour identifier et commercialiser un produit (nous verrons plus loin qu'en France, la Commission de Liaison Interprofessionnelle du Livre en fait autant), et que c'est à cette aune que Nielsen juge de la complétude et de la fraîcheur des métadonnées associées à un ouvrage. Le standard BIC Basic comprend les éléments suivants :

- ISBN

[44] Ces dimensions sont l'exactitude, la complétude, la fraîcheur et la cohérence. Voir par exemple : BERTI, Laure et THION, Virginie. *Fiche ExQI Comprendre : les dimensions de la qualité des données*. [En ligne]. Août 2013. Disponible sur : <http://exqi.asso.fr/site/medias/641739FicheDimensionsQualite.pdf> (consulté le 20 août 2018)

- Titre
- Catégorie de produit
- Catégorie thématique BIC principale (le BIC propose une classification thématique)
- Marque éditoriale
- Date de publication
- Image de couverture
- Nom de fournisseur (au moins un)
- Disponibilité
- Prix mentionnant la part de TVA
- Droits de vente

Dans cette liste, une présence et une absence sont remarquables : la mention de l'image de couverture, qui contrairement aux autres données n'est pas textuelle et nécessite donc un mode de transmission particulier ; les dimensions et le poids, dont l'importance pour le revendeur a été rappelée plus haut (mais nous pouvons former l'hypothèse qu'il s'agit de disposer d'un standard valable aussi bien pour le livre matériel que pour le livre numérique).

Si l'écrasante majorité des titres dans le périmètre de l'étude Nielsen comporte une image et des données BIC complètes, il apparaît néanmoins que la présence d'une image est corrélée avec des ventes deux fois plus élevées pour ISBN présentant des données BIC incomplètes, et que la présence des données complètes et d'une image de couverture est corrélée avec des ventes moyennes par ISBN près de quatre fois plus élevées par rapport aux ISBN sans image de couverture et aux données BIC incomplètes. Un autre critère testé par l'étude est lié au fait que « les titres connaissant typiquement leurs ventes les plus fortes dans les semaines qui suivent immédiatement la publication, quand le marketing et les activités de promotion sont à leur plus haut ». Le standard BIC est de ce point de vue que les données essentielles soient présentes seize semaines avant la date de publication. 70% des titres étudiés remplissent cette condition, qui est corrélée à une augmentation de ventes d'environ 6%.

Au-delà du standard BIC Basic, l'étude Nielsen s'intéresse également à un critère de conformité au format ONIX, critère qui porte sur la transmission des métadonnées dans un format ONIX valide et qui est plus exigeant du point de vue bibliographique puisqu'il prend en compte la présence de données descriptives des ouvrages. Là encore les différences sont très marquées puisque la conformité à l'ONIX est corrélée avec des ventes par ISBN deux fois plus élevées, et la transmission des métadonnées dans les délais standards (toujours seize semaines) est ici corrélée à une augmentation supplémentaire des ventes de 25%.

En ventilant par grands « genres » (Fiction, Non-fiction, Non-fiction spécialisée et Littérature pour enfants), les différences sont à nouveau spectaculaires. La base de comparaison n'est plus tout à fait la même, puisque les titres qui soit n'ont pas d'image, soit n'ont pas des données BIC complètes (ce qui fusionnent ces deux facteurs) sont comparés à des titres comportant une image et des données BIC complètes. Le

rapport de vente moyenne est toutefois de un à quatre pour le genre « Fiction », de un à près de trois pour la non-fiction commerciale, de un à plus de deux pour la littérature pour enfants, de un à un et demi seulement pour la non-fiction spécialisée. Les mêmes rapports se retrouvent dans les différents genres pour la comparaison des titres non conformes au standard ONIX avec ceux à la fois conformes à l'ONIX et dont les métadonnées sont délivrées en temps opportun.

Nielsen, se restreignant ensuite aux métadonnées conformes à l'ONIX, étudie l'effet du nombre d'éléments descriptifs additionnels (la conformité à l'ONIX exige qu'il y en ait au moins un), en rappelant que « des éléments descriptifs additionnels sont précieux pour la plupart des titres, et ajoutent à la complétude et à la richesse des données. Ceci se traduit par une facilité d'accès accrue à la fois pour les acheteurs professionnels de livres et pour les consommateurs. Les éléments descriptifs ici considérés sont : une description courte, une description longue, une notice biographique d'auteur et des revues critiques. La table des matières, qui n'est pas forcément pertinente pour certains titres et d'une manière générale peu communiquée dans les métadonnées, est laissée de côté. Un accroissement des ventes est clairement lisible en passant de la présence de zéro à quatre de ces éléments descriptifs dans les métadonnées. En croisant les facteurs « genre » et « nombre d'éléments descriptifs », les augmentations de vente corrélées sont claires pour tous les genres, et particulièrement accentuées dans le genre Fiction. Pour l'emprunt en bibliothèque, sur lequel porte également l'étude, les mêmes tendances sont présentes.

La conclusion de l'étude se veut épistémologiquement précautionneuse, en rappelant qu'il n'est pas en l'occurrence possible de « mesurer une causalité directe, mais seulement d'identifier la corrélation entre la présence de métadonnées et les ventes »(ibid., p. 18), en signalant toutefois que le constat de cette corrélation en utilisant différentes mesures des métadonnées des ouvrages et en segmentant les données de différentes manières donne des raisons de croire qu'il y a des indicateurs claires de ce lien. La conclusion que nous pouvons en tirer, en ce qui nous concerne, est que l'accent est mis sur les métadonnées que nous avons appelées « distinctives » d'un ouvrage : l'image de couverture et les éléments descriptifs (descriptions, notice biographique d'auteur, revues critiques), sont toutes de cette sorte. Les métadonnées que nous avons appelées « de regroupement » sont également présentes, nous pouvons le voir dans le standard BIC Basic (la catégorie thématique BIC, ainsi que la marque éditoriale), mais ne sont pas mises à part. Leur influence propre est donc difficile à déterminer, même si nous pouvons néanmoins remarquer que ce sont les plus susceptibles d'intéresser le consommateur final, la plupart des autres données du BIC Basic relevant plus du domaine des données « métier », utiles aux professionnels mais pas discriminantes pour opérer un choix. Remarquons enfin, avant d'y revenir plus loin, l'importance capitale de l'image de couverture, et celle également du format ONIX.

L'étude Nielsen reste centrée sur le rapport des métadonnées du livre à l'objectif de vente, et s'adresse donc prioritairement aux éditeurs ou aux distributeurs de livre.

Les enjeux économiques liés aux métadonnées dans le secteur commercial du livre existent toutefois en dehors de cette problématique, comme le montre bien l'article de Souad Odeh et Ghislaine Chartron *Acteurs et économie des métadonnées du livre en France : analyse et avenir* (19, Odeh et Chartron), qui présente pour nous l'intérêt de déplacer le questionnement de la place des métadonnées dans la vente de livres vers celui du commerce des métadonnées elles-mêmes. Leur étude, essentiellement consacrée au cas du livre numérique, offre néanmoins un panorama des acteurs du commerce marchand et non-marchand des métadonnées et pose des questions qui s'appliquent pour l'essentiel au cas du livre matériel : « le modèle économique des métadonnées des livres en France articule le non-marchand et le marchand, ce qui n'est pas nouveau mais une tension grandit quant à la valorisation possible dans un contexte où la gratuité est de plus en plus revendiquée. Qui sont aujourd'hui les principaux acteurs de chaque type d'offre ? Comment évaluer leur positionnement ? Quelle valeur économique est accordée aux différents services ? » (ibid., p. 23). Les « principaux acteurs de l'offre marchande » distingués sont Dilicom (dont nous reparlerons bientôt), Électre (dont nous reparlerons bientôt également), Decitre (entreprise de librairie) et Tite-Live (entreprise proposant des logiciels pour l'industrie des médias). Quant aux « cibles » de ce marché, ce sont « les bibliothèques, les libraires, les sites de commerce culturel en ligne ainsi que les éditeurs » (ibid., p. 24), qui présentent des besoins différenciés. « Les bibliothèques achètent les métadonnées pour assurer trois fonctions principales : la gestion des acquisitions, le catalogage des nouveautés et la gestion des ressources numériques […] Les détaillants des livres sollicitent l'offre marchande des métadonnées pour l'identification de l'offre éditoriale, le suivi des nouveautés, la commande des livres, l'alimentation de leur propre base de données et la diffusion de leur catalogue sur leur site Web (classement de l'offre, recommandation) […] Les éditeurs ont aussi besoin d'assurer la visibilité de leur offre éditoriale en fournissant à tous les acteurs de la chaîne de livre des métadonnées de qualité » (ibid., p.25). Les auteurs proposent un tableau qui met en regard des besoins en métadonnées les offres possibles et leurs valeurs respectives, en se référant au « concept "d'analyse de la valeur" introduit aux États-Unis par Lawrence Delos Miles et développé en Europe dans les années 1960[45] » (ibid., p. 24). Comme le signale ensuite les auteurs, « parallèlement à une production marchande, il existe en France une offre de métadonnées produite par des acteurs publics, principalement des bibliothèques qui créent des notices catalographiques à des fins de gestion de collections. Deux acteurs publics principaux produisent des métadonnées catalographiques, la Bibliothèque nationale de France (BnF) et l'Agence bibliographique de l'enseignement supérieur (ABES) » (ibid., p. 26). Dans la suite de l'article, des convergences sont envisagées entre les différents acteurs (nous y

[45] L'analyse de la valeur d'un produit rapporte la satisfaction qu'apporte le produit au besoin du client au coût de la solution mise en œuvre. Deux leviers existent donc pour augmenter la valeur d'un produit : trouver des solutions moins coûteuses sans faire baisser la satisfaction fonctionnelle et/ou faire augmenter la satisfaction fonctionnelle sans engager de solutions plus coûteuses.

reviendrons en 3.2) et un schéma est proposé des relations entre ces acteurs, marchands et non marchands, ainsi que des réseaux sociaux numériques et Google. En effet, « Google, le grand moteur de recherche sur le texte intégral, reconnaît l'importance des métadonnées pour organiser l'accès au contenu numérisé. Les métadonnées sont utilisées pour l'identification de livres et la création de liens. Elles sont fournies à Google en format MARC par les bibliothèques partenaires du projet de numérisation ou en format ONIX par les éditeurs avec lesquels un contrat a été négocié […] Enfin, Babelio, réseau socionumérique du livre, est l'exemple d'un nouvel acteur producteur de métadonnées sociales (avis d'internautes, critiques de professionnels) émises par les lecteurs de son réseau et pouvant être intégrées dans les services de bibliothèques ou d'éditeurs. Le service Babelthèque permet ainsi aux bibliothèques d'enrichir leur OPAC (Online Public Access Catalog) et leur site internet en important le contenu produit par les membres de la communauté de lecteurs Babelio.com » (ibid., p. 29).

2.1.3. La place de Dilicom dans la chaîne du livre

Selon l'article 3 des statuts de la société Dilicom, elle a pour mission « la réalisation de toutes opérations destinées à faciliter les rapports entre les distributeurs, les éditeurs et les libraires par les moyens de l'informatique, de la télématique, et de toute autre technologie, plus particulièrement pour accélérer les processus administratifs du commerce du livre (commandes, factures, etc…) et la diffusion des catalogues électroniques ». Cette définition juridique donne une vision du rôle de Dilicom, mais pour mieux comprendre le rôle et la place de Dilicom dans le secteur commercial du livre, un rappel historique n'est pas inutile. L'histoire de Dilicom commence en 1989, mais elle succède à EDILECTRE, filiale de la base bibliographique Electre. Comme le rappelle l'ALIRE (Association des Librairies Informatisées et utilisatrices de Réseaux Electroniques), « ELECTRE a revendu aux distributeurs et aux libraires (via l'ALIRE) sa filiale créée pour l'activité de commandes EDI, devenue EDILECTRE, puis transformée en DILICOM. Fortement déficitaire les premières années, DILICOM est devenue une SAS équilibrée, voire rentable. Toutefois, depuis l'origine, aucune distribution de bénéfices n'a été faite aux actionnaires de DILICOM. Les bénéfices et gains de productivité ont toujours fait l'objet de baisses de tarifs pour les utilisateurs, d'investissements ou d'augmentation de services et de messages pour le même coût à l'exemple du nouveau modèle économique décidé en 2006 : la seule ligne de commande payée supporte les coûts de l'ensemble des autres messages EDI, dont la facture dématérialisée. Les gains sont redistribués aux usagers des services de DILICOM » (20, ALIRE). Il s'agit comme nous le voyons d'une structure commerciale assez originale, dont les actionnaires (distributeurs et associations de l'interprofession dont l'ALIRE) sont les clients. Son activité principale est l'Échange de Données Informatisées (EDI) entre les distributeurs et les revendeurs. Les flux transitent dans

les deux sens. Dilicom reçoit des revendeurs leurs commandes de réassort, qu'elle retransmet aux distributeurs, et reçoit des libraires les mises à jour des fiches-produit du Fichier Exhaustif du Livre (FEL), qu'elle retransmet aux revendeurs. Le FEL, base de données à vocation commerciale est à distinguer d'Électre, comme le rappelle l'ALIRE : « tandis qu'ELECTRE est une base « bibliographique » (issue de Bulletin du Livre de France et du Catalogue des Livres disponibles) consultable sur abonnement, le FEL est une base « commerciale » appuyée sur l'Échange de Données informatisé (EDI) entre les libraires et les distributeurs, base consultable gratuitement pour tout revendeur sur le site Internet de DILICOM. Il faut noter la différence entre ce fichier commercial, très simplifié mais garanti par les distributeurs, et les fichiers bibliographiques, qui comprennent de nombreuses rubriques très élaborées (traduction, résumés, critiques, etc.) » (ibid.). Notons qu'il existe aussi un FEL « numérique », pour les livres numériques. Toutefois nous avons vu plus haut la différence entre les chaînes du livre matériel et du livre numérique, ce qui explique le rôle moins central de Dilicom dans ce contexte, les plateformes de vente des éditeurs ou des distributeurs, voire la présence d'auteurs autoédités et auto-distribués sur des *MarketPlace*s comme celui d'Amazon permettant une vente plus directe aux consommateurs finaux. Aux deux activités de Dilicom que nous avons citées, nous pourrions ajouter entre autres la transmission d'autres messages EDI, comme les avis d'expédition, et d'autres services, comme les factures dématérialisées, le logiciel PLUME de prise de commande pour les représentants, la gestion du Prêt Numérique en Bibliothèques. A la fois en raison du rôle qui a été le nôtre au sein du Pôle Projets et gestion des bases de données, et du sujet que nous abordons dans le présent travail, c'est bien entendu l'activité de gestion du FEL qui retiendra particulièrement notre attention. Si nous continuons un instant d'examiner les deux activités principales de Dilicom, nous pouvons remarquer qu'elles représentent une volonté collaborative de mutualisation de la part des distributeurs et des revendeurs. Comme le remarque l'ALIRE, « il y a des centaines de milliers de titres disponibles en France, et chaque jour ce sont des milliers de mises à jour qui sont nécessaires. Compte tenu du caractère très hétérogène de l'édition et de la distribution, il est très difficile pour un libraire d'être certain de mettre à jour correctement ses fichiers. Voilà pourquoi DILICOM est chargée de rassembler puis de diffuser des catalogues électroniques, à vocation commerciale, de façon à permettre à un libraire de bien gérer ses commandes » (ibid.). Nous avons étudié plus haut la structure de la distribution et de l'édition de livres, qui laissait apparaître un grand nombre d'acteurs très diversifiés. Nous pourrions en dire autant de la revente de livres. Nous comprenons aisément la complexité que représenterait le fait pour chaque libraire de passer ses commandes à chaque distributeur, et pour chaque distributeur de transmettre les mises à jour des fiches-produit des ouvrages à chaque libraire. L'existence de Dilicom permet aux revendeurs de disposer d'un destinataire unique auquel transmettre leurs commandes de réassort, et aux distributeurs de disposer d'un destinataire unique de leurs mises à jour de fiches-produit, la valeur ajoutée de Dilicom étant notamment, dans ces deux

processus, d'accepter différents formats en entrée de la part des expéditeurs et de fournir différents formats en sortie selon les souhaits des récipiendaires. Cette mutualisation, si nous la jugeons du point de vue des systèmes d'information, fait de Dilicom une partie d'un système d'information logistique, cette logistique étant collaborative puisqu'elle réunit des acteurs de différentes entreprises.

Dilicom peut donc être vue comme un carrefour, un *hub*, entre distributeurs et libraires, au sein de l'interprofession (Syndicat National de l'Édition, Syndicat de la Librairie Française, Commission de Liaison Interprofessionnelle du Livre - CLIL, ALIRE) et au-delà, puisqu'elle est également en interaction avec la BnF dans l'échange de métadonnées, ou avec EDItEUR, qui maintient le format ONIX - dont nous allons immédiatement parler. Elle a une activité de normalisation et de standardisation des métadonnées, notamment par le rôle qu'elle joue au sein de la Commission FEL de la CLIL. Donnons pour finir quelques chiffres sur le volume du FEL : il comprend plus de 2,2 millions de fiches-produit actives pour les livres physiques et plus de 500.000 notices pour les livres numériques ; il est alimenté par plusieurs millions de mises à jour par an (quatre millions en 2013) ; il contient les fiches-produit de 3.500 distributeurs, pour 9.500 marques éditoriales (21, Backert).

2.2. Le format ONIX

2.2.1. ONIX et EDItEUR

Nous avons envisagé dans la première partie les standards, soit les formats standardisés de métadonnées, d'un point de vue général. Nous en venons maintenant à un standard en particulier, que nous avons appris à connaître au cours de notre alternance. Un standard, nous l'avons signalé, concerne une communauté d'utilisateurs réunis par les problématiques communes. Il évolue au cours du temps pour s'adapter à ces problématiques et aux usagers, ce qui suppose un organisme qui assure cette évolution. Dans le cas de l'ONIX, cet organisme s'appelle EDItEUR. Il nous semble intéressant de laisser EDItEUR présenter ce standard et l'interaction des différents acteurs, comme il le fait dans sa Foire Aux Questions (22, EDItEUR) :

« Qu'est-ce qu'ONIX ?

ONIX – plus spécifiquement 'ONIX Livres' – est une spécification standard de communication des métadonnées de livres et de livres numériques entre éditeurs, intermédiaires divers comme les distributeurs, les grossistes et les entreprises de service de données, et les revendeurs dans la chaîne d'approvisionnement du livre. Les métadonnées – les informations sur chaque livre – sont devenues vitales à la bonne marche de la chaîne d'approvisionnement et à l'efficacité du marketing, du merchandising, de la vente de livres et de produits liés. Étant donné la quantité énorme de livres et de livres numériques disponibles sur le marché et le large éventail d'informations sur chacun d'entre eux, une méthode standardisée de communication de ces informations entre partenaires de la chaîne d'approvisionnement est cruciale. Elle diminue les coûts et accélère le flux d'informations par son adaptation aux échanges de données fortement automatisés. Parce qu'ONIX est prévu et optimisé pour la communication entre ordinateurs, il n'est pas particulièrement facile à lire et à interpréter pour les humains. Et parce qu'ils sont centrés sur la communication, les fichiers ONIX sont habituellement appelés des 'messages ONIX'. Ces messages transitent entre des bases de données, pas entre des personnes »[46] (ibid.). S'il est précisé qu'il s'agit ici d'ONIX Livres, c'est parce qu'ONIX est décliné en une famille de standards qui comprend également ONIX Serials et ONIX for Publication Licenses, comme il l'est mentionné plus bas.

« Comment ONIX Livres est-il apparu ?

ONIX Livres a été à l'origine développé conjointement par le groupe de travail Enjeux du Numérique [*Digital Issues*] de l'Association des Éditeurs Américains (AAP) et EDItEUR, en réponse à l'importance et à la valeur croissantes prises pour les éditeurs et les vendeurs de livres en ligne de métadonnées de très bonne qualité. ONIX signifie *ONline Information eXchange* [échange d'informations en ligne], et la première version

[46] Nous traduisons.

a été publiée en janvier 2000. Cette première version combinait des idées sur les métadonnées de travaux antérieurs comme la spécification 'BIC Basic' de *Book Industry Communication*, le projet financé par l'Union Européenne <indecs> et le dictionnaire de données EPICS d'EDItEUR. Elle était également influencée par la spécification XML publiée en 1998 par le *World Wild Web Consortium*. ONIX Livres visait à réduire la difficulté de gérer, distribuer et mettre à jour de grands volumes de métadonnées riches et dynamiques. Aujourd'hui, ONIX Livres n'est qu'un membre de standards internationaux basés sur XML prévus pour aider la communication d'ordinateur à ordinateur entre parties impliquées dans la création, la distribution, l'octroi de licences ou autre manière de rendre disponible la propriété intellectuelle publiée, physiquement ou numériquement. C'est de loin le plus largement adopté et mis en œuvre des membres de la famille, mais il existe d'autres spécifications pour le commerce des périodiques (journaux académiques), pour l'octroi de licences de bibliothèques, et autres usages spécialisés » (ibid.). Si le nom de l'organisme qui maintient le standard n'y suffisait pas, nous voyons que ce dernier est à l'attention des éditeurs prioritairement, qui sont au début de la chaîne dans la transmission des métadonnées du livre. Dans le contexte de notre alternance, l'enjeu a toutefois l'adoption de ce format par les distributeurs, assez peu d'éditeurs fournissant déjà leurs métadonnées aux distributeurs dans ce format.

« Qui est actuellement responsable du standard ONIX Livres ?
Bien qu'ONIX Livres (à partir de maintenant seulement 'ONIX') ait été au départ développé conjointement par l'AAP et EDItEUR en collaboration avec *Book Industry Communication* (BIC) au Royaume-Uni et le *Book Industry Study Group* (BISG) aux États-Unis, les développements ultérieurs ont été de la responsabilité d'EDItEUR. ONIX est à présent solidement établi dans le monde comme le standard du commerce du livre en matière de communication de 'métadonnées de produit riches' – le type de métadonnées nécessaire à l'appui de la vente de livres dans la chaîne d'approvisionnement, notamment pour la vente en ligne. Les développements en cours sont gérés par EDItEUR, supervisés par un Comité de Pilotage International comprenant des représentants de groupes d'utilisateurs de plus de quinze pays, dont l'Australie, la Belgique, la Chine, le Canada, la Finlande, la France, l'Allemagne, l'Italie, le Japon, les Pays-Bas, la Norvège, la Russie, l'Espagne, la Suède et la République de Corée, ainsi que des représentants américains et britanniques du BISG et de la BIC. Le Comité de Pilotage se réunit deux fois par an, aux Foires de Londres et de Francfort, et il a la responsabilité d'assurer que le standard se développe en regard des besoins des utilisateurs d'ONIX » (ibid.). C'est ici la capacité d'enrichir les métadonnées qui est mise en avant, comme c'est le cas dans la présentation générale du format ci-après.

« Le Message d'Information Produit ONIX pour les Livres est *le* standard international pour représenter et communiquer l'information produit de l'industrie du livre sous une forme électronique.

ONIX est un standard basé sur le XML permettant des métadonnées du livre riches, fournissant aux éditeurs, revendeurs et à leurs partenaires dans la chaîne d'approvisionnement une moyen cohérent de communication d'informations riches sur leurs produits. Il est expressément destiné à être utilisé mondialement, et n'est limité à aucun langage ou aux caractéristiques d'aucun commerce du livre national. Il est largement utilisé à travers la chaîne d'approvisionnement du livre et du livre électronique en Amérique du Nord, en Europe et en Australasie, et il est de plus en plus adopté dans la zone Asie Pacifique.
En tant que standard basé sur le XML, chaque nouvelle version d'ONIX for Books consiste en une définition de type de document (DTD) XML ou sinon en un schéma XSD, conjointement avec une documentation associée qui décrit le contenu des données d'un message ou d'un fichier de données dans le standard ONIX. EDItEUR fournit ces spécifications, plusieurs outils XML, ainsi que des conseils sur la manière de mettre en œuvre ONIX, et l'utilisation de tous ces documents est gratuite aux termes d'une licence hautement permissive. Aucun enregistrement, droit d'enregistrement ou adhésion n'est demandée pour mettre en œuvre ONIX.
ONIX n'est pas en soi une base de données, ni même un modèle de base de données — c'est une façon de se communiquer les données entre bases de données — mais plusieurs membres d'EDItEUR et d'autres organisations fournissent des logiciels standards ou des applications web pour la gestion de produits qui adoptent le standard de communication ONIX » (22, EDItEUR). Les mentions de la permissivité de la licence et de l'absence de droit d'enregistrement ou d'adhésion doivent être ici comprises dans la logique d'un standard, qui abaisse les « coûts » (au sens large) d'adoption pour renforcer son statut de standard, tandis que la promesse est celle de métadonnées plus riches. C'est ce point que nous allons aborder à présent.

2.2.2. L'intérêt du standard ONIX Livres

Afin de mettre en évidence les possibilités d'enrichissement des métadonnées qu'offre le format ONIX, il nous faut d'abord présenter rapidement les formats qui l'ont précédé. Nous pouvons à cette fin nous appuyer sur un document établi par la CLIL et Dilicom, la « Fiche produit du livre » (24, CLIL), dans sa version 3.0 de novembre 2013, qui se présente comme un « Mode d'emploi pour la rédaction et la lecture des fiches informatisées des ouvrages » et signale à sa première page que « ce document, établi par la Clil et Dilicom, précise la définition des termes utilisés dans la Fiche Produit du Livre (en langage Gencod V4). Son utilisation doit permettre d'améliorer la communication entre les différents acteurs du circuit du Livre » (ibid.). Le Gencod V4 est l'un des trois formats « historiques » d'échange de données, avec le Prodis 4 et le « format plat ». La principale caractéristique de ces trois formats que nous devons retenir ici est qu'ils sont positionnels, c'est-à-dire que c'est la position dans une ligne d'une certaine valeur qui permet de savoir de quelle propriété de la ressource elle est

la valeur. La conséquence de cette structuration est bien entendu qu'il est difficile de modifier la position ou la taille d'un champ. Mais quels sont ces champs ? Ils sont au nombre de trente-quatre et mêlent métadonnées commerciales et métadonnées bibliographiques.

Métadonnée bibliographique	Taille / valeur	Commentaire
Auteur	20 car.	Des règles définissent la manière d'exprimer un auteur personne physique, une collectivité auteur, un auteur anonyme, plusieurs auteurs, un collectif, un auteur inconnu
Livre scolaire	1, 0 ou rien	
Collection		
Collection sérielle	4 car. Alphabétiques et 6 car. numériques	Signale une collection dont chaque ouvrage a un numéro. La liste des collections est déposée auprès de Dilicom
Éditeur	15 car.	Marque éditoriale figurant sur la première de couverture
Libellé étendu	100 car.	Titre exact figurant sur la première de couverture
Présentation éditeur		Caractéristique physique du produit (livre relié, broché, jeu, CD, etc.)
Produit lié		Indique un lien de remplacement entre deux produits
Publics	01 ou 02	Tout public / réservé aux enseignants
Thèmes	4 car. num.	Liste de thèmes établie par la Clil

Tableau 4 : Les métadonnées bibliographiques dans les formats positionnels des fiches-produit (CLIL, 2013)

Métadonnée commerciale	Taille / valeur	Commentaire
Code EAN 13	13 car.	*European Article Numbering* : c'est l'ISBN précédé d'un préfixe (978 ou 979 généralement)
Code ISBN	10 car.	*International Standard Book Number*
Peut être commandé par le revendeur	1 ou 0	Commandabilité à l'unité
Composants		Cas d'un article composé

Date d'application		Prise d'effet de la mise à jour
Date de fin de commercialisation		
Date de parution		Date de la première publication
Dimensions	3 x 4 car. num.	Épaisseur, hauteur, largeur
Disponibilité	2 car. num.	Faculté d'obtention
Impression à la demande	I (i maj.) ou rien	Impression à la commande uniquement
ISBN de l'éditeur		
Libellé caisse	20 car.	Titre abrégé
Libellé standard	30 car.	Titre courant
Motif de suppression		
Nombre de références		Nombre de composants d'un article composé
Poids	7 car.	
Présentation magasin		
Prix		D'abord type de prix puis prix selon le type défini
Produit (type de)		Conditionnement particulier du produit (unité, lot, etc.)
Référence fournisseur		Code produit attribué par le fournisseur
Retour	1 car num.	Indique un accord commercial de retour chez le distributeur
Symbolisation	1 car. num.	Représentation sous forme de code à barres
TVA		
Type de lot	0 ou 1	Insécable / sécable

Tableau 5 : Les métadonnées commerciales dans les formats positionnels des fiches-produit (CLIL, 2013)

Nous voyons que la Fiche Produit du livre fait la part belle aux métadonnées commerciales, qui représentent près de 70% de l'ensemble. Cette proportion reflète à la fois l'histoire et le rôle de Dilicom. Comme nous l'avons vu, Dilicom est issue d'une filiale d'Électre, qui est une base de données bibliographiques. Il y a donc eu un partage des rôles, Électre restant sur son cœur de métier avec les données bibliographiques et Dilicom s'attachant préférentiellement aux métadonnées commerciales, ne conservant que des données bibliographiques assez basiques (nous sommes d'ailleurs très proches du standard BIC Basic évoqué par l'étude Nielsen), liés aux impératifs minimaux de la vente en librairie – la mention des thèmes étant dans la logique de la définition de « rayons ». D'autre part, l'auteur, l'éditeur et le libraire, nous l'avons vu en évoquant le paradigme de l'économie de l'attention, ont un

fort intérêt à des métadonnées riches. En revanche le distributeur peut constituer un goulet d'étranglement de ce point de vue, dans la mesure où son intérêt le porte essentiellement à privilégier les métadonnées utiles aux commandes et à la logistique des produits, mais ne vise pas directement à capter l'attention des consommateurs.

Pour comprendre la transformation que représente le passage au format ONIX, nous devons comparer les quelques lignes (d'une à trois selon les formats) d'une Fiche Produit telle que nous venons de la décrire à une notice au format ONIX. Nous présentons une notice ONIX « type » dans l'annexe 1. Pour donner une idée générale de la manière dont est structurée une notice au format ONIX, nous reprenons une planche de la présentation de la Directrice Générale de Dilicom, faite à la BnF et intitulée « ONIX, une norme pour communiquer entre familles professionnelles ? » (21, Backert).

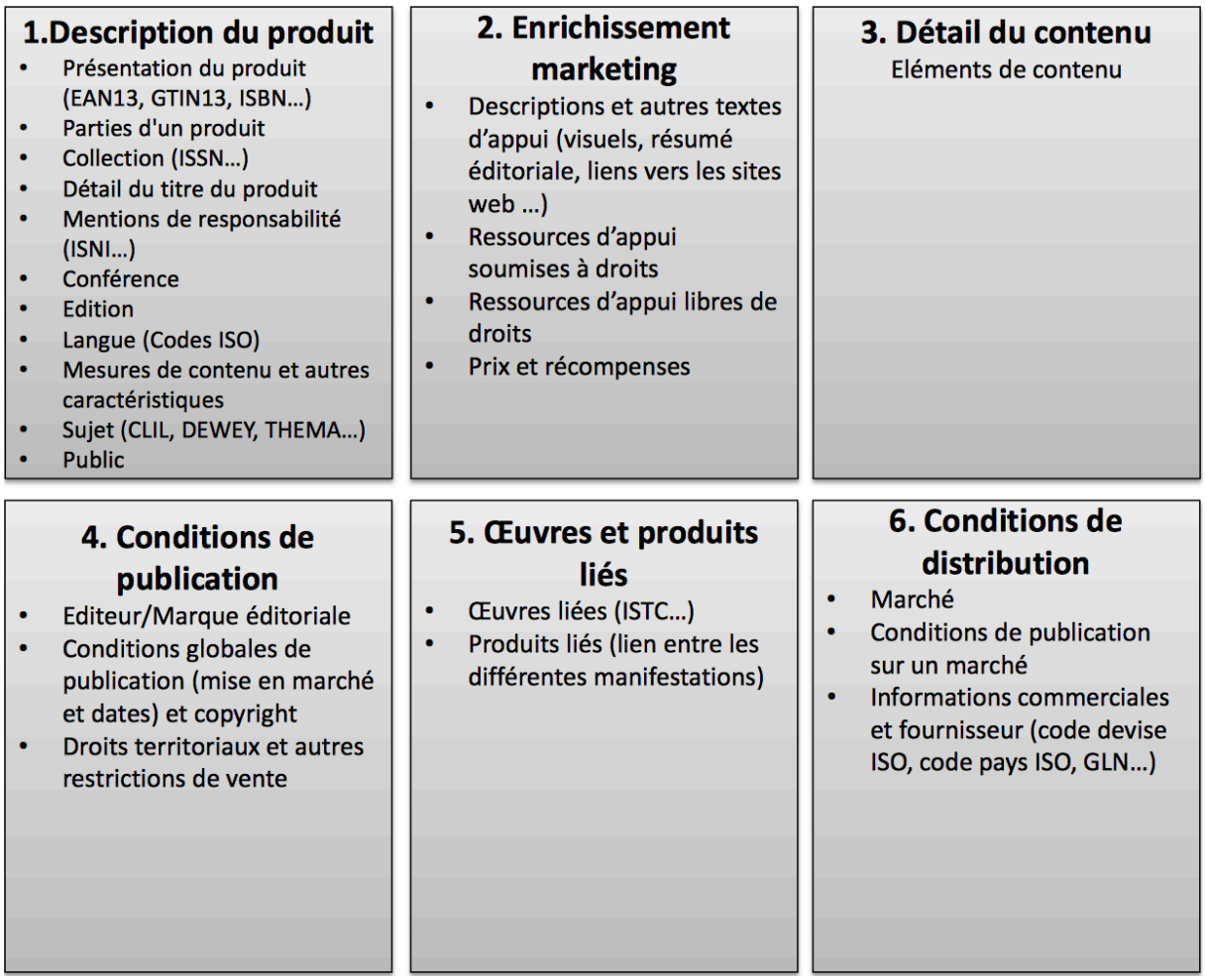

Figure 8 : Les six blocs d'ONIX Livres (V. Backert., 2014)

Une présentation par EDItEUR plus détaillée et traduite en français est proposée par le Cercle de la Librairie sur son site (25, EDItEUR). Ce qui n'apparaît pas dans sa pleine extension est le nombre fortement accru de métadonnées possibles : plus de deux cents, certaines exigées et d'autres optionnelles. Nous pouvons comparer avec la Fiche Produit et constater que la plupart des métadonnées

de la Fiche Produit se retrouvent dans le bloc 1. De plus, ONIX étant basé sur XML, il est possible pour un grand nombre de métadonnées de spécifier plusieurs valeurs. Prenons l'exemple des noms de contributeurs : un grand nombre peut être indiqué, avec pour chacun leur rôle ; ou encore l'exemple des marchés et des prix sur chacun de ces marchés, qui permet avec la même notice de sortir des frontières françaises.

Venant avant les blocs décrits ci-dessus, nous devons signaler l'existence d'un en-tête porteur de métadonnées « techniques » comme l'émetteur de la notice, ou un identifiant de celle-ci. Une fois décrite la structure et évoquées les propriétés, venons-en aux valeurs des propriétés, qui sont spécifiées dans des vocabulaires contrôlés basés pour certains sur des normes ISO et appelés *codelists*. EDItEUR les présente ainsi : « Les listes de codes – parfois appelées vocabulaires contrôlés – sont une part vitale d'ONIX. Elles font partie de la sémantique partagée d'un message ONIX. Les listes de codes contiennent une liste de valeurs de codes indépendantes des langues (souvent numériques) qui sont utilisées dans les messages ONIX, ainsi qu'un libellé court pour chaque code et parfois un commentaire plus long pour définir le sens du code. En principe, les libellés et les commentaires peuvent être dans n'importe quelle langue : bien que le sens des presque tous les codes soit originellement défini en anglais, la traduction des libellés et/ou des commentaires dans une autre langue ne modifie pas le sens du code lui-même. Le code BB dans la liste 150 signifie 'livre relié', mais signifie également '精装' (jīngzhuāng shū) et 'Gebundene ausgabe'. Ainsi, le sens d'un message ONIX utilisant une valeur de code particulière est compréhensible et non ambigu, quels que soient les langages utilisés par l'émetteur et le récepteur des données » (26, EDItEUR).

2.2.3. Les problèmes que pose l'adoption de l'ONIX

Nous évoquions la différence entre les quelques lignes d'une fiche-produit traditionnelle et une notice au format ONIX, qui compte généralement environ deux cents lignes. La qualité descriptive accrue d'ONIX, et son caractère autodescriptif (puisque les balises précisent les propriétés qu'elles contiennent), modifient l'ordre de grandeur du poids en kilooctet d'une notice. Dès lors qu'un fichier contient, comme il est courant, plusieurs dizaines de milliers de notices, les processus, voire les infrastructures qui étaient traditionnellement en place peuvent devoir être modifiées. D'autre part, de ce point de vue - aussi bien que du point de vue de la constitution ou de la réception et de la lecture d'un catalogue au format ONIX - de nouvelles compétences doivent éventuellement être développées au sein des entreprises.

Si nous nous adoptons un point de vue plus technique sur le standard lui-même, nous pouvons reprendre les difficultés que note Véronique Backert dans l'utilisation d'ONIX :

« • Un standard qui jongle entre une description globale du marché du livre mais qui essaye aussi d'intégrer les spécificités de chaque pays : risques de construire une

arborescence incohérente liée à une juxtaposition de pratiques et de demandes de différents pays ;
• Des descriptions ou des interprétations différentes pour un même objet mais syntaxiquement correctes (description des gratuits, description des prix par pays...) ;
• Un complexité liée à la sophistication du modèle : problème de l'homogénéité des données (plusieurs manières différentes de renseigner un contributeur) ;
• Des contraintes fortes liées à la fréquence des mises à jour des listes de codes et des versions d'ONIX (semestrielles) ;
• L'impossibilité parfois d'établir des correspondances entre les différentes versions ;
• Difficulté de mettre en place des contrôles ;
• Le respect du standard ne garantit pas de la valeur qualitative des données transmises (syntaxe <> sémantique) ! » (21, Backert).

Enfin, un point important à envisager pour des acteurs commerciaux est la question du coût de la mise en place d'un nouveau format. Les modifications multiples, en termes d'infrastructure ou de ressources humaines, que nécessite la mise en place de ce nouveau format peuvent amener les distributeurs à s'interroger sur l'existence ou pas d'un intérêt financier dans cette démarche.

Si nous retournons à l'*Overview* que propose EDItEUR du standard ONIX sur son site, le paragraphe *Business benefits* présente comme nous pouvons nous y attendre une image très positive : « Pour les éditeurs, l'expérience a montré qu'ONIX Livres procure deux avantages commerciaux. En tant que format de communication, il permet de livrer une information produit riche dans la chaîne d'approvisionnement sous une forme standardisée, aux grossistes et aux distributeurs, aux plus grands revendeurs, aux agrégateurs de données, et aux sociétés affiliées. Cela réduit fortement les frais de support, puisque les éditeurs n'ont plus besoin de fournir les données dans autant de formats uniques. Dans bien des cas, un seul flux de données peut être utilisé par tous les partenaires de la chaîne d'approvisionnement d'un éditeur. Et en fournissant un modèle pour le contenu et la structure d'une fiche-produit, ONIX a contribué à stimuler l'introduction de meilleurs systèmes d'information internes, capables de réunir toutes les 'métadonnées' nécessaires pour la description et la promotion de titres neufs ou de fonds. Dans certains pays, ce modèle est utilisé dans des systèmes d'accréditation de qualité de la donnée. Les mêmes données ONIX de base peuvent aussi être utilisées pour produire des fiches de renseignement élaborées, des catalogues et autre matériel promotionnel, de nourrir les sites web des éditeurs, et de satisfaire les besoins de la chaîne d'approvisionnement étendue » (23, EDItEUR). Cette présentation appelle plusieurs remarques. D'abord, le cadre qui est présenté ne correspond pas tout à fait à celui que nous avons eu l'occasion d'observer lors de notre alternance, puisque s'il peut arriver que des éditeurs fournissent aux distributeurs des notices au format ONIX, ce n'est pas un cas fréquent et l'effort de produire des notices au format ONIX a plutôt été porté sur les distributeurs. Nous pouvons y voir un effet de la structuration spécifiquement française de la chaîne du livre. Ensuite, le discours est adressé à des éditeurs, en leur montrant qu'ils ont intérêt

à transmettre au format ONIX les métadonnées riches qu'ils détiennent sur les ouvrages, ce qui n'est pas aussi évident qu'il y paraît dans cette présentation. Enfin, remarquons que dans le cas des distributeurs avec lesquels nous avons collaboré lors du passage à l'ONIX, il est indéniable qu'un effet secondaire - mais non négligeable - du processus a été de les obliger à « nettoyer » leurs bases de données d'articles qui y figuraient alors qu'ils n'auraient pas ou plu dû.

Mais pour en revenir au deuxième point, qui consiste à affirmer que le format ONIX constituerait un avantage commercial pour les éditeurs, ce qui d'autre part bénéficierait aux autres acteurs de la chaîne d'approvisionnement (notamment les gros revendeurs), nous pouvons émettre un doute sur cette manière de présenter les intérêts en jeu. Ce doute est illustré par l'article d'Endre Beky, directeur régional des ventes chez Elsevier, dans son article de 2004 *ONIX: Is there a return on investment for all publishers?* (27, Beky). L'auteur s'y demande : « Que sera le futur d'ONIX ? Certains soutiennent qu'une acceptation plus large de la part de l'industrie mènera à plus d'efficacité dans le transfert des données des livres, ce qui bénéficiera finalement à la vente des livres. ONIX doit encore surmonter plusieurs défis : son coût de mise en œuvre est élevé, les grossistes et les agrégateurs utilisent encore différents tableurs et différentes sources d'information. En outre, MS Excel est une solution alternative amplement disponible et acceptée par l'industrie. Les petits éditeurs ont besoin de plus de formation et d'une démonstration évidente du retour sur investissement d'ONIX. Une diffusion de l'aide des organisations professionnelles est fortement recommandée afin d'encourager le processus d'adoption parmi les petits acteurs. C'est le rôle de l'AAP [*American Association of Publishers*] et du BISG [*Book Industry Study Group*] de faire que leurs effectifs pensent à ce qui est bon pour tous les membres, à travers toute l'industrie. En tant qu'organisation, le BISG peut recommander, enseigner et promouvoir le standard. Il ne peut pas, en revanche, faire pression ou exiger que les organisations l'utilisent. Les éditeurs, les revendeurs et les grossistes prennent leurs décisions pour eux-mêmes et apparemment la majorité ne voit pas encore les récompenses possibles. Les acteurs de ce marché doivent déterminer le rapport coût bénéfice et aussi longtemps que les distributeurs et les revendeurs sont d'accord pour recevoir d'autres formats, ONIX ne s'imposera pas sur le marché. Après tout, les éditeurs devraient se concurrencer les uns les autres avec de bons livres, pas avec des codes informatiques multiples »[47] (ibid.). Même si cet article est à présent ancien, force est de constater que le paysage français actuel présente de fortes similitudes avec ce qui est décrit ici, et que le clivage entre les plus gros distributeurs ou revendeurs, intéressés par la mise en place du format ONIX, et les plus petits acteurs pour lesquels c'est un processus complexe, coûteux et dont ils ne voient pas forcément l'intérêt. Une forte différence entre les paysages étasunien et français est précisément l'existence d'une structure comme Dilicom, qui peut « recommander, enseigner et promouvoir le standard », mais aussi en tant que

[47] Nous traduisons.

plateforme interprofessionnelle mettre en place le standard si l'interprofession dans sa majorité le décide.

Nous voudrions toutefois explorer plus avant les arguments qu'avance Endre Beky contre ONIX. Il explique que « si les métadonnées peuvent être créées, intégrées dans un tableur Microsoft (MS) Excel, et envoyé aux distributeurs, alors quel est le réel avantage, la raison impérieuse des éditeurs de taille petite à moyenne pour investir dans cette technologie ? En théorie, l'avantage compétitif de l'ONIX serait que tous les éditeurs pourraient envoyer la même information commerciale sur leurs titres à tous les partenaires commerciaux de la chaîne du livre en même temps, mettre à jour les informations de vente en temps réel avec la même facilité qu'on appuie sur un bouton. Est-ce que cela aura à long terme un effet sur le modèle actuel de vente de livres ? La réponse est peut-être, ou au mieux un positif oui.

Que nous l'acceptions ou non, nous vivons dans un monde où le marketing, le design, le packaging sont aussi importants que le produit lui-même. La même tendance existe pour l'industrie du livre, avec quelques exceptions comme l'édition littéraire. Plusieurs agrégateurs de contenu en ligne, y compris les gens de NetLibrary[48], soutiennent que "le contenu est mort, l'accès est ce qui importe". Aujourd'hui, un éditeur de livres grand public – même s'il publie les meilleurs livres - ne peut pas se permettre de *ne pas* commercialiser ses titres sur Amazon, Baket & Taylor, Barnes & Noble, Ingram, ou tout autre réseau de vente important. Se battre pour l'attention limitée des consommateurs est plus crucial que jamais.

Quatre ans après ses débuts, ONIX est encore plus ou moins le privilège des plus grands éditeurs et de quelques-uns de taille moyenne, qui peuvent se permettre d'avoir un département informatique et le soutien des départements marketing et éditorial pour incorporer ONIX dans leur façon de travailler. Selon l'estimation d'un fournisseur de solutions ONIX, 50 à 60 éditeurs étasuniens jusqu'à présent envoient des fichiers ONIX à leurs partenaires commerciaux. Selon moi, le reste de l'industrie de l'édition étasunienne ne l'adoptera pas tant que son coût de mise en place n'aura pas baissé de manière drastique et que les éditeurs ne verront pas d'avantage financier » (ibid.). Incidemment, remarquons qu'Endre Beky fait ici référence à l'économie de l'attention, pour justifier que les ouvrages doivent se trouver sur les principaux réseaux de vente. Nous avons quant à nous invoqué ce cadre de pensée pour étudier la question de l'enrichissement des métadonnées. Nous voyons que les deux se rejoignent autour de la question d'ONIX comme standard : comme le remarque l'auteur, l'impulsion pour l'adoption de l'ONIX n'est pas venue des éditeurs mais des plus importants réseaux de vente. Ceux-ci souhaitent en effet disposer d'un standard pour ne pas avoir à traiter autant de formats de métadonnées que d'éditeurs ou de distributeurs. Leur taille les rendant incontournables, les éditeurs les plus importants doivent adopter ce standard de métadonnées pour gagner de la visibilité

[48] NetLibrary, fournisseur d'e-books, a été racheté en 2011 par la compagnie étasunienne EBSCO Information Services, qui fait notamment commerce de ressources bibliothécaires.

en étant présents sur ces réseaux, au cas où ils n'auraient pas souhaité adopter ce format pour gagner plus de visibilité par la richesse de métadonnées que ce même format permet. Pour en revenir au contexte français, la position d'intermédiaire de Dilicom dans l'échange de données informatisées entre distributeurs et revendeurs lui fait « absorber » une partie du coût du passage à l'ONIX pour les distributeurs, non pas en termes financiers mais en termes de formation à l'usage de ce nouveau format, à travers l'accompagnement des distributeurs et des libraires, comme nous allons le montrer immédiatement.

2.3. Mettre en place un nouveau format

Comme nous l'avons dit, nos deux années d'alternance au sein de Dilicom nous ont permis de participer à une période particulièrement importante de la vie de l'entreprise et plus largement du secteur commercial du livre, qui a été celle de l'adoption d'un nouveau format standardisé pour les métadonnées du livre[49]. La transition était engagée avant notre arrivée et se poursuit, une telle transformation est une œuvre de longue haleine, mais nous avons eu la chance d'accompagner des distributeurs qui effectuaient le passage d'un format à l'autre et de devoir également accompagner, et parfois trouver des solutions, pour les revendeurs qui, même s'ils aspiraient pour certains d'entre eux à cette transformation, n'en devaient pas moins adapter leurs systèmes d'information. C'est à cette nouvelle donne pour les métadonnées que nous allons maintenant nous attacher, dans ce qu'elle implique d'abord pour la collecte des notices au format ONIX, puis ensuite pour la diffusion des notices au format ONIX par Dilicom.

2.3.1. Collecter les notices au format ONIX

Afin de bien comprendre les enjeux et le contexte du passage au standard ONIX, rappelons que ce standard à prétention internationale est d'abord conçu à destination des éditeurs. Dans le contexte français, ce sont les distributeurs qui ont été les acteurs de la mise en place de ce format. Même si certains éditeurs, loin d'être majoritaires, leur transmettaient déjà leurs notices dans ce format, eux-mêmes ne l'utilisaient pas. Le standard ONIX représente deux défis pour les distributeurs : d'une part, il fait la part belle aux métadonnées bibliographiques, d'autre part, il inscrit les métadonnées commerciales dans un contexte international. Si les distributeurs sont très avertis concernant les métadonnées bibliographiques, elles ne ressortissent pas de leur cœur de métier et ne constituent donc pas leur principale préoccupation. Entre les éditeurs et les libraires, qui peuvent tous deux aspirer à disposer de métadonnées bibliographiques riches, les distributeurs peuvent constituer sinon un goulet d'étranglement, du moins l'occasion d'une perte de richesse dans l'expression des métadonnées. D'autre part, des plus grosses structures de distribution disposant d'un service informatique rompu aux technologies de l'information et de la communication les plus récentes, jusqu'aux plus petits éditeurs auto-distribués pour lesquels l'utilisation d'un tableur reste difficile, des distributeurs dont les produits sont essentiellement les livres qu'on trouve classiquement dans une librairie aux distributeurs qui proposent une gamme de produits extrêmement diversifiée, les

[49] Nous parlons ici du livre imprimé. Quelques distributeurs formaient une « avant-garde » restreinte. Dans le cas du livre numérique, qui compte très peu d'acteurs en termes de distribution, mais beaucoup plus versés dans les technologies numériques, le passage à l'ONIX avait eu lieu quelques années auparavant.

situations rencontrées sont à chaque fois singulières et nécessitent une prise en compte des problématiques propres à chacun.

L'accompagnement des distributeurs a donc débuté avec un chacun d'entre eux par des rendez-vous en tête à tête permettant de fixer ce qui était attendu et les différentes difficultés qui pouvaient être rencontrées. Un document essentiel de ce processus a été le *Guide pratique ONIX* édité par la Commission FEL (Fichier Exhaustif du Livre) de la CLIL (Commission de Liaison Interprofessionnelle du Livre), dont le sous-titre est « Règles pour la rédaction et l'intégration des métadonnées au format ONIX 3.0 pour le commerce du livre » (28, CLIL). La commission FEL rassemble à la fois des éditeurs, des distributeurs, des libraires, mais également des « membres associés » d'Alire, de la BnF, de Dilicom, d'Électre, du Syndicat de la Librairie Française et du Syndicat National de l'Édition. L'objectif de ce guide pratique, intitulé dans ses premières version « guide des bonnes pratiques », est selon le « mot du président », « de normaliser les données nécessaires à la description et à la commercialisation des livres imprimés, numériques et autres produits culturels référencés dans le FEL » (ibid., p. 10). Rappelant le « souci toujours plus grand de visibilité et d'accessibilité [des] ouvrages », il rappelle que pour améliorer leur découvrabilité, il paraît « primordial que ces données soient strictement normalisées, claires et appliquées par l'ensemble de la profession ». Le président précise ensuite : « pour atteindre cet objectif, notre commission […] s'est appuyée sur le partage des expériences de chacun en coordination avec les différentes commissions de l'interprofession. Ces travaux, menés à partir de la norme ONIX 3.0 sont centrés sur les données "vitales" et destinés à tous les intervenants de la "chaîne" du livre. Nous avons choisi de vous présenter nos travaux sous forme de fiches qui définissent les règles de codification à appliquer pour chacune des métadonnées, et qui fournissent des exemples illustrant l'utilisation de ces règles » (ibid., p. 10). Vient ensuite un tableau présentant ces « données vitales », c'est-à-dire une « liste de données minimales que l'éditeur ou le distributeur doit s'engager à fournir dans le cadre du Fichier Exhaustif du Livre » (ibid., p. 12). Elles ont au nombre de 79, dont 28, les plus importantes, sont surlignées en gris (voir ANNEXE 2). Le reste du document est pour l'essentiel consacré à expliquer ou préconiser des manières d'exprimer certaines métadonnées. Nous pouvons donc noter d'abord que ce guide a bien une motivation « pédagogique », en ceci qu'il donne des explications et des exemples quand la documentation produite par EDItEUR, essentiellement en anglais, est parfois aride. Remarquons ensuite qu'en définissant ce socle de données, en prescrivant dans certains cas l'usage de certaines balises plutôt que d'autres, de certaines valeurs plutôt que d'autres, il construit une sorte de « standard dans le standard ». Il est bien entendu loisible à un éditeur ou un distributeur d'utiliser en plus des balises préconisées pour exprimer par exemple un nom les autres balises qui le permettent ; ce que nous voulons pointer ici est le besoin ressenti d'aménager au sein d'un standard aussi étendu et ouvert que l'ONIX, pour des raisons évidemment pratiques, une manière commune de l'utiliser qui redouble la standardisation.

Selon les solutions choisies pour mettre en place le format ONIX dans leur structure, recours à leur prestataire informatique, utilisation d'un logiciel spécialisé, solution *ad hoc* développée en interne, l'accompagnement des distributeurs a évidemment pris des allures très différentes. De ce point de vue, il est apparu notable que le temps d'accompagnement entre les premiers rendez-vous et le déploiement en production de la réception, pour un distributeur particulier, ne dépend pas uniquement de la taille ou de la complexité de son catalogue, mais que la capacité technologique du distributeur est à prendre en compte. Comme ces deux dimensions sont généralement en relation inversement proportionnelle, accompagner un « petit » ou un « grand » distributeur n'a pas demandé d'une façon qui soit remarquable plus ou moins d'efforts.

Une place particulière doit ici être faite à une métadonnée particulière, non textuelle : la première de couverture. L'étude Nielsen nous a permis de constater la corrélation entre la présence de cette métadonnée spécifique et la vente d'un ouvrage. Dans le cas du livre numérique, cette métadonnée est présente dans une balise spécifique (*SupportingResource*) sous la forme d'un lien http vers une ressource fournie par le distributeur. Dans le cas du livre matériel, antérieurement à l'adoption du format ONIX, un circuit parallèle était mis en place pour la collecte de cette métadonnée particulière, qui pouvait être fournie par l'éditeur, le distributeur ou le diffuseur d'un ouvrage selon les cas sous la forme d'un fichier image numérique. La généralisation de l'emploi du format ONIX pour le livre matériel également, permet à présent d'intégrer la collecte de cette métadonnée dans la collecte de la notice ONIX.

2.3.2. Diffuser les notices au format ONIX

Notons de façon liminaire que pour diffuser les notices au format ONIX, il faut pouvoir les stocker et les traiter. Le passage au format ONIX a donc supposé pour Dilicom la mise en place de nouveaux outils technologiques. Alors que les notices dans les formats positionnels pouvaient être intégrées dans des bases de données relationnels, en recourant donc à des systèmes de gestion de bases de données (SGBD) relationnels, les notices au format ONIX, basé sur le XML, en raison du caractère semi-structuré de ce langage de balisage, ne peuvent pas sans difficulté être traitées avec ce type de SGBD. Cela aurait supposé un *mapping* complexe, des procédures lourdes, et un risque de perte d'information, quand un SGBD NoSQL semble mieux adapté ; c'est d'ailleurs la solution qui a été mise en œuvre. Le standard ONIX a également poussé les membres de l'équipe du pôle Projets et Gestion de base de données de Dilicom à développer de nouvelles compétences dans des langages de requête permettant d'extraire de l'information de documents ou de collections de documents en XML, comme XPath ou XQuery.

Afin de pouvoir proposer la diffusion au format ONIX, Dilicom a procédé auprès des revendeurs à un sondage sur les habitudes de traitement des mises à jour des

notices qui leur sont quotidiennement transmises. Ce sondage portait sur les horaires auxquels ils effectuaient ces traitements, sur le nombre maximal de mises à jour et sur le nombre de diffusions quotidiennes de mises à jour qu'ils étaient en mesure de traiter. C'était également une opportunité de savoir quels revendeurs, d'une manière plus générale, étaient susceptibles de recevoir des mises à jour au format ONIX. Nous avons insisté jusqu'à présent sur la diversité des distributeurs et des éditeurs, mais la diversité des libraires et autres revendeurs n'est pas moindre. Outre la librairie archétypique telle que nous pouvons nous la représenter, citons le point-presse qui a une offre de livres, les chaînes de magasins de grande distribution de produits culturels, les chaînes de magasins de grande distribution généraliste (qui sont depuis longtemps des vendeurs importants de livres) et les plateformes de commerce électronique. En outre, les solutions technologiques adoptées, si elles concernent ici la réception et le traitement des notices, sont — comme pour les distributeurs — différenciées. Beaucoup de librairies recourent à des prestataires informatiques, les grands revendeurs disposent de solutions en interne. Toutefois, parmi les revendeurs, tous ne sont pas à même de traiter des notices en ONIX et ne souhaitent donc pas abandonner les formats positionnels. Cela suppose pour Dilicom de diffuser de façon différenciée selon les revendeurs les notices au format ONIX ou dans les formats positionnels, ce qui ne va pas sans poser quelques difficultés. Pour beaucoup de métadonnées, il s'agit de mettre en place un *mapping* qui réduit la richesse des valeurs exprimées au format ONIX mais ne pose pas de difficulté majeure. Pour certaines métadonnées, la tâche est plus ardue. Nous pouvons prendre l'exemple des bandes dessinées faisant partie d'une série numérotée dont chaque opus à son titre propre. Pour les formats positionnels, qui n'incluent pas les métadonnées de collection ou de tomaison, les éditeurs et les distributeurs avaient coutume de proposer un titre rassemblant le nom de la série, le numéro dans la série et le titre de l'opus. Dans la mesure où le format ONIX permet d'exprimer ces différentes valeurs dans différentes balises, continuer de faire ainsi aboutit, pour les revendeurs qui utilisent l'ONIX, à un redoublement, dans le titre, de l'information qu'ils ont par ailleurs. A l'inverse, demander aux éditeurs et aux distributeurs de cesser de faire ainsi pourrait faire, en mappant dans le format positionnel le seul titre, que les revendeurs utilisant les formats positionnels n'aient plus que le titre de l'opus, sans disposer du nom de la série ou du numéro dans la série. Afin de pouvoir fournir ces deux types de revendeur, Dilicom doit donc mettre en place une concaténation des valeurs de l'ONIX pour le nom de la série, le numéro de la série et le titre de l'opus, afin de produire un titre qui puisse alimenter le format positionnel. Comme nous l'avons déjà souligné, cette capacité de Dilicom à recevoir en « entrée » et à diffuser en « sortie » plusieurs formats est à la fois sa « marque de fabrique », ainsi qu'une part de sa raison d'être et de sa capacité à créer de la valeur ajoutée pour le secteur commercial du livre.

Conclusion

Le monde des bibliothèques parle des livres comme d'objets informationnels, quand le secteur commercial du livre en parle comme de produits. Il nous semble qu'il serait erroné d'en conclure qu'ils auraient des abords complètement différents du livre, notamment en ce qui concerne les métadonnées bibliographiques. En ce sens, le secteur commercial du livre a tout intérêt à se rapprocher encore du monde des bibliothèques, en pleine « transition bibliographique », « des catalogues vers le web de données », pour reprendre le titre de la page d'accueil du site www.transition-bibliographique.fr (émanation de la Bibliothèque nationale de France et de l'agence bibliographique de l'enseignement supérieur). Cela peut amener le secteur commercial du livre à envisager de nouvelles évolutions des standards qu'il utilise.

Si la manière d'aborder le livre ne nous semble pas radicalement différente, ce qui fait toutefois la différence du secteur commercial du livre est la diversité des acteurs qui le composent. De même qu'on parle de chaîne du livre, il faut parler de chaîne des métadonnées du livre. Et selon l'image consacrée, la force d'une chaîne est exactement celle de son maillon le plus faible. Ceci reste vrai quel que soit le standard dans lequel sont exprimées les métadonnées, même si un changement de standard peut amener temporairement la fragilisation d'une chaîne des métadonnées ancrée dans un fonctionnement bien établi. Dès lors, l'enrichissement des métadonnées du livre suppose que chaque acteur de la chaîne du livre soit stimulé dans une démarche d'adoption d'un standard et de qualité des métadonnées, ce que la place particulière de Dilicom dans le secteur commercial du livre français lui permet d'accomplir.

3. Troisième partie

Alors que le standard ONIX est encore d'usage récent dans le secteur commercial du livre, il peut paraître prématuré d'envisager de nouvelles modifications touchant au format des métadonnées du livre. Pourtant, nous avons vu dans la première partie qu'à travers les questionnements relatifs aux métadonnées, un même horizon était envisagé par les différents auteurs que nous avons cités, celui des données liées. D'autre part, une source d'enrichissement des métadonnées paraissait encore trop peu exploitée, celle des métadonnées « sociales » issues du web collaboratif. Afin de nous projeter dans un avenir possible pour le secteur commercial du livre, nous souhaitons à présent envisager la place que les deux, paradigme des données liées et métadonnées sociales, pourraient occuper dans l'enrichissement des métadonnées.

3.1. Le paradigme des données liées

3.1.1. L'idée

Le terme de « données liées » a été introduit par Tim Berners-Lee en 2006. Dans un article de 2009 qu'il co-écrit avec Christian Bizer et Tom Heath, *Données liées – Un peu d'Histoire* , la définition suivante en est proposée : « Le terme Données Liées renvoie à un ensemble de bonnes pratiques pour publier et connecter des données structurées sur le Web »[50] (29, Bizer *et al.*). Même si le Web nous est familier depuis une vingtaine d'années, il nous semble utile, avant d'explorer plus avant la notion de Données Liées, de faire un retour sur l'histoire du Web lui-même, dans laquelle Tim Berners-Lee a un rôle prépondérant, mais dont on fait généralement remonter la vision anticipée à l'article de Vannevar Bush « *As We May Think* ». Nous souhaitons citer le début de cet article, Vannevar Bush situant sa proposition dans des enjeux d'organisation de la connaissance et de constitution de la culture d'une manière qui entre en résonance avec la problématique dans laquelle nous avons inscrit notre étude de l'enrichissement des métadonnées : « Il y a une montagne croissante de recherche. Mais il devient manifeste qu'aujourd'hui nous nous enlisons, avec l'extension de la spécialisation. Le chercheur est frappé de sidération par les trouvailles et les conclusions de milliers d'autres travailleurs – conclusions qu'il ne peut trouver le temps de comprendre, encore moins de mémoriser, à mesure qu'elles apparaissent. Pourtant, la spécialisation devient de plus en plus nécessaire au progrès, et l'effort pour rapprocher les disciplines est en proportion superficiel.
D'un point de vue professionnel, nos méthodes de transmission et d'examen des résultats de la recherche sont vieux de plusieurs générations et sont devenus

[50] Nous traduisons

totalement inadéquats à leur finalité. Si l'on pouvait évaluer le temps cumulé à écrire des travaux universitaires et celui à les lire, le ratio entre ces quantités de temps pourrait bien nous surprendre. Ceux qui tentent consciencieusement de se tenir au fait de la pensée actuelle, même dans des champs restreints, par une lecture continue et attentive, pourraient bien craindre une évaluation quantifiée montrant combien, des efforts du mois précédent, pourrait être produit sur demande. Le concept des lois de la génétique de Mendel a été perdu pour le monde pendant une génération parce que sa publication n'a pas atteint les quelques personnes qui étaient capables de la comprendre et de la prolonger ; et cette sorte de catastrophe est sans doute répétée à l'identique pour nous, les connaissances vraiment significatives se perdant dans la masse de l'inconséquent.

La difficulté semble être non pas tant que nous publions exagérément au regard de l'étendue et de la diversité de nos intérêts actuels, mais plutôt que la publication a été étendue bien au-delà de notre capacité à en faire un réel usage. La somme de l'expérience humaine s'est étendue à un rythme prodigieux, et les moyens que nous employons pour nous frayer un chemin dans le labyrinthe qui en résulte jusqu'à l'item momentanément important sont restés les mêmes qu'à l'époque des gréements carrés »[51] (30, Bush). Pour sortir de cette difficulté, Vannevar Bush imagine une machine fictive permettant d'augmenter la mémoire, le « memex » (*memory extender*), en affichant des livres et en offrant la possibilité de créer des liens entre ces livres ainsi que de garder des annotations à propos de ses lectures. Dans les années 1960, Ted Nelson met en œuvre cette idée avec des fichiers « hypertexte », découpés en morceaux, les morceaux étant reliés d'une manière qui rende possible de naviguer entre eux. Dans les années 1980, Tim Berners-Lee transpose cette idée dans un réseau de machines, un document pouvant pointer vers un document présent sur une autre machine du réseau. Ce rapide survol nous met en position de mieux comprendre le fond sur lequel s'inscrit la proposition d'un Web de données. En effet, le Web peut être vu comme un Web de documents, chaque page Web étant un document lié par des liens hypertextes à d'autres documents, c'est-à-dire d'autres pages. Comme le développent Bizer, Heath et Berners-Lee, « le World Wide Web a radicalement modifié la façon dont nous partageons l'information en abaissant la barrière à la publication et à l'accès aux documents comme parties d'un espace informationnel global. Les liens hypertextes permettent aux usagers de traverser cet espace informationnel en utilisant des navigateurs Web tandis que les moteurs de recherche indexent les documents et analysent la structure des liens entre eux pour en inférer une potentielle pertinence pour les recherches des utilisateurs. Cette fonctionnalité a été permise par la nature générique, ouverte et extensible du Web, qui est également une caractéristique clé de la croissance sans limite du Web. Malgré les indiscutables bénéfices que fournit le Web, les mêmes principes qui ont permis au Web de documents de prospérer n'ont jusqu'à récemment pas été appliqués aux données. Traditionnellement, les données

[51] Nous traduisons

publiées sur le Web ont été rendues disponibles sous forme de dépôts bruts dans des formats comme le CSV ou le XML, ou exprimées en tableaux HTML, en sacrifiant beaucoup de leur structure et de leur sémantique. Dans le Web hypertextuel conventionnel, la nature des relations entre deux documents liés est implicite, le format des données, c'est-à-dire le HTML, n'étant pas suffisamment expressif pour permettre que des entités individuelles décrites dans un document puissent être connectées par des liens typés à des entités en rapport avec les premières. Toutefois, dans les dernières années, le Web a évolué d'un espace informationnel global à un espace où à la fois les documents et les données sont liés » (29, Bizer *et al.*). Deux points importants sont à noter ici : d'une part la fragmentation d'un document en entités individuelles, d'autre part le typage, c'est-à-dire la qualification sémantique, des liens entre les entités. De ces deux points de vue, si nous avons vu en Vannevar Bush un « grand ancêtre » du Web de documents, un plus grand ancêtre encore, pour le Web de données cette fois-ci, pourrait être invoqué ici en la personne de Paul Otlet, qui imagine en 1934, à la fin de son *Traité de documentation*, le « livre téléphoté », qu'il décrit ainsi : « ici, la Table de Travail n'est plus chargée d'aucun livre. À leur place se dresse un écran et à portée un téléphone. Là-bas, au loin, dans un édifice immense, sont tous les livres et tous les renseignements, avec tout l'espace que requiert leur enregistrement et leur manutention, avec tout l'appareil de ses catalogues, bibliographies et index, avec toute la redistribution des données sur fiches, feuilles et en dossiers, avec le choix et la combinaison opérée opérés par un personnel permanent bien qualifié. Le lieu d'emmagasinement et de classement devient aussi un lieu de distribution, à distance avec ou sans fil, télévision ou télétaugraphie. De là, on fait apparaître sur l'écran la page à lire pour connaître la réponse aux questions posées par téléphone, avec ou sans fil. Un écran serait double, quadruple ou décuple s'il s'agissait de multiplier les textes et les documents à confronter simultanément ; il y aurait un haut-parleur si la vue devait être aidée par une donnée ouïe, si la vision devait être complétée par une audition. Une telle hypothèse, un Wells certes l'aimerait. Utopie aujourd'hui parce qu'elle n'existe encore nulle part, mais elle pourrait bien devenir la réalité de demain pourvu que se perfectionnent encore nos méthodes et notre instrumentation. Et ce perfectionnement pourrait aller peut-être jusqu'à rendre automatique l'appel des documents sur l'écran (simples numéros de classification, de livres, de pages) ; automatique aussi la projection consécutive, pourvu que les données aient été réduites en leurs éléments analytiques et disposées pour être mises en œuvre par les machines à sélection » (31, Otlet).

Le lien entre le Web sémantique de Tim Berners-Lee et la vision de Paul Otlet est d'ailleurs fait par l'auteur collectif R. T. Pédauque dans l'ouvrage *La redocumentarisation du monde* : « Pour éclairer la relation entre le numérique et le social, on peut faire un parallèle entre le rôle du document-papier imprimé dans l'émergence des sociétés modernes et celui du document numérique dans les transformations sociales auxquelles nous assistons. C'est ce que nous appelons la *documentarisation*. La première documentarisation a accompagné l'industrialisation,

la mise en place de l'État au sens moderne ou encore, parmi d'autres illustrations possibles, la montée des sciences positives. Une figure comme le belge Paul Otlet et sa tentative de fonder une "documentologie" au début du XX^{ème} siècle est emblématique du besoin de rationnaliser la prolifération documentaire qui s'est installée. Aujourd'hui, certains pensent que nous sommes entrés dans un nouveau processus de modernisation. De la même manière, le document, en basculant dans le numérique, se transforme, accompagne et amplifie ces changements de plus grande ampleur. Tim Berners-Lee et sa volonté de développer un "Web sémantique" joue un rôle comparable à celui de Paul Otlet un siècle plus tôt. Nous assistons, sans doute, bien à une redocumentarisation, c'est-à-dire une nouvelle forme de documentarisation qui reflète ou tente de refléter une organisation post-moderne de notre rapport au monde, repérable aussi bien dans les sphères privée, collective, et publique. Comme dans la précédente modernisation le document participe au processus et y joue le même rôle-clé, mais il s'est transformé au point que l'on peut se demander s'il s'agit encore de la même entité » (32, Pédauque). Nous avons déjà rencontré le terme de « redocumentarisation » en 1.2.3, tel que défini par Manuel Zacklad, comme documentarisation seconde. La redocumentarisation concernait donc un document. Ici, nous voyons que le terme désigne un mouvement général des sociétés contemporaines. Les deux sens ne sont pas nécessairement contradictoires, la redocumentarisation au sens de Manuel Zacklad pouvant constituer une caractéristique de la redocumentarisation au sens de R.T. Pédauque, ce qui nous permet d'esquisser un lien entre les deux sources d'enrichissement des métadonnées que nous avons envisagées dans la première partie de ce travail : les données liées et les métadonnées générées par les utilisateurs. L'article introductif d'Evelyne Broudoux et Claire Scopsi au dossier du numéro 36 de la revue Études de Communication, intitulé *L'enjeu des métadonnées dans un contexte de « redocumentarisation »*, explicite ce lien : « de l'hypothèse, formulée par Pédauque, d'une redocumentarisation nous retiendrons tout d'abord la recomposition des médiations documentaires, par le biais notamment de la télédocumentation qui donne l'initiative au lecteur, jusqu'à donner l'impression d'un effacement des médiations. La malléabilité du document numérique, la création de multiples outils permettant aux usagers d'agir sur le document, conduisent à un rééquilibrage des rôles d'auteur-émetteur et de lecteur-récepteur, tout en modifiant notre perception de la notion de document original. Ce phénomène rend plus crucial la communication, en marge du document lui-même, d'éléments décrivant pourquoi, comment et par qui les états successifs de ce contenu ont été élaborés » (33, Broudoux et Scopsi). Ces éléments sont bien entendu des métadonnées, tout autant qu'une autre catégorie évoquée par les auteurs dans l'article, de contenus générés par les usagers : « la production croissante de "Users Generated Content" (UGC) ou données produites par l'usager, dans des sites aux caractéristiques assez diverses grossièrement regroupées sous le terme de web 2.0 (blogs, sites de partage de ressources, applications collaboratives…) produit une forme particulière de métadonnées exploitables pour l'accès à

l'information : les tags appliqués par les utilisateurs soit dans une logique "communautaire » (l'usager fait don de ses métadonnées à une communauté pour construire un système de recherche performant), soit dans une logique personnelle (l'usager applique des tags sur une ressource pour la retrouver plus commodément) soit dans une logique promotionnelle (pour inciter un public plus nombreux à accéder à la ressource) » (ibid.). Nous considérons les trois formes d'UGC qui viennent d'être décrites également intéressantes dans l'optique qui est la nôtre.

A travers les textes de Vannevar Bush et de Paul Otlet, nous pouvons retracer l'extension de notre problématique de l'abondance d'information, qui ne concernait dans l'Antiquité que quelques personnes, les bibliothécaires d'Alexandrie, puis à l'époque d'Otlet et de Bush les universitaires et les chercheurs, et à présent la plupart des habitants des pays développés. Nous pourrions remarquer que cette problématique a été fortement accrue, pour tout un chacun, par le développement du Web. Ce serait ignorer d'une part que le déluge informationnel préexistait au Web et que le Web l'accroît mais aussi le rend que plus évident, ignorer d'autre part qu'il nous donne des outils pour que ce déluge ne soit pas subi. C'est encore plus vrai avec le web de données d'une part, et avec le web 2.0 d'autre part. Les deux peuvent être mis en relation comme deux aspects de la « redocumentarisation du monde », et nous verrons par la suite qu'une alliance entre les deux peut être souhaitable, qu'une utilisation par le deuxième des outils développés par le premier peut permettre l'enrichissement des métadonnées d'une manière exponentielle. Mais quels sont les outils intellectuels et technologiques des données liées ?

3.1.2. Les solutions

Dans la mesure où le Web consiste pour l'essentiel à rendre disponible une ressource, un document, ou à aller consulter une ressource présente sur une machine éloignée, nous pouvons considérer que l'architecture du Web repose pour l'essentiel sur trois concepts : l'adressage des ressources, le protocole de transfert d'informations entre machines, le langage de représentation des ressources. L'adressage d'une ressource est obtenu par l'emploi d'*Uniform Resource Locator* (URL), qui permet à la fois de l'identifier et de la localiser. Le protocole de transfert est l'*HyperText Transfer Protocol* (HTTP). Le langage pour représenter les ressources et indiquer les éventuels liens de cette ressource avec d'autres est l'*HyperText Markup Language* (HTML). Qu'en est-il dans le Web de Données ?

L'élément de base en est l'*Uniform Resource Identifier* (URI), qui permet d'identifier toutes sortes d'entités, qu'elles soient de nature numérique, comme un document numérique, qu'elles appartiennent au monde physique, comme des objets, des personnes, des organisations, ou même au monde idéel, comme des concepts. L'URI peut n'être qu'un identifiant, mais elle peut également posséder à sa racine un

élément de localisation dans un réseau, par exemple le Web : elle devient alors une URL. Pour comprendre d'une phrase la différence d'utilisation entre URL et URI, nous pouvons reprendre la formule de Fabien Gandon, de l'INRIA, selon laquelle « L'URL permet d'identifier ce qui existe sur le Web ; l'URI permet d'identifier sur le web ce qui existe ». « Résoudre » une URI, c'est-à-dire mettre en œuvre la méthode appropriée pour y accéder, permet de rendre celle-ci « actionnable », par exemple d'accéder à une représentation de la ressource (on dit : « déréférencer » l'URI). Pour reprendre notre exemple de la première partie, « Victor Hugo », « *Les Misérables* », mais aussi « est l'auteur de », sont tous trois des ressources, et à ce titre doivent disposer chacun d'un URI.

Le langage pour représenter les ressources diffère de celui du Web de documents. Il y a en fait plusieurs syntaxes (on parle aussi de « sérialisations ») possibles pour représenter les ressources, qui se basent sur un même modèle de données fourni par le *Resource Description Framework* (RDF). Le RDF, comme son nom l'indique, n'est pas un langage mais un *cadre* qui définit la structure d'une représentation des données. Cette structure est le triplet Sujet-Prédicat-Objet. Sujet ? Victor Hugo - Prédicat ? Est l'auteur de - Objet ? *Les Misérables*. Ce triplet peut être par exemple exprimé dans la syntaxe RDF/XML. Il peut être vu comme un graphe orienté et étiqueté, le sujet et l'objet étant des nœuds et la propriété l'arête qui relie les deux nœuds. D'autres arêtes sont susceptibles de partir de la ressource qui est ici sujet : « est né en » ; ou y arriver (la ressource devenant objet dans un autre triplet) : « est l'enfant de ». Tout ajout de triplet reliant un nouveau nœud à un nœud déjà existant augmente la taille du graphe. De même, tout triplet reliant des nœuds appartenant à des graphes auparavant disjoints, lie ces graphes. Des bases de données spécifiques existent pour stocker ces triplets, appelés des *triplestores*.

Afin de rechercher des informations dans les graphes stockés dans les triplestores, il est possible de formuler des requêtes grâce au langage SPARQL (*SPARQL Protocol and RDF Query Language*). Celui-ci peut être rapproché du langage de requête *Structured Query Language* (SQL), approprié aux bases de données relationnelles. Afin de donner accès à ses données de façon simple, il est possible de proposer un *SPARQL Endpoint*, c'est-à-dire une URL qui accepte les requêtes SPARQL et en renvoie le résultat.

Le langage RDFSchema (RDFS) permet de documenter et d'organiser le vocabulaire utilisé dans des descriptions RDF, tout d'abord en associant un vocabulaire à un espace de nommage (qui permet de préciser le sens des termes employés), mais aussi en décrivant les classes de ressources et leurs relations hiérarchiques, et notamment les propriétés (par exemple en précisant les types de ressources qui sont liées par une propriété). En ce sens, RDFS est un vocabulaire qui fournit l'opportunité de créer d'autres vocabulaires plus complexes permettant de structurer les descriptions. De tes vocabulaires pourront par exemple être exprimés dans le langage OWL (*Web Ontology Language*). Avec RDFS, mais surtout avec OWL, nous retrouvons la notion d'ontologie abordée dans la première partie, qui —

rappelons-le — permet de traduire la représentation, le modèle que nous nous faisons d'un certain domaine et permet d'opérer à partir de ce modèle des inférences logiques.

Cette description de quelques standards fondamentaux du web de données est extrêmement succincte et mériterait à la fois de nombreuses extensions et des approfondissements. Nous voulions ici donner quelques éléments lexicaux permettant de se faire une première représentation de la manière dont fonctionne le Web de données. Remarquons que la proposition et le maintien de ces standards par le *World Wide Web Consortium* (W3C) est le gage de l'interopérabilité que promet et promeut le Web de données. Nous voudrions enfin opérer quelques distinctions qui nous semblent utiles entre données ouvertes, données liées et données en RDF, en nous appuyant sur un billet de Christopher Gutteridge, consultable sur le blog *Southampton Web and Data Innovation Team* : « "Données ouvertes" est une politique ; "Données Liées" est une approche et "RDF" est une structure de données. **Données ouvertes** : les données ouvertes sont des données que vous pouvez utiliser plus ou moins gratuitement. Elles sont généralement disponibles sur le web, et recourent à des formats non-propriétaires comme XML, CSV. Une définition extrémiste en est : des données avec des règles claires de droit d'auteur et une licence libre (qui permet la réutilisation commerciale), disponibles par une URL ou une API bien documentée sans aucune restriction, dans des formats qui sont complètement ouverts (c'est-à-dire sans souci de patente, etc.). Une définition plus modérée en est : "disponibles comme données sur le web sous une forme telle que les gens peuvent faire des choses avec". Certaines Données Ouvertes sont aussi des Données liées et en RDF, mais probablement moins de la moitié d'entre elles. **Données liées** : les Données Liées sont des données qui contiennent des liens avec d'autres jeux de données. Elles vont généralement utiliser des URI qui peuvent être résolues, pour découvrir plus de faits. Il n'est pas essentiel que les URI puissent être résolues, mais il est toutefois très utile d'avoir deux jeux de données qui utilisent les mêmes identifiants. Les URI sont sans ambiguïté. Toutefois, certaines données n'ont pas beaucoup d'intérêt à être liées, ou les coûts sont trop élevés et dissuadent de le faire. Les données liées sont souvent ouvertes, mais ne le sont pas nécessairement — par exemple, vous pouvez disposer de données confidentielles liées à d'autres sources de données. Un bon exemple est le calendrier de cours d'un étudiant, qui est confidentiel, mais relié à des données sur les salles et les modules qui sont ouverts. Presque toutes les Données Liées sont exprimées en RDF mais on pourrait avoir des liens en XML, KML, CSV, etc. Simplement, RDF est conçu en vue de permettre des liens. **RDF** : RDF est une structure de données utile pour créer des données interopérables. Il y a un certain nombre de formats de fichiers pour échanger ces données. Le plus répandu est le RDF/XML. Le meilleur (à mon avis) est Turtle. Le plus simple est N-Triples, où l'on écrit les données avec un fait par ligne. On peut aussi exprimer des données RDF embarquées dans du HTML comme "RDFa". La structure de RDF rend trivial de fusionner des données de sources multiples — ce sont tous des triplets. RDF suppose également que vous voudrez lier les données vous-mêmes, ou que d'autres voudront

faire des liens avec vos données. Vous pouvez publier des données RDF qui deviennent des données liées à mesure que des gens s'y relient, comme en publiant des pages sur le web. RDF est juste une manière de structurer les données et en tant que tel n'est pas toujours ouvert ni toujours lié. **Données Liées Ouvertes** : (LOD) c'est un terme répandu, et comme on peut l'imaginer elles sont aussi en RDF. L'idée est de ne pas être découragé par le fait de lier. Ajoutez des liens quand ils ajoutent de la valeur à vos données et vont aider des gens utilisant vos données (vous-même inclus) à faire plus avec elles »[52] (34, Gutteridge). Le fait de lier les données, même s'il n'est pas consubstantiellement lié au fait de les publier sur le Web, prend son sens dans ce cadre, celui du Web de données, et permet à ce dernier d'accomplir ses promesses.

3.1.3. Les promesses

Si nous en revenons au texte de Bizer, Heath et Berners-Lee, le ton peut sembler très mesuré quand il s'agit d'envisager ce que permet le Web de données : « Ce web des données permet de nouveaux types d'applications. Des navigateurs de données liées génériques permettent aux usagers de commencer à naviguer dans une source de données puis de naviguer par les liens dans des sources de données qui y sont liées. Des moteurs de recherche parcourent le web en suivant des liens entre sources de données et fournissent des possibilités de recherche élaborées sur des données agrégées, de la même manière qu'une base de données locale est interrogée aujourd'hui. Le Web de Données ouvre également de nouvelles possibilités pour des applications spécifiques à un domaine. Contrairement aux applications composites [*mashups*] du Web 2.0 qui sont restreintes à un ensemble fixé de sources de données, les applications de Données Liées opèrent sur un espace de données global non prédéterminé. Cela leur permet de fournir des réponses plus complètes à mesure que de nouvelles sources de données apparaissent sur le Web » (29, Bizer *et al.*). Nous voyons dans la première partie de ce passage l'expression de la transposition aux sources de données de l'idée fondatrice du Web de documents : établir des liens entre sources de données se trouvant dans des machines liées en réseau (avec en plus la possibilité de requêtes élaborées). Dans la deuxième partie, la différence est marquée par rapport à ce qui existe déjà dans le web 2.0, différence qui ne tient pas tant à la fonctionnalité elle-même, qui peut être déjà existante, qu'à la manière dont cette fonctionnalité est ancrée dans des jeux de données et peut donc évoluer selon un dynamisme propre.

Ce ton que nous avons qualifié de mesuré est effectivement modeste si nous le comparons à l'article séminal de Tim Berners-Lee, en 2001, quand il annonce qu'« une nouvelle forme de contenu Web compréhensible par les ordinateurs va déchaîner une

[52] Nous traduisons.

révolution de nouvelles possibilités »[53] et imagine avec ses coauteurs un scénario présentant des personnages contrôlant les objets et accédant à toutes les informations qui leur sont nécessaires avec la plus grande facilité. Ou encore lorsqu'il emploie en 2007 l'expression de Graphe Global Géant [*Global Giant Graph*]. Cette différence de ton n'est sans doute pas sans rapport avec les différences de dénomination. Fabien Gandon explique dans l'article collectif *Enjeux et technologies : des données au sens*, que « ce qui rend parfois difficile l'abord du domaine, c'est que l'initiative dans laquelle il s'inscrit est présentée sous des appellations différentes qui semblent privilégier, pour chacune d'elles, un aspect particulier de l'architecture globale :

- le « web de données » (*web of data*) insiste sur la possibilité qui nous est offerte d'ouvrir les silos de données de toutes tailles, depuis l'application individuelle de notre carnet d'adresses jusqu'aux immenses bases de génomique, et de les échanger, de les relier, de les mélanger selon nos besoins ;

- l'expression « données ouvertes liées » (*linked open data*) met l'accent sur l'opportunité d'exploiter des données ouvertes dans nos applications et rappelle qu'une grande valeur ajoutée réside dans l'utilisation et la réutilisation des URI pour joindre des assertions de différentes provenances mais portant sur un même sujet ;

- le « gigantesque graphe global » (*Global Giant Graph*) remet en perspective ces milliers de liens entre données distribués sur le Web et le fait que, à travers les points de jointure que sont les URI, la structure de données qu'ils tissent est un graphe d'envergure mondiale ;

- enfin, le « web sémantique » (*semantic web*) met en avant la possibilité d'échanger les schémas de nos données et la sémantique associée afin d'enrichir la gamme des traitements automatiques qui peuvent leur être appliqués » (35, Bachimont *et al*., 30).

En ce sens, parler de Web de données plutôt que de Web sémantique peut être une manière de prévenir les déceptions face aux espoirs soulevés par l'idée du web sémantique. A l'heure où nous écrivons, de nombreux commentateurs estiment que le Web sémantique n'a pas rempli ses promesses. Les raisons avancées peuvent tenir à la prolifération des ontologies ou à la complexité d'en établir une, à la perception d'un Web qui se « referme » — allant donc à l'encontre du projet du Web sémantique, aux contraintes légales liées aux licences. Nous pouvons quant à nous risquer l'hypothèse que si le Web sémantique ne se développe pas au même tempo où s'est développé le Web des documents, c'est peut-être qu'il est amené à croître en premier lieu dans ses usages professionnels. Si certaines des critiques adressées au Web sémantique peuvent être entendues, il serait prématuré d'y voir une impasse alors que Google, avec le *Google Knowledge Graph* — l'apport d'informations sémantisées lors de l'utilisation de son moteur de recherche étant d'ailleurs « invisible » pour la plupart

[53] BERNERS-LEE, Tim, HENDLER, James et LASSILA, Ora.*The Semantic Web*.[En ligne] Scientific American. Mai 2001. Disponible sur : <https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf> (Consulté le 20 août 2018)

des utilisateurs, et Facebook, avec l'*Open Graph*, mettent en place des initiatives en ce sens. Jean Charlet et Gérald Kembellec, dans leur article *Du Web sémantique au web de données, quels enjeux professionnels*, remarquent de façon liminaire que « si quelques industriels se sont approprié le web de données dès ses prémices, principalement en recherche et développement, ses implications concrètes restent peu connues du grand public, ce qui ne signifie pas qu'il n'y ait pas d'impact tangible. En effet, les utilisations des principes du web de données sont présentes dans plusieurs aspects de l'Internet d'aujourd'hui, à commencer par leur prise en compte dans les récentes évolutions de l'algorithme du leader des moteurs de recherche » (36, Charlet et Kembellec). Cela nous confirme dans l'idée que le web de données pourrait devenir primordial et omniprésent dans le monde professionnel sans bénéficier de la même exposition qu'a eu le web de documents ; ce qui s'explique aussi par le fait, noté par les auteurs, que « plus les volontés d'interaction — et possiblement de raisonnement — augmentent, plus le modèle doit être formel pour permettre le développement informatique d'applications » (ibid.). Là encore, nous ne retrouvons pas la facilité de mise en œuvre qui a fait le succès du Web de documents. Les auteurs concluent l'article pour introduire au dossier qui le suit dans la revue, en signalant que « ce dossier met ainsi en évidence que l'articulation entre les contenus, principalement textuels, du Web et des ontologies est en quelque sorte l'aboutissement de la vision du web sémantique tel que promu initialement par Tim Berners-Lee. La transition du web sémantique vers le web de données se fera par qualification desdites données grâce à des schémas de description communs qui permettront de promouvoir des données liées » (ibid.). C'est dans cette articulation que se trouvent les deux plus importantes promesses du paradigme des données liées : l'organisation des connaissances par des ontologies, et la possibilité qu'elle ouvre de métadonnées lisibles, « compréhensibles » sémantiquement, par les machines. De ce point de vue, si nous risquions une métaphore mathématique, nous pourrions dire que l'ajout une à une de métadonnées descriptives d'une ressource peut être vu comme une addition, que le croisement de jeux de métadonnées portant sur une même ressource serait une multiplication, et que le lien entre métadonnées dans le paradigme des données liées serait une exponentiation.

3.2. Données liées sociales dans le secteur commercial du livre

3.2.1. Des métadonnées liées et socialement construites : une utopie ?

« Les métadonnées sont “des données à propos de données” – des informations comme les mots-clés, le nombre de pages, le titre, le nombre de mots, un résumé, une localisation, une unité de gestion des stocks, un ISBN, et ainsi de suite. Les métadonnées précises, générées par les hommes, ont été récemment très tendance, particulièrement dans le monde du XML. Un scénario typique ressemble à ça : plusieurs fournisseurs se réunissent et se mettent d’accord sur un standard de métadonnées – une définition de type de document ou un schéma – dans un certain domaine, disons les machines à laver. Ils se mettent d’accord sur un vocabulaire commun pour décrire les machines à laver : la taille, la contenance, la consommation d’énergie, le prix. Ils créent des bases de données lisibles par les machines à partir de leurs stocks, qui sont tout ou partie disponibles pour des agents de recherche et d’autres bases de données, afin qu’un consommateur puisse entrer les paramètres de la machine à laver qu’il cherche et interroger simultanément plusieurs sites pour obtenir une liste exhaustive des machines à laver disponibles qui satisfont ses critères. Si tout un chacun souscrivait à un tel système et créait de bonnes métadonnées en vue de décrire leurs biens, leurs services et leurs informations, il serait trivial de chercher sur Internet des résultats très pertinents, contextualisés : un fan pourrait trouver toute la musique téléchargeable dans un genre donné, un fabricant pourrait efficacement découvrir des fournisseurs, les voyageurs pourraient facilement choisir un hôtel pour leur voyage à venir.

Un monde de métadonnées complètes, fiables serait une utopie. C’est aussi un rêve fumeux, fondé sur l’illusion, l’arrogance des *nerds* et des opportunités commerciales hystériquement exagérées. […] Il y a au moins sept obstacles insurmontables entre le monde tel que nous le connaissons et la méta-utopie. Je vais les énumérer ci-dessous :

- Les gens mentent. Les métadonnées existent dans un monde compétitif. Les fournisseurs sont en compétition pour vendre leurs biens, les zozos sont en compétition pour transmettre leurs théories cinglées (mea culpa), les artistes sont en compétition pour l’audience […]
- Les gens sont paresseux […] Ici, dans la Tout d’Ivoire de l’Info, nous comprenons l’importance de créer et d’entretenir d’excellentes métadonnées pour nos informations. Mais les civils de l’info sont remarquablement cavaliers avec leurs informations. Votre tante paumée vous envoie des emails sans indiquer d’objet, la moitié des pages de Geocities sont intitulées “SVP donnez un titre à cette page” et votre patron range tous ses fichiers sur son bureau avec des titres aussi utiles que “UNTITLED.DOC” […]

- Les gens sont stupides. Même lorsqu'il y a un bénéfice à créer de bonnes métadonnées, les gens refusent avec acharnement de mettre du soin et de l'application dans la création de leurs métadonnées […]
- Mission Impossible – connais-toi toi-même […] Les gens sont de médiocres observateurs de leurs propres comportements […]
- Les schémas ne sont pas neutres […] Toute hiérarchie d'idées implique la prééminence de certains axes sur d'autres […]
- Les métriques influencent le résultat […] Se mettre d'accord sur une aune commune pour mesurer ce qui est important dans un quelconque domaine privilégie nécessairement les items qui ont un bon score sur cette dimension, indépendamment de la pertinence globale de ces items […]
- Il y a plus d'une manière de décrire quelque chose […] Des gens raisonnables peuvent à jamais être en désaccord sur la manière de décrire quelque chose [...]

Doit-on jeter les métadonnées, alors ?

Bien sûr que non. Les métadonnées peuvent être tout à fait utiles, si l'on sait en prendre et en laisser. La méta-utopie ne verra jamais le jour, mais les métadonnées sont souvent un bon moyen de formuler des hypothèses approximatives sur les informations qui flottent sur Internet.

Certaines sortes de métadonnées implicites sont terriblement utiles, en fait. Google exploite les métadonnées du World Wide Web : en examinant le nombre de liens pointant vers une page (et le nombre de liens pointant vers chaque *linker*), Google peut déduire des statistiques sur le nombre d'autorités du Web qui croient que cette page est suffisamment importante pour s'y relier, et par conséquent peut produire des estimations extrêmement fiables de la bonne réputation des informations sur cette page.

Cette sorte de métadonnées d'observation est beaucoup plus fiable que ce que les humains créent afin que leurs documents soient trouvés. Elle se débarrasse du baratin commercial, de l'illusion, et des conflits de vocabulaires.

Plus largement, cette sorte de métadonnées peut être conçue comme une sorte de pédigrée : qui pense que ce document a de la valeur ? A quel point les jugements de valeur de cette personne ont été corrélés aux miens par le passé ? Cette sorte de validation implicite des informations est une bien meilleure candidate à être une panacée de la récupération d'informations que tous les schémas du monde conjugués »[54] (37, Doctorow).

Après avoir entrevu les promesses des données liées et au moment d'imaginer ce que ce paradigme et les métadonnées sociales pourraient apporter au secteur commercial du livre, ce texte peut faire l'effet d'une douche froide. Qui l'a écrit et quand ? Il s'agit de Cory Doctorow, blogueur, journaliste et auteur de science-fiction, et le texte date de 2001. Le ton est volontairement relâché et mordant. Nous pourrions

[54] Nous traduisons.

considérer qu'il s'agit d'un billet d'humeur sans grande importance. Néanmoins, ce texte pointe de façon suffisamment pertinente l'illusion d'une « utopie » d'un monde de métadonnées abondantes et de qualité pour que deux des ouvrages que nous avons cités dans la première partie de notre travail sentent le besoin de lui répondre. L'article de Tony Gill *Metadata and the Web* considère le premier argument de Cory Doctorow comme le plus probant. « Il est facile pour des éditeurs Web peu scrupuleux d'embarquer des "balises META spam" – des métadonnées descriptives délibérément trompeuses ou malhonnêtes – dans leur pages Web »[55] (38, Gill, p. 33). Mais l'argument est finalement écarté : « parce que la plupart des moteurs de recherche n'utilisent pas les métadonnées embarquées, il n'y a en général aucune motivation pour la vaste majorité des éditeurs Web honnêtes à consacrer du temps et des efforts à ajouter cette information potentiellement utile à leurs pages » (ibid.). Quant aux autres points, ils sont considérés comme moins convaincants au regard des pratiques des musées, des bibliothèques et des archives : « Les bibliothécaires, les équipes de documentalistes des musées, et les archivistes sont généralement des professionnels de l'information consciencieux, qualifiés, et ne sont pas le plus souvent malhonnêtes, paresseux ou stupides. Ils ont une longue tradition dans l'usage d'ensembles d'éléments de métadonnées standards […], de schémas de classification, de vocabulaires contrôlés et de règles de catalogage propre à une communauté […] pour décrire des ressources selon des manières standardisées qui ont été développées au cours de décennies d'efforts collaboratifs de construction de consensus. En réalité, ils ont démontré la valeur des métadonnées créées par des humains pendant des siècles » (ibid., p.34). Cette réaffirmation de la valeur des standards semble totalement ignorer les questions soulevées par Doctorow sur la diversité des descriptions possibles d'un objet et sur l'absence de neutralité des schémas de classification. Quant à l'éloge des métadonnées « autorisées » par des professionnels de l'information, il semble louper le fond de l'argument de Doctorow qui s'intéressent au contraire à la création de métadonnées hors des cercles des professionnels de l'information. Toutefois, Tony Gill poursuit en remarquant qu'« un autre récent développement dans le champ des métadonnées qui affaiblit significativement le propos de Doctorow sont les folksonomies […] Si une personne appose le terme *impressionnisme* à un site Web, ça ne dit pas grand-chose. En revanche, si plusieurs centaines de personnes utilisent ce terme et que c'est l'étiquette la plus couramment utilisée pour ce site, il y a fort à parier que le site Web est sur l'impressionnisme et l'art impressionniste ». Mais la suite du propos et la conclusion de l'article montre que ce léger écart par les folksonomies visait seulement à appuyer l'utilité et la légitimité – contre Doctorow – de données générées par des humains : « Tous les standards nécessaires et les composants technologiques pour faciliter le partage de connaissance au sein d'une communauté sont à présent en place […] En combinant ces différentes composantes de manière créative pour donner accès au riche contenu

[55] Nous traduisons.

d'information qui se trouve dans les musées, les bibliothèques et les archives, il devrait être possible de bâtir un Web Sémantique global distribué de contenu numérique culturel et les outils de recherche verticalement intégrés adaptés pour aider les usagers à trouver le contenu qu'ils y cherchent » (ibid., p. 37). L'intérêt que nous supposons sincère pour les folksonomies est ici employé comme boutefeu contre les arguments de Cory Doctorow, mais le point final de l'article montre que le modèle de pensée de l'auteur conserve l'usager dans un rôle de consommateur passif de métadonnées « autorisées » parce que créées par des professionnels de leur domaine et encadrées par des standards – le web sémantique apparaissant comme un standard souhaitable.

Alemu et Stevens, dans *An Emergent Theory of Digital Library Metadata*, donnent une place entièrement différente aux arguments de Doctorow. Celui de la non-neutralité des schémas intervient dans le chapitre sur les vocabulaires contrôlés, alors que les auteurs remettent en question la prétention de non-recouvrement et de complétude des catégories qui constituent ces vocabulaires. C'est l'occasion pour eux de rappeler qu'« en substance, il y a quatre groupes de créateurs de métadonnées : les bibliothécaires, les auteurs, les usagers et les machines »[56] (12, Alemu et Stevens, p. 20) . Tout en reconnaissant que « les métadonnées créées par les bibliothécaires sont considérées être d'une qualité assez élevée, en particulier en ce qui concerne la précision, la complétude et la cohérence », ils pointent deux écueils : la difficulté pour cette manière de créer des métadonnées de « passer à l'échelle » à mesure que les collections grandissent et – rejoignant ici Doctorow – la possibilité de différences d'interprétation d'un livre, aboutissant à des descriptions différentes. En notant, toujours avec Doctorow, que les auteurs peuvent biaiser la représentation du contenu de leurs livres pour les mettre en valeur dans des recherches, ils introduisent une distinction essentielle pour la suite de leur propos entre les métadonnées *a priori* que fournissent les auteurs et les bibliothécaires, et ce qu'ils définissent plus tard comme des métadonnées *post hoc* qui sont le fait des usagers. Nous nous étions posés la question, au point 1.1.1., de savoir si les métadonnées venaient forcément *après* les données qu'elles caractérisaient, conformément au préfixe grec que contient le mot. Les métadonnées *a priori* seraient plutôt pour l'usager des « co-données » ou des « épi-données », et seules les métadonnées *post hoc* mériteraient vraiment leur nom. Comme le rappellent Alemu et Stevens « quand on parle du processus de création des métadonnées, il est important de garder à l'esprit que les standards de métadonnées actuels supposent que les auteurs créent les œuvres, que les libraires créent les métadonnées et que les usagers accèdent à des objets informationnels. Par conséquent, les métadonnées basées sur des standards sont majoritairement générées a priori, soit avant que les usagers aient accès à un objet informationnel en particulier, comme un livre » (ibid., p. 21). La cohérence et la précision des métadonnées, dans cette approche par les standards, sont portées par la structure

[56] Nous traduisons.

des métadonnées, et l'attention est portée sur leur provenance et leur qualité. Ceci ne va pas sans contrepartie, notamment le fait que le vocabulaire employé ne soit pas celui des créateurs de contenus ou des usagers, que les terminologies soient dépassées et ne représentent plus la vision du monde des usagers. Quel est alors « le futur des standards de métadonnées » ? Si Alemu et Stevens considèrent que le modèle RDF et que les langages RDF Schema et OWL permettent de meilleures descriptions des ressources, ils constatent néanmoins que ces solutions ne sont encore pas ou peu mises en place dans le monde des bibliothèques : « il faut noter que récemment, en 2011 […] la British Library, La Bibliothèque Nationale de France et le Bibliothèque Numérique Europeana ont dévoilé leurs plans d'ouvrir leurs fonds patrimoniaux sous forme de données liées. Cependant, un rapport publié par le World Wide Web Consortium en octobre 2011 reconnaît la faible adoption des données liées dans les bibliothèques, alors qu'il recommandait aux bibliothèques d'inclure les principes des données liées qui incluent, entre autres, les identifiants uniformes de ressource (URI), RDF, le protocole et le langage de requête SPARQL et OWL. Le rapport indique l'importance de rendre les données bibliographiques des bibliothèques ouvertes et librement accessibles sous une forme qui soit "partageable, extensible et facilement réutilisable". Alors que le web sémantique n'a pas fait d'avancée significative dans le domaine des bibliothèques, son rôle potentiel pour l'encodage, la représentation et le partage des métadonnées est signalé. Malgré l'intérêt croissant pour ces nouvelles approches, un large débat est toujours en cours sur la question de savoir si des changements incrémentaux des modèles et des formats de notices traditionnellement centrés sur les bibliothèques peuvent suffire ou s'il se trouvent des raisons convaincantes de les abandonner tout ou partie et d'adopter de nouvelles approches » (ibid., p. 27). Lorsqu'on s'intéresse au web sémantique, force est de constater que les travaux des bibliothèques dans le domaine sont nombreux et informés. Si comme le soutiennent les auteurs un débat est en cours, les partisans des données liées n'en restent pas aux discours, comme en témoigne le projet de la BnF data.bnf.fr, basé sur le web sémantique et proposant un SPARQL *endpoint*. Toutefois, pour reprendre le fil de l'argumentation d'Alemu et Stevens, il est à noter que les arguments de Doctorow qu'ils reprennent sont précisément ceux que Tony Gill laissait de côté, et qui leur permettent de porter la critique sur les standards, sans vouloir s'en débarrasser complètement : « actuellement, les approches basées sur les standards sont confrontées à des défis comme l'illisibilité pour une machine, le manque d'intégration et d'interopérabilité avec les métadonnées hors des bibliothèques, la duplication de métadonnées, la lenteur à s'adapter aux besoins des usagers en termes de vocabulaires d'usagers (termes de recherche) et le manque de relations avec les usagers. Certains de ces défis peuvent être résolus par des moyens techniques, comme le changement des formats de métadonnées des bibliothèques de MARC à des formats compatibles avec le web, d'autres sont conceptuels et d'autres encore sont sociaux et institutionnels » (ibid., p. 28). L'argument le plus fondamental des auteurs contre les standards semble néanmoins le rapport aux usagers : « Les

principes actuels des métadonnées montrent des limites au regard des besoins changeants des usagers, de l'existence d'interprétations multiples et des changements dans les tendances technologiques, comme les médias sociaux et le web 2.0 » (ibid.).

Finalement, le seul point sur lequel Tony Gill d'un côté, Alemu et Stevens de l'autre, pourraient se retrouver pour contester Cory Doctorow est la pertinence des métadonnées générées par des humains. Chez Alemu et Stevens, toutefois, cela inclut de façon centrale les usagers, comme nous allons le voir dans le modèle qu'ils proposent.

3.2.2. Le modèle « enrichir et filtrer »

Voyant dans le web 2.0 une manière d'aborder les métadonnées comme socialement construites, plutôt qu'un ensemble de technologies pour étiqueter, évaluer, critiquer ou recommander, Alemu et Stevens s'appuient sur l'ouvrage de Tim O'Reilly *What is web 2.0*[57] pour rappeler les concepts qui sous-tendent le web 2.0 et sont selon eux pertinents pour les métadonnées des bibliothèques. Il s'agit :

- de la collaboration bilatérale active, le web 2.0 pouvant être utilisé comme un plateforme de collaboration bilatérale permettant et encourageant la participation des usagers ;
- des usagers comme co-créateurs, ou *prosumers*[58], mot-valise formé à partir de *proactive* et de *consumer*, la participation des usagers devenant un avantage compétitif ;
- de la Sagesse des Foules, autre nom pour l'intelligence collective à grande échelle, le tout valant plus que la somme des parties (récolter les fruits de cette « sagesse des foules supposant de la coordination) ;
- de la participation variable, aucune contribution n'étant considérée comme trop modeste,
- de l'ouverture, la barrière d'entrée pour les contributeurs potentiels étant le plus possible abaissée.

Qu'en est-il dans ce modèle du contrôle de la qualité ? Selon les auteurs, citant Clay Shirky[59], il n'intervient plus avant la publication mais après : « dans un monde où publier ne coûte rien, publier quelque chose ne dit rien de sa qualité. C'est ce qui se passe après que c'est publié qui compte. Si les gens n'y renvoient pas, d'autres gens

---

[57] O'REILLY, Tim. (2005). *What is web 2.0: Design patterns and business models for the next generation of software*. Disponible sur : <http://oreilly.com/web2/archive/what-is-web- 20.html> (Consulté le 20 août 2018)

[58] Nous pourrions traduire par « prosommateurs ».

[59] SHIRKY, Clay. (2005). *Ontology is overrated: Categories, links, and tags*. Clay Shirky's writings about the internet. Disponible sur : <http://shirky.com/writings/herecomeseverybody/ontology_overrated.html> (Consulté le 20 août 2018)

ne le liront pas » (ibid., p.34). D'autre part, si l'on pense à une œuvre collective ouverte, prenons l'exemple habituel de Wikipédia, l'argument est de considérer que la communauté des usagers elle-même se régule, se corrige, permettant d'éviter les contributions dont la mauvaise qualité ferait baisser celle de l'ensemble.

Alemu et Stevens revendiquent quant à eux une approche mixte des métadonnées, qui prend en compte les avantages et les limitations de l'approche basée sur les standards et de l'approche d'une construction sociale des métadonnées. Les métadonnées du premier type sont *a priori* et permettent en premier lieu que les objets informationnels soient trouvables et découvrables, les métadonnées du second type (étiquetages, critiques d'usagers, évaluations, recommandations) sont *post hoc* et viennent améliorer les premières. Dans cette optique, enrichir les métadonnées signifie enrichir les métadonnées *a priori* en recourant à des métadonnées socialement construites, *post hoc*. Le « principe d'enrichissement des métadonnées » dépend d'abord de la diversité d'interprétation pertinentes que recèle les métadonnées, mais aussi de la granularité des métadonnées, qu'il faut entendre à la fois comme étendue des objets informationnels et niveau de détail de leur description. Pour que les usagers puissent contribuer aux métadonnées et les réutiliser, celles-ci doivent être liées. C'est l'un des intérêts majeurs du paradigme des données liées qui est ici convoqué : celui de pouvoir fragmenter une notice de métadonnées en une série d'énoncés atomiques qui peuvent être reconvertis, remotivés, recomposés pour décrire les objets informationnels. Ce paradigme peut également permettre de repérer des structures d'usage des ressources, en voyant quelles sont les terminologies et les interprétations proposées. Il permet également la classification à facettes[60], introduite par le bibliothécaire indien Ranganathan. Enfin, « des métadonnées enrichies de liens contextuels et pertinents permettraient aux usagers de naviguer avec fluidité entre des bases de données disparates et des fournisseurs d'information externes comme d'autres bibliothèques et des moteurs de recherche. En identifiant de manière globale et unique des entités (comme des œuvres, des gens, des lieux, des événements), des éléments ou des propriétés de métadonnées (auteur, titre, sujet, relations) et les valeurs correspondantes (instances), l'interconnexion offre une multitude de voies d'enrichissement des objets informationnels, qui faciliterait la découverte d'information et améliorerait l'expérience de l'usager dans les bibliothèques numériques » (ibid., p. 75). Le complément nécessaire de l'interconnexion est l'ouverture, c'est-à-dire la liberté d'accès et de réutilisation des métadonnées. Nous avons vu en 3.1.1 les différents degrés d'ouverture des métadonnées définis par Tim Berners-Lee.

Si les arguments d'Alemu et Stevens sont convaincants, en ce qui concerne la question du « passage à l'échelle » (que les approches des métadonnées basées sur les standards peuvent avoir du mal à négocier au fur et à mesure que les collections grandissent), mais aussi en ce qui concerne l'intérêt de l'ouverture aux métadonnées

[60] En décrivant un document selon plusieurs facettes (Personnalité, Matière, Énergie, Espace, Temps), on peut le faire appartenir à différentes classifications arborescentes classiques, en même temps.

socialement construites, il semble que nous soyons revenus à notre point de départ : celui de l'abondance informationnelle redoublée de celle de l'abondance de métadonnées sur les objets informationnels, cette deuxième abondance étant revendiquée par les auteurs comme très étendue, diverse par les interprétations et les dénominations. Comment les auteurs traitent-ils ce problème ? Durant les entretiens qu'ils ont menés « avec des bibliothécaires, des experts des métadonnées et des usagers des bibliothèques, le filtrage des métadonnées est apparu comme un principe important des métadonnées, que les bibliothèques et les développeurs de système devraient considérer en concevant et développant des interfaces de découverte » (ibid., p. 91). Ce filtrage, pour être efficace, doit comme le notent les auteurs être enraciné dans les besoins des usagers. Mais en quoi consiste-t-il exactement ?

Les auteurs constatent au cours de leurs entretiens une attente de simplicité lors d'une recherche dans une base de données. Ils considèrent toutefois que cette simplicité doit être permise techniquement par l'interface, plutôt que par la simplicité des métadonnées elles-mêmes, ce qui est le cas dans le filtrage *a priori* exercé par les bibliothèques. Ils notent également des besoins différenciés selon que la recherche est exploratoire, vise la découverte d'ouvrages dans un champ donné, ou qu'elle vise spécifiquement l'accès à une ressource déjà connue. Selon eux, « les métadonnées enrichies par l'usage combiné des approches des métadonnées basées sur les standards et des métadonnées socialement construites pourraient être filtrées pour l'usager afin de répondre à ses besoins individuels d'information, plutôt que par une généralisation *a priori*. Un tel filtrage devrait aussi permettre d'heureuses trouvailles [*serendipitous discoveries*] en plus des recherches d'un item connu. Le principe du filtrage des métadonnées défend l'importance de séparer le contenu des métadonnées de leur présentation contextuelle. Il introduit l'idée de contextualisation et/ou personnalisation *post hoc* des métadonnées avant la présentation, en se basant sur les besoins et les préférences des usagers » (ibid., p. 96). Il s'agit donc de recueillir des informations sur l'usager afin de lui proposer des recommandations pertinentes, ainsi que de laisser l'usager fixer des préférences et choisir la disposition, concernant l'affichage des résultats. Le modèle d'Alemu et Stevens entraîne une modification du rôle des bibliothécaires : dans la mesure où les usagers deviennent des « co-créateurs proactifs des métadonnées », les bibliothécaires doivent quant à eux devenir des « architectes de systèmes de métadonnées », plutôt que des créateurs de contenus des métadonnées. « Ainsi, les bibliothécaires deviennent les experts dans l'offre de structure, de granularité et d'interopérabilité des métadonnées *post-hoc*, plutôt que dans l'entretien du contenu, ce qui implique en retour une reconceptualisation de la qualité des métadonnées » (ibid., pp. 99-100). Les auteurs invitent ici à abandonner une vision perfectionniste des métadonnées pour une vision beaucoup plus pragmatique, basée sur l'utilité des métadonnées pour retrouver ou découvrir l'information, et considèrent que le rôle accordé aux usagers dans la création des métadonnées doit aussi s'étendre à la gestion et à l'entretien. Le modèle « enrichir et filtrer » apparaît donc comme séparant le contenu des données, qui doit être sous le

signe de l'enrichissement, de l'interface par laquelle on y accède, qui doit être sous celui du filtrage ; comme s'attachant à l'à-propos de la description et de l'interconnexion du contenu des métadonnées plutôt qu'aux caractéristiques physiques du support de l'objet informationnel (même si ces derniers, en tant que métadonnées administratives, peuvent avoir de l'importance pour assurer la provenance ou l'authenticité des métadonnées) ; comme défendant « l'enrichissement et le filtrage comme un processus non-déterministe » (ibid., p. 101) par l'utilisation de vocabulaires *post-hoc* en perpétuelle évolution et non de taxonomies et de vocabulaires contrôlés décidés *a priori* et par l'observation plutôt que par l'anticipation des besoins des usagers.

3.2.3. Modeste proposition au secteur commercial du livre

Nous avons beaucoup emprunté, dans les lignes qui ont précédé, à une littérature émanant du monde des bibliothèques et des sciences de l'information. Qu'il s'agisse des possibilités d'enrichissement des métadonnées ouvertes par le paradigme des données liées ou de celles ouvertes par le web 2.0, les réflexions et les publications y sont nombreuses, et informées. Malgré les problèmes techniques, structurels, institutionnels, personne ne semble se résigner à la « métadaube » promise par Cory Doctorow. En regard, le secteur commercial du livre semble assez silencieux. Il nous semble qu'il y a là comme une forme de paradoxe. D'un côté, le monde des bibliothèques, traditionnellement attaché à une certaine « vérité bibliographique », ou à tout le moins à une pratique historiquement ancrée, sur la manière de cataloguer un ouvrage, fait preuve d'une grande ouverture intellectuelle sur les folksonomies, le *tagging* social, les évaluations ou les revues critiques des usagers, qui fleurissent dans un univers auquel ces bibliothèques ne sont pas directement connectées. De l'autre, le secteur commercial du livre, dont la pratique foisonnante et mouvante appelle un modèle de métadonnées qui soit capable d'évoluer souplement, et qui d'autre part possède le rapport aux clients qui permet de mettre en place le type de *feedbacks*, de collaboration, qui nourrissent l'intelligence collective du web 2.0, fait effort pour mettre en place des standards contraignants.

Nous avons vu que les métadonnées dans le secteur commercial pouvaient être divisées théoriquement (même si pratiquement elles sont transmises ensemble) en métadonnées bibliographiques et en métadonnées commerciales. Cette distinction entre les deux peut et doit nous semble-t-il être conservée. Les métadonnées commerciales, comme la disponibilité ou le taux de TVA, ne gagneraient pas nécessairement à être ouvertes, même si elles gagneraient à être liées. Les métadonnées bibliographiques, en revanche, gagneraient à être liées, ouvertes, et enrichies de métadonnées produites par les lecteurs, selon le modèle promu par Alemu et Stevens. Nous avons vu précédemment que les données liées ne sont pas

nécessairement ouvertes : les données commerciales pourraient être liées mais rester « fermées », quand les données bibliographiques seraient, elles, liées et ouvertes.

Il semble peu réaliste de mettre en place les données liées à travers toute l'interprofession : comme nous l'avons vu, le passage au format XML pour transmettre les métadonnées vient à peine d'avoir lieu, et ne semble pas devoir concerner avant longtemps l'ensemble des distributeurs ni des revendeurs. En revanche, toute l'interprofession gagnerait à ce que Dilicom adopte le paradigme des données liées et propose une plateforme permettant à tout un chacun de consulter les métadonnées bibliographiques, et de fournir ou de consulter des métadonnées sociales, et permettant aux partenaires professionnels (par négociation de contenu) d'accéder aux métadonnées commerciales – en plus des métadonnées bibliographiques – et à la possibilité de les télécharger. Nous pouvons imaginer également d'autres possibilités, par exemple de signaler – à condition que les libraires acceptent de faire part de leurs stocks et que l'usager du site accepte de donner sa localisation – quelles ont les librairies les plus proches disposant de l'ouvrage recherché, ce qui irait dans le sens d'initiatives déjà existantes comme Paris Librairies (des libraires ne disposant pas d'un ouvrage qui indiquent au client dans quelle librairie proche il peut le trouver). Un système de recommandation paraît en revanche peu envisageable au regard de l'exigence de neutralité de Dilicom.

Si nous pensons que l'interprofession gagnerait dans l'adoption de ce paradigme, nous pouvons nous demander si Dilicom y gagnerait. Quelques éléments de réponse peuvent être apportés par l'article de Tassilo Pellegrini *Semantic metadata in the publishing industry – technological achievements and economic implications*. L'auteur étudie sous l'angle du pilotage stratégique l'adoption des Données Liées par deux éditeurs, Wolters Kluwer et Reed Elsevier. Concernant l'impact technologique des métadonnées liées, l'auteur signale que « l'approche par les Données Liées offre des bénéfices significatifs, comparée aux pratiques conventionnelles de gestion des données. Ceux-ci sont, selon Auer[61] :

- Déréférençabilité. Les identifiants (URIs) ne sont pas seulement utilisés pour identifier les entités, mais puisqu'ils peuvent être utilisés comme des URLs, ils permettent aussi de localiser et récupérer des ressources décrivant et représentant ces entités sur le Web.
- Cohérence. Quand un triplet RDF contient des URIs de différents espaces de nommage en position de sujet et d'objet, ce triplet établit un lien entre une entité identifiée par le sujet (et décrite dans le jeu de données source utilisant espace de nommage A) avec l'entité identifiée par l'objet (décrite dans le jeu de donnée cible utilisant l'espace de nommage B). Par ces liens RDF typés, les items de données sont interconnectés efficacement et de façon cohérente.

[61] AUER, Sören. *Creating knowledge out of interlinked data*. In *Proceedings of WIMS'11*. 2011, pp. 1–8.

- Intégrabilité. Puisque toutes les sources de Données Liées partagent le modèle de données RDF, qui est basé sur un unique mécanisme pour représenter l'information, il est très simple d'arriver à une intégration syntaxique et une intégration sémantique simple de différents jeux de Données Liées. Une intégration sémantique de plus haut niveau peut être réalisée en employant un schéma et des techniques de concordance d'instances et d'exprimer les concordances trouvées à nouveau comme alignements de vocabulaires RDF et d'ontologies en termes de triplets additionnels.
- Actualité. Les Données liées peuvent facilement être publiées et mises à jour, ce qui facilite une disponibilité rapide. En outre, une fois qu'une source de Données Liées est mise à jour, on peut instantanément y accéder et l'utiliser, puisque le processus chronophage et source d'erreurs extraction-transformation-chargement n'est pas nécessaire » (39, Pellegrini, pp. 11-12).

Comparant dans la suite de son article les deux éditeurs, Tassilo Pellegrini observe que « Wolters Kluwer vise à nourrir un environnement business ouvert inspiré des principes de l'innovation libre [*open innovation*]. Ils ont commencé par expérimenter l'offre de certaines ressources avec une politique de licences doubles, stimulant ainsi une dynamique de communauté et des pratiques de business collaboratif […] À l'inverse, Reed Elsevier se base sur des mécanismes de contrôle strict dans la gouvernance du processus de création de la valeur. Ils utilisent un modèle de licence très strict, arguant que pour des raisons d'assurance de la qualité, ils doivent exercer un contrôle strict sur leur environnement business et les collaborateurs associés. Ils sont toutefois conscients des opportunités commerciales offertes par l'innovation libre, mais n'ont pas encore adopté cette culture » (ibid., p. 18). Malgré ces différences marquantes, notamment sur les aspects évidemment fondamentaux des licences portant sur les contenus et les métadonnées, « les Données liées servent dans les deux de couche d'intégration technologique et organisationnelle, affectant les flux et les pratiques de travail, et provoquant de nouveaux produits et de nouveaux modèles commerciaux » (ibid.).

Il n'est pas étonnant que ce soit ici des éditeurs qui aient franchi le pas. Ils sont en effet les premiers créateurs des métadonnées des ouvrages qu'ils éditent. Si l'angle adopté était ici celui du pilotage stratégique, nous pouvons le compléter du témoignage de Jean-Paul Jorda dans son article *L'enjeu du web de données pour les éditeurs*, à propos du « rôle d'enrichissement » de l'éditeur : « Le rôle d'un éditeur est de produire une version de référence de l'article, acceptée par les pairs, identifiée comme telle et, pour ce qui concerne EDP Sciences, de très grande qualité, quel que soit le modèle de publication (Open access ou pas). Cette version de référence, publiée sur le site de la revue, est incomparablement plus étoffée que les prépublications disponibles par ailleurs. Accessible sous plusieurs formats (PDF, EPUB et HTML), elle a donné lieu à un travail de composition réalisé par des professionnels et validé par les auteurs. Elle est profondément insérée dans le web des sciences, à travers les citations, les mots-clés, les auteurs, les données scientifiques, les réseaux sociaux. Nous enrichissons

ainsi le web de données et assurons la visibilité de l'article, de ses auteurs, de leurs institutions, des financeurs et de la revue en son sein.
Pour cela, nous devons identifier l'article (attribution d'un DOI) mais aussi, et avec leur aide, les auteurs, les financements et certaines données scientifiques. Nous devons également diffuser cet article et les données afférentes via des envois automatiques à plus de 30 partenaires dans différents formats, y compris pour des republications en Open Access. Ces données sont également mises à disposition sur notre plateforme à l'aide de divers protocoles et formats pour permettre le moissonnage et l'indexation par des tiers.
Enfin, à l'aide de partenaires, nous devons pérenniser sur le long terme la version électronique des articles. Ce travail est le résultat d'une chaîne de traitement humaine et technique complexe, en évolution permanente » (40, Jorda). Si Jean-Paul Jorda met ici l'accent sur le rôle d'enrichissement de l'éditeur, le fait de disposer d'un identifiant unique permet que l'échange de métadonnées ait lieu dans les deux sens et que chacun des acteurs puisse enrichir ses métadonnées de celles des autres. Ainsi, à mesure que la BnF adopte elle-même les données liées, l'adoption de ce paradigme par Dilicom permettrait d'enrichir les métadonnées du livre beaucoup plus facilement que ce n'est actuellement le cas, l'utilisation des notices de la BnF aux formats UNIMARC et INTERMARC obligeant à des *mappings* parfois complexes, en utilisant les identifiants *Archival Resource Key* (ARK) de la BnF.

Lorsque Souad Odeh et Ghislaine Chartron, dans l'article *Acteurs et économie des métadonnées du livre en France : analyse et avenir*, envisagent les convergences et les divergences possibles, dans l'optique d'une mutualisation des métadonnées de référence, la première divergence envisagée est précisément l'utilisation par les bibliothèques de métadonnées MARC quand le secteur commercial dispose de métadonnées au format ONIX. L'utilisation commune du standard RDF lèverait cette opposition, mais en soulèverait une deuxième les auteurs remarquant que certains acteurs, comme Électre, craindraient « une concurrence directe du secteur public, surtout sur son marché des bibliothèques » (19, Odeh et Chartron, p. 28). Il nous semble que le souhait de conserver des formats non interopérables pour éviter ce type de concurrence n'est pas une option tenable à moyen terme. Les auteurs notent par ailleurs que « des acteurs commerciaux comme Dilicom et TITE-LIVE sont au contraire des utilisateurs des métadonnées de la BnF surtout pour les autorités auteurs leur permettant d'assurer la qualité des métadonnées qu'ils produisent. L'ouverture sur le Web de données pourrait alors renforcer davantage la productivité de ces acteurs ». Nous en sommes persuadés. Odeh et Chartron envisagent à la fin de leur article une complémentarité dans la production de métadonnées, les éditeurs produisant des « métadonnées orientées œuvre » quand les bibliothèques produiraient des « métadonnées orientées autorité » et les réseaux sociaux des « métadonnées orientées recommandation ». Concernant les métadonnées relatives aux œuvres, en effet, les éditeurs « sont les mieux placés pour les produire dès la source. Les éditeurs français en sont conscients mais ils n'auront pas tous le savoir-faire nécessaire à la

production de métadonnées de qualité (complètes, correctes, mises à jour, formatées à la demande), leur partenariat avec des sociétés telles que Dilicom, Electre ou de nouveaux intermédiaires tels que GiantChair, Lekti, devrait s'affirmer » (ibid., p. 30). « Par ailleurs, les métadonnées produites par les éditeurs gagneront en qualité si elles se basent sur des référentiels et des vocabulaires d'autorité. Exposés en Linked Data et utilisés comme pivots dans des projets et applications, ils facilitent la recherche et la mise en relation des ressources » (ibid.), ce qui fait des bibliothèques comme la BnF les acteurs les plus crédibles dans la production de métadonnées d'autorité. Enfin, quant aux métadonnées issues des réseaux sociaux, elles permettent la recommandation, qui « s'avère une stratégie précieuse pour mieux capter l'attention des usagers dans une inflation d'offre de contenus. Les systèmes de recommandation bâtis sur des collectes de données variées se sont développés dans les secteurs du commerce en ligne, et tout particulièrement dans le secteur du livre, de la musique et des films… »[62]. Ce qui amène les auteurs à conclure que « la chaîne de valeur des métadonnées du livre en France devrait se répartir en trois types d'activités principales : activité de production assurées majoritairement par les éditeurs (en interne ou en sous-traitance), activités d'enrichissement effectuées par les bibliothèques (liens et autorités) et les réseaux sociaux numériques du livre (recommandation sociales d'experts et de lecteurs) et activités d'agrégation et de personnalisation réalisées par des intermédiaires spécialisés. La réalité économique et les défis du numérique devraient encourager l'ensemble des acteurs à développer une vision globale pour une gestion efficace du workflow des métadonnées du livre, mutualisant les efforts apportés par l'ensemble des professionnels » (ibid.). La vision que propose cette conclusion peut sembler un vœu pieux, cet effort de rationalisation supposant que les nombreux obstacles évoqués au cours de l'article soient levés et que les différents acteurs se coordonnent. Il nous semble que la place particulière de Dilicom dans la chaîne du livre et la base de collaboration sur les données d'autorité qu'elle a déjà avec la BnF placent cette entreprise dans une situation privilégiée pour être l'opératrice de cette mutualisation, à quelques conditions : adopter le standard RDF dans l'expression des métadonnées, se rapprocher des éditeurs et ouvrir les métadonnées du livre à celles produites par les réseaux sociaux.

Pour celles-ci, même si ce n'est pas à notre connaissance le cas actuellement, il nous semble que les éditeurs pourraient avoir une place et s'inspirer plus largement de démarches de marketing adoptées dans d'autres secteurs économiques. Ainsi, dans son article *Émergence des prospectivistes 2.0, le cas des planneurs stratégiques*, Maria Mercanti-Guérin explique que « les caractéristiques du Web 2.0 renforcent le travail collaboratif en assouplissant et simplifiant la production et diffusion de l'information. Ces nouvelles formes de relation sont rendues possibles par un certain nombre d'outils largement utilisés par les communautés virtuelles :

[62] Sur ce dernier point, voir KEMBELLEC Gérard, CHARTRON Ghislaine et SALEH Imad. *Les systèmes de recommandation*. Paris : Hermès, ISTE, 228p.

- Les outils de partage et de collaboration : bookmarks sociaux, réseaux sociaux, multimédia, syndication
- Les outils de diffusion : wikis, blogs et flux RSS
- Les outils de recherche et de collecte de l'information : blogs et flux RSS, tags, moteurs de recherche collaboratifs, bases de données, outils sémantiques
- Les outils d'étude : bulletin board et focus group online, web-reporting, géomarketing à travers les applications composites, analyse des blogs introspectifs
- Les outils de surveillance et de veille : détection et agrégation de flux RSS, utilisation d'outils type Digimind

La nouveauté de ces outils est contestée par bon nombre d'experts pour qui le Web 2.0 n'est qu'un enrichissement des interfaces utilisateur fondées sur les technologies éprouvées du Web 1.0. Néanmoins, sa fonction de partage des connaissances, d'identification des sources et d'évaluation des productions par les internautes eux-mêmes en fait un espace où le marketing et la communication sont omniprésents, ses fonctionnalités se mariant aux nouveaux concepts de ciblage et de segmentation : logique de CRM (*customer relationship management*), buzz marketing ou marketing du bouche à oreille, marketing de la valeur client, dialogue avec la marque, blogs corporate, communication interactive. De ce fait, la prospective marketing devient un champ privilégié de recherche sur le Web 2.0. » (41, Mercanti-Guérin). L'éditeur se ferait alors planneur stratégique, puisque selon l'auteur, ce dernier « enrichit sa réflexion prospective par l'utilisation de portails participatifs sur lesquels sont déposées des actualités proposées par les internautes. Ces catégories peuvent être divisées en thématique et bénéficient d'un modérateur. Ainsi le Digg Like consacré au Marketing, Marketingrama bénéficie d'un classement des articles les plus lus et les mieux notés (782 items à ce jour), des commentaires afférents et d'un étiquetage via un système de tags. L'étude des commentaires et des articles permet de mesurer en temps réel la popularité d'un produit mais également les souhaits non satisfaits en matière de consommation. A noter également l'émergence des cahiers blancs circulant librement sur les blogs. (...)

Une des dernières thématiques abordées par les blogs des planneurs stratégiques concerne le processus de création et son séquençage. Si la *copy strategy* est dans certaines agences rédigée sous forme de Wiki, les planneurs du Web 2.0 réclament de "briser la chaîne séquentielle du positionnement, brief, création pour remettre l'idée au cœur de la création" (Dumont in slideshare.net, 2007). La souplesse du Web 2.0 permet, comme nous l'avons vu précédemment, de tester presque en direct des scenarii et des concepts. Fondé sur le principe de la folksonomie ou peuplonomie, le Web 2.0 repose sur la subjectivité des internautes en termes de classification et de ce fait représente un indicateur des aspirations profondes des consommateurs et de leurs centres d'intérêt » (ibid.). Même si l'orientation de l'article est celle du marketing, nous ne sommes pas ici très éloignés des préconisations d'Alemu et Stevens.

Nous remarquions un certain « silence » du secteur commercial du livre sur ces questions. Celui-ci n'est toutefois que relatif, et peut-être les questionnements et les débats se situent ailleurs que sur le devant de la scène, comme en témoigne le document *Métadonnées, web sémantique : quels enjeux pour les professionnels du livre ?*, compte-rendu d'une table ronde lors des Assises du livre numérique, le 21 mars 2014[63]. Les acteurs de l'interprofession se sont emparés du sujet et semblent convaincus de sa pertinence, au moins en ce qui concerne le livre numérique, autour duquel les enjeux d'accessibilité sont plus forts, comme nous l'avons vu dans l'article de Souad Odeh et Ghislaine Chartron. En ce sens, de même que le livre numérique a été pionnier dans l'adoption du format ONIX, il pourrait l'être dans l'adoption du standard RDF et la mutualisation des métadonnées avec d'autres acteurs du secteur du livre, y compris non marchands.

Un dernier enjeu serait bien entendu que la plateforme dont nous avons esquissé le projet, présentant sur le web les métadonnées bibliographiques des ouvrages commercialisés, soit bien référencée, et que rechercher des informations sur un livre la fasse apparaître dès la première page. De ce point de vue aussi, la structuration sémantique des métadonnées a un rôle à jouer. Dans son article *Que voit réellement Google de la sémantique des pages web ?*, Gérald Kembellec rappelle les deux principales catégories parmi les méthodes d'exposition des contenus, Plein Old Semantic HTML et triplets RDF, avant de rappeler que « sur son blog officiel, [Google] présente depuis 2012 l'idée que le moteur recense des concepts et non des chaînes de caractère », pour conclure ainsi : «si la structuration des données et métadonnées des documents par la sémantique n'est pas l'essentiel de l'indexation par Google, elle devient une part importante de la stratégie Search Engine Optimisation (SEO) encouragée par la firme » (42, Kembellec).

[63] Disponible sur le site du SNE : <https://www.sne.fr/app/uploads/2017/11/Metadonnees_web-semantique_Assises_mars14.pdf> (Consulté le 20 août 2018)

3.3. Pistes vers des données liées dans le secteur commercial du livre

3.3.1. A quoi ressemblerait une ontologie du secteur commercial du livre ?

Nous n'avons pas dans le cadre de ce travail l'ambition de proposer une ontologie pour le secteur commercial du livre. Nous voulons seulement esquisser quelques pistes dans ce sens, en notant de façon préliminaire que si nous considérons qu'à bien des égards, le secteur commercial du livre gagnerait à s'inspirer des réflexions et du travail accompli par les bibliothèques en matière de données liées, une ontologie au moins en partie « originale » est nécessaire dans le secteur commercial, qui rencontre des problématiques auxquelles les bibliothèques n'ont pas à faire face. D'autre part, une telle ontologie devrait s'appuyer, ou s'inscrire dans la continuité, du standard ONIX, afin de rendre plus aisé le passage aux données liées.

Une première piste nous est fournie par l'article de Gautier Poupeau intitulé *Les éditeurs et les métadonnées : ONIX* (qui date du 6 avril 2006), sur son blog *Les petites cases*. C'est la deuxième partie de l'article qui retient notre attention, lorsque Gautier Poupeau nous dit : « pour finir, je dirais que ONIX aurait intérêt à passer à RDF ce qui permettrait une meilleure formalisation et de reprendre des bouts d'autres vocabulaires existants, plutôt que de réinventer la roue. […] Voyons maintenant concrètement ce qui peut nous intéresser dans cette norme […] Sur les 26 parties qui composent ONIX, 11 sont intéressantes pour les publications électroniques sur le Web en accès libre et gratuit. Évidemment, nous n'avons pas besoin des informations relatives à la gestion des stocks, à la description physique du produit… Voici donc ce que j'ai pu relever ». Il énumère ensuite onze points qui lui semblent intéressants. Nous les citons ci-dessous, mais les évolutions du format depuis 2006 rendent certaines de ses remarques caduques :

- « Le **format du produit** » : nous ne retrouvons pas les Code Lists actuelles ce que Gautier Poupeau en dit.
- « Le **groupe de titre**, qui permet d'indiquer les renseignements sur le titre, sous-titre, titre original dans le cas d'une traduction, abréviation dans le cas d'un titre long… ».
- « Le **groupe de description d'une publication numérique** (comme quoi les éditeurs vont finir par s'y mettre, ne désespérons pas…) » : là aussi, le paysage éditorial a changé depuis la rédaction de l'article.
- « Le **groupe de description du site Web** du produit est prévu normalement pour indiquer un lien vers un site Web promotionnel. Ils n'ont pas prévu à cet endroit que le site Web soit lui-même la publication ce qui en dit long sur les pensées des éditeurs traditionnels sur ce vecteur de diffusion » : l'auteur aurait peut-être trouvé de quoi le satisfaire dans les 45 valeurs actuelles de la liste de codes n°73 intitulée « website role » — nous pensons notamment à la valeur 29, « Web page for full content », décrite ainsi : « Utilisez cette valeur dans le composé <Website> […] quand vous transmettez un lien vers une page web à

laquelle un produit numérique est disponible pour téléchargement ou accès en ligne » (26, EDItEUR).

- « Les **informations sur la/les langues utilisées**. Pas grand-chose à dire, si ce n'est qu'il faut utiliser la norme ISO 693-2/B » : c'est bien le cas.
- « Les **sujets du produit**. Vous pouvez indiquer les sujets selon la taxonomie qui vous fait plaisir en indiquant laquelle vous utilisez »
- « Les **groupes de publics** permet d'indiquer selon une classification précise les publics concernés par le produit »
- « Les **différents textes d'accompagnement**, ce qui recoupe quatrième de couverture, extrait de critiques, résumé… »
- Le **groupe d'information relatives à la collection** dans laquelle prend place le produit permet de renseigner le nom de la collection et le numéro du produit dans la collection. »
- « Le **groupe de renseignements sur les "contributeurs"**, c'est exactement le même système que Contributor dans Dublin Core. Il permet donc de spécifier précisément le rôle du "contributeur", son nom et les renseignements qui peuvent le concerner (biographie, affiliation, position professionnelle…) »
- « Le **groupe de renseignements sur l'éditeur** (au sens de *publisher*) et la publication permet de renseigner le nom de l'éditeur, la ville de publication et la date de publication ». (43, Poupeau)

Remarquons en premier lieu que ce sont les métadonnées descriptives bibliographiques, comme il le précise d'emblée, auxquelles s'attache l'auteur (qui vient du monde des archives). Ce sont effectivement les métadonnées qui sont le plus susceptibles d'intéresser d'autres professionnels du livre. Dans l'optique d'une ouverture du secteur commercial du livre aux données liées ouvertes (*Linked Open Data* - LOD), il peut être tout à fait envisageable de prévoir des ouvertures « à géométrie variable » pour les métadonnées, les métadonnées commerciales les plus sensibles étant réservées aux partenaires commerciaux auxquelles elles étaient traditionnellement adressées, et les métadonnées bibliographiques étant elles totalement ouvertes. En deuxième lieu, nous constatons que des points de convergence doivent pouvoir être trouvés entre ONIX et des ontologies existantes, ne serait-ce que sur la propriété "Contributeur".

Nous ne sommes toutefois pas beaucoup plus avancés sur la constitution d'une ontologie pour le secteur commercial du livre. Dans le blog de Mondeca[64] intitulé *Leçons de choses,* Bernard Vatant offre dans le billet « Vous lisez quoi cet été » quelques réflexions autour du livre, partant de la question banale « Tu as lu ça ? » pour se demander comment il est possible de désigner de façon non ambiguë le « ça » dont il est question : « C'est toute la richesse et la flexibilité du langage et de l'esprit humain de pouvoir gérer instantanément toutes ces subtilités et l'ambiguïté

[64] Mondeca est un éditeur de logiciels. Les données liées et la gestion d'ontologies sont au cœur de son activité.

fondamentale du référent de la question (le livre), de pouvoir passer sans dommage de l'objet physique désigné à mon regard à des représentations mentales plus larges. Comment traduire cette flexibilité dans des langages formels, des vocabulaires qui nous permettent de communiquer sans ambigüité avec les machines ? Cette traduction nécessite des objets clairement identifiés, typés et décrits qui peuvent expliciter le « ça » de différentes façons. Pour ce faire, les bibliothécaires ont mis au point un cadre de représentation assez élaboré nommé FRBR[65], qui distingue quatre niveaux de représentation du livre (Œuvre, Expression, Manifestation et Item). Découpage intéressant et utile, mais dont le nombre de niveaux et leur distinction parait finalement assez arbitraire, et dont la terminologie peut paraître déroutante a priori, même pour les documentalistes. BIBFRAME dont nous avons déjà parlé réduira d'ailleurs le nombre de niveaux à trois seulement, pour simplifier, mais la démarche générale de FRBR est intéressante et peut sans doute être utilisée dans d'autres domaines […].

Le modèle FRBR est-il suffisant ou même pertinent quand je considère le livre dans un tout autre contexte que la classification et la recherche documentaire comme la vente en ligne ? Un livre sur un site marchand, c'est un produit, c'est-à-dire une référence et une description, je l'achète comme n'importe quel autre produit en ligne. Je ne sais pas quel exemplaire m'arrivera par la poste. Du point de vue FRBR, ce que j'achète serait à la rigueur au niveau de la Manifestation, identifiée par un ISBN par exemple. Mais le vendeur stocke et me propose des données qui ne sont pas du tout de l'ordre de FRBR : nombre d'exemplaires en stock, état neuf ou occasion, pourcentage de remise, délai d'acheminement, options de livraison, recommandations (ceux qui ont acheté ce livre ont aussi consulté …), et une fois expédié, c'est un exemplaire (Item) qui est géré par le service de suivi, avec un transporteur, un numéro de colis, une date de livraison prévue … Dans ce contexte, le livre sera sans doute mieux décrit par un vocabulaire comme GoodRelations.

Un autre schéma de représentation serait nécessaire pour décrire la chaîne de fabrication du livre, un autre pour le processus de sélection chez un éditeur (ce manuscrit non encore publié est un livre en puissance, mais est-ce déjà un livre), un autre encore pour le processus de restauration s'il s'agit d'un livre ancien et précieux. On le voit, il ne peut exister de représentation unique du livre, chaque contexte d'information utilise son mode de représentation et d'identification. Vouloir les unifier dans une représentation unique de ce qu'est un livre est un exercice de compromis difficile, entre les deux extrêmes de l'accumulation (la réunion de toutes les représentations risque d'être incohérente) et du plus petit modèle commun (l'intersection de toutes les représentations risque d'être vide). C'est pourquoi on peut se poser la question du périmètre d'utilisation d'une représentation à tout faire comme http://schema.org/Book qui à la fois ne capture pas tous les aspects du concept, et se

[65] FRBR, pour *Functional Requirements for Bibliographic Records* (Spécifications fonctionnelles des notices bibliographiques), propose une modélisation des documents, personnes, sujets, ainsi que de leurs attributs et de leurs relations. NdA

trouve encombrée par héritage de son parent http://schema.org/CreativeWork de propriétés dont la pertinence dans certains des contextes évoqués plus haut est plus que douteuse.
Ce qui est indispensable par contre ce sont des interfaces de traduction entre la représentation de la bibliothèque et celle du site marchand, entre le site marchand et son fournisseur, entre la bibliothèque et l'atelier de restauration, et à ces interfaces le partage d'identifiants et de vocabulaires minimaux communs. Une architecture composée de représentations pertinentes dans un contexte, et d'interfaces explicitant les règles de traduction d'un contexte à l'autre, est certainement l'avenir de l'architecture des systèmes d'information sémantique. D'où l'importance de la gouvernance des biens communs que représentent des vocabulaires variés et interconnectés, écosystème que nous avons commencé à décrire dans le projet *Linked Open Vocabularies*, et auquel le W3C s'intéresse de près avec sa nouvelle proposition de *Vocabulary Services* » (44, Vatant).

Effectivement, le secteur commercial des livres n'a à faire qu'aux items ou aux manifestations du modèle FRBR, comme le dit très bien Bernard Vatant. Néanmoins, dans une optique de mise en relation d'une manifestation avec d'autres, par exemple une édition d'une traduction avec une autre édition de la même traduction, ou avec d'autres expressions, par exemple des traductions dans d'autres langues de la même œuvre traduite ou des films tirés du livre original, il peut être utile de « remonter » au niveau de l'expression ou à celui de l'œuvre. En ce sens, le rapprochement avec les bibliothèques est évidemment souhaitable, d'autant que la « transition bibliographique » dans laquelle sont engagées les bibliothèques depuis 2015 a pour objectif affiché d'aller vers le web de données. Même si les bibliothèques et l'industrie du livre ont des tempos différents, logique patrimoniale et logique commerciale obéissant à des exigences bien différentes, un accord « ontologique » entre les deux au niveau bibliographique, soit par une ontologie commune, soit par l'anticipation *by design* d'un interfaçage à venir, serait éminemment profitable aux deux parties. Quant à GoodRelations, cette ontologie se présente explicitement comme orientée vers le e-commerce. Peut-être faudrait-il alors se tourner vers des ontologies dédiées aux chaînes d'approvisionnement. Nous conservons néanmoins l'idée tout à fait acceptable d'une hybridation d'ontologies, les unes traitant des aspects bibliographiques du livre et les autres traitant des aspects commerciaux et de la chaîne d'approvisionnement.

3.3.2. A quoi ressemblerait une notice de livre conçue selon RDF ?

Afin de savoir à quoi ressemblerait une notice du secteur commercial du livre selon RDF, nous pouvons nous appuyer sur une notice RDF de la BnF pour un ouvrage actuellement disponible sur les rayons des librairies. Il suffit pour cela :

- D'aller sur la page web du SPARQL Endpoint de la Bnf : http://data.bnf.fr/sparql/

- De rechercher par une requête SPARQL très simple les ouvrages qui indiquent un *European Article Numbering* sur 13 chiffres[66] (EAN13), qui est l'identifiant d'un ouvrage dans le secteur commercial (nous nous sommes limités à 10000 ressources pour ne pas avoir un fichier trop lourd) :

```
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
SELECT ?s ?o
WHERE {
  ?s bnf-onto:ean ?o.
}
LIMIT 10000
```

- De croiser avec un catalogue de distributeur et de choisir un ouvrage commun aux deux listes
- De préciser notre requête pour n'obtenir que la notice BnF correspondant à cet EAN :

```
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
SELECT ?s
WHERE {
  ?s bnf-onto:ean "9782258076624".
}
```

- De récupérer la notice (en spécifiant « Auto » comme forme du résultat dans l'interface, elle est automatiquement téléchargée)

La notice obtenue est au format RDF/XML. Nous l'appelons par facilité de langage une notice, mais il s'agit plutôt de la description d'une ressource qui est une édition de l'œuvre. Pour plus de lisibilité, nous l'avons traduite en N3 sur un traducteur RDF en ligne. Elle se divise en quatre blocs.

```
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/> .
@prefix bnfroles: <http://data.bnf.fr/vocabulary/roles/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix marcrel: <http://id.loc.gov/vocabulary/relators/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdagr1: <http://rdvocab.info/Elements/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdrel: <http://rdvocab.info/RDARelationshipsWEMI/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

[66] L'EAN13 n'est autre que le code à barres. Il reprend, sans les tirets, la suite de chiffres de l'ISBN13, souvent plus connu hors du secteur commercial.

Dans ce premier bloc sont déclarés les préfixes qui sont utilisés par la suite. Ces préfixes sont des abréviations, qui permettent par exemple de noter une propriété « dctermes:created » plutôt que http://purl.org/dc/terms#created.

```
http://data.bnf.fr/ark:/12148/cb41181815x a <http://www.w3.org/2004/02/skos/core#Concept> ;
   dcterms:created "2008-01-10" ;
   dcterms:modified "2008-02-06" ;
   foaf:focus <http://data.bnf.fr/ark:/12148/cb41181815x#about> .
```

Ce deuxième bloc qualifie la notice, désignée par son URI. Cet URI intervient dans quatre triplets : celui qui définit la description de la ressource comme un « concept » au sens du vocabulaire SKOS (*Simple Knowledge Organization System*, vocabulaire permettant de définir des systèmes de classification), celui qui donne sa date de création, celui qui donne sa date de dernière modification, et celui qui donne son « focus » au sens du vocabulaire *foaf*, c'est-à-dire l'entité associée au concept.

```
<http://data.bnf.fr/ark:/12148/cb41181815x#Expression> a
                                        <http://rdvocab.info/uri/schema/FRBRentitiesRDA/Expression> ;
   bnfroles:r550 <http://data.bnf.fr/ark:/12148/cb119048528#about> ;
   bnfroles:r70 <http://data.bnf.fr/ark:/12148/cb11902711x#about> ;
   marcrel:aui <http://data.bnf.fr/ark:/12148/cb119048528#about> ;
   marcrel:aut <http://data.bnf.fr/ark:/12148/cb11902711x#about> ;
   dcterms:contributor <http://data.bnf.fr/ark:/12148/cb11902711x#about>,
      <http://data.bnf.fr/ark:/12148/cb119048528#about> ;
   dcterms:language <http://id.loc.gov/vocabulary/iso639-2/fre> ;
   dcterms:type <http://purl.org/dc/dcmitype/Text> ;
   = <http://data.bnf.fr/ark:/12148/cb41181815x#frbr:Expression> .
```

Le troisième bloc se situe au niveau de l'expression de l'œuvre. Nous sommes ici dans le modèle FRBR, qui définit une hiérarchie — prenons l'exemple d'un livre — de l'exemplaire que nous pouvons tenir en main (item) à la publication à laquelle il appartient (manifestation), à son contenu intellectuel ou artistique (expression), jusqu'à la création abstraite à laquelle se rattache ce contenu (œuvre). A ce niveau peuvent être définis certains éléments qui resteront vrais quelles que soient les « manifestations » : l'auteur et l'auteur de l'introduction au sens du vocabulaire MARC, qui sont des contributeurs au sens du vocabulaire Dublin Core, qui ont les rôles d'auteur et de préfacier au sens du vocabulaire bnfroles. Leurs noms n'apparaissent pas, mais seulement leurs identifiants ARK.

```
<http://data.bnf.fr/ark:/12148/cb41181815x#about> a
                                        <http://rdvocab.info/uri/schema/FRBRentitiesRDA/Manifestation> ;
   bnf-onto:FRBNF 41181815 ;
   bnf-onto:ean "9782258076624" ;
   bnf-onto:firstYear 2007 ;
   bnf-onto:isbn "978-2-258-07662-4" ;
```

```
    dcterms:date "2007" ;
    dcterms:description "1 vol. (IX-977 p.)" ;
    dcterms:publisher "Paris : Omnibus , impr. 2007" ;
    dcterms:title "Pièces courtes, monologues, vaudevilles et comédies" ;
    bibo:isbn13 "978-2-258-07662-4" ;
    rdagr1:dateOfPublicationManifestation <http://data.bnf.fr/date/2007/> ;
    rdagr1:note "Note : Réunit : \"Amour et piano\" ; \"Gibier de potence\" ; \"Fiancés en herbe\" ; \"Un
bain de ménage\" ; \"Notre futur\" ; \"Les pavés de l'ours\" ; \"Dormez, je le veux\" ; \"Léonie est en
avance\" ; \"Hortense a dit : je m'en fous\" ; \"L'homme de paille\" ; \"J'ai mal aux dents\" ; \"Trop
vieux\" ; \"Patte en l'air\" ; \"Le petit ménage\" ; \"Le juré\" ; \"Les fiancés de Loches\" ; \"Le mariage de
Barillon\" ; \"Champignol malgré lui\" ; \"Le système Ribadier\" ; \"L'âge d'or\"" ;
    rdagr1:placeOfPublication "Paris" ;
    rdagr1:publishersName "Omnibus" ;
    rdrel:expressionManifested <http://data.bnf.fr/ark:/12148/cb41181815x#Expression> ;
    rdfs:seeAlso <http://catalogue.bnf.fr/ark:/12148/cb41181815x> .
```

Ce quatrième bloc se situe au niveau de la manifestation, qui est la « chose » que décrit la notice, à propos de laquelle elle est formulée (c'est le sens du « #about » ajouté à l'identifiant ARK de la notice). C'est ici que nous trouvons dans les triplets, comme objets, des valeurs littérales, comme l'identifiant FRBNF de l'ouvrage, son EAN, son année de publication, son titre, mais aussi l'URL de la notice bibliographique de l'ouvrage dans le catalogue BnF.

La description de cette notice peut sembler complexe, et à bien des égards elle l'est, notamment en raison de l'usage du modèle FRBR, dont nous avons vu plus haut qu'il n'était pas nécessairement adapté au secteur commercial. La complexité du format ONIX, qui peut aller jusqu'à six niveaux de profondeur dans le XML, n'est pas à négliger non plus. L'intérêt de RDF est que la complexité de la structure conceptuelle est déplacée dans les ontologies, mais disparaît de la structure de la notice elle-même, qui ne contient que des triplets, ce qui facilite à la fois le stockage et les requêtes.

3.3.3. Publier des données liées

Notre ambition est ici de proposer une manière de feuille de route pour la publication de données liées, ce qui nous permettra d'offrir une vision plus concrète de la mise en œuvre de ce paradigme. Nous pouvons dans un premier temps rappeler une célèbre note de Tim Berners-Lee dans laquelle il formule, en 2006, quatre règles pour faire que le web s'étende : « comme le web de l'hypertexte, le web des données est avec des documents sur le web. Cependant, contrairement au web de l'hypertexte, dans lequel les liens sont des ancres de relations [*relationships anchors*] dans des documents hypertextes écrits en HTML, pour les données ils relient des choses arbitraires décrites par RDF. Les URIs identifient toutes sortes d'objets ou de concepts. Mais que ce soit pour HTML ou pour RDF, les mêmes exigences s'appliquent pour que le web croisse :

1. Utilisez des URIs comme noms pour les choses
2. Utilisez des URIs http afin que les gens puissent chercher ces noms
3. Quand quelqu'un recherche une URI, fournissez de l'information utile, en utilisant les standards (RDF*, SPARQL)
4. Incluez des liens à d'autres URIs, afin qu'ils puissent découvrir plus de choses »[67] (45, Berners-Lee)

Nous sommes ici dans un discours qui tient à la fois de la méthode et des bonnes pratiques. De notre point de vue actuel, quelques années plus tard, ces deux directions se sont développées. Une page sur le site du Word Wide Web Consortium est consacrée aux « bonnes pratiques pour publier des données liées »[68], qui explique par exemple ce que sont des « bonnes URIs », donne des recommandations sur l'usage des vocabulaires ou sur les manières de donner accès aux données une fois qu'elles sont liées. D'autre part, un livre comme celui de Tom Heath et Christian Bizer, datant de 2011, qui explique en détail comment publier des données liées afin de « faire évoluer le Web en un Espace de Données Global » (46, Heath et Bizer), auquel nous reprenons le schéma ci-dessous présentant différents cas de figures selon le type de données qu'on souhaite lier et la manière dont on souhaite les publier.

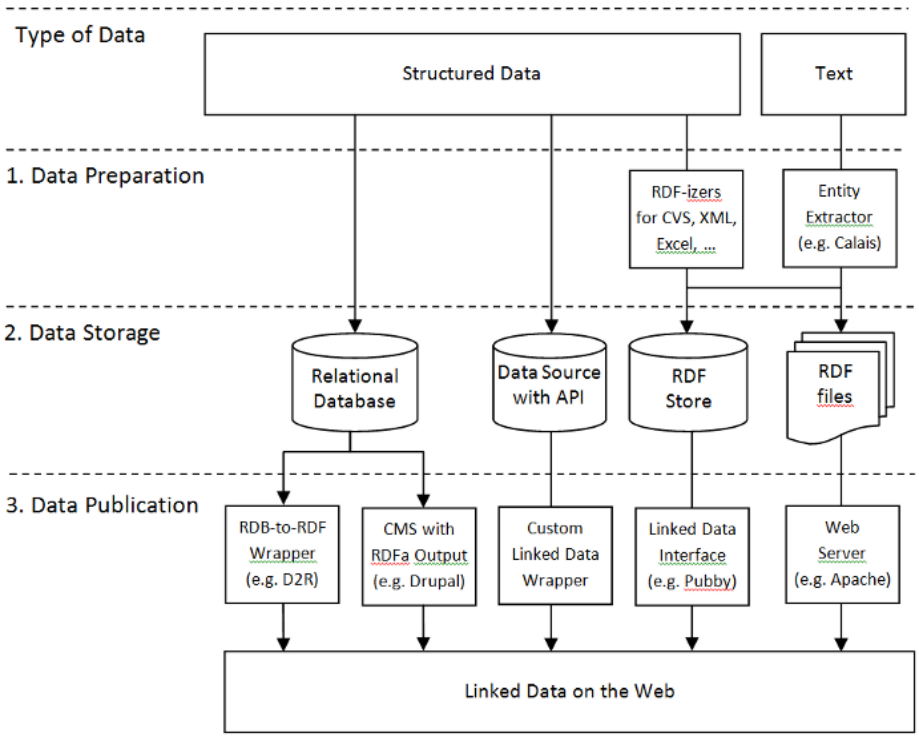

Figure 9 : Options et workflows de publication de Données Liées (C. Bizer et T. Heath, 2011)

Dans l'optique qui est la nôtre, nous considèrerons que les données que nous aurions à lier sont les métadonnées présentes dans les notices en ONIX, soit des données semi-structurées en XML (traduites éventuellement en JSON). Il s'agirait

[67] Nous traduisons.
[68] W3C. *Best Practices for Publishing Linked Data*. Disponible sur : < https://www.w3.org/TR/ld-bp/> (Consulté le 20 août 2018)

toutefois dans un premier temps de revenir sur la forme sous laquelle les données se présentent, de les analyser pour préparer la conception de la forme générale des URI à utiliser. Un deuxième temps serait celui de la modélisation des données amenant à la création d'une ontologie permettant de représenter le domaine des données. Notre expérience à Dilicom nous laisse penser que cette étape présenterait une grande complexité, en raison de la diversité du secteur, que nous avons à plusieurs reprises pointée, et des transformations qui sont en cours dans le domaine de l'édition, à mesure que les éditeurs individuels auto-distribués et les MarketPlaces se développent. Une ontologie du secteur devrait assurément prendre en compte ce type d'évolution. D'autre part, il est possible pour élaborer une telle ontologie de s'appuyer sur des ontologies existantes, comme nous l'avons vu en 3.3.1. Une troisième étape serait la production de données en RDF, en accord avec l'ontologie mise en place, à partir des données existantes. Cette étape supposerait donc un *mapping* à partir des notices en XML, par exemple en recourant à un langage de type XSLT (*eXtensible Stylesheet Language Transformation*). La liaison proprement dite interviendrait en quatrième lieu, en utilisant des propriétés comme owl:sameAs ou rdfs:seeAlso (que nous avons déjà rencontré dans la notice RDF de la BnF pour lier celle-ci à la notice du catalogue BnF « traditionnel »), après avoir identifié soigneusement les jeux de données propres à être liés. Viendrait ensuite l'étape du stockage des notices en RDF, puis celle de leur publication, supposant que le mode de stockage choisi soit connecté au web. L'accès permis par la publication peut être double, à la fois pour les machines (c'est un minimum) et pour les humains (par exemple par l'intermédiaire d'un SPARQ Endpoint). Cette dernière étape suppose que les jeux de données publiés soient accompagnés de métadonnées, afin de faciliter leur capacité à être réutiliser et éventuellement signaler leur provenance, permettre leur traçabilité. L'étape de la publication doit aussi s'accompagner d'une étape de « publicité », au sens où il ne suffit pas que les données soient accessibles : encore faut-il que les gens qui seraient susceptibles d'utiliser ces données ou y seraient intéressés le sachent. L'ajout d'un fichier sitemap.xml au site web où les jeux de données sont disponibles, par exemple, est de nature à favoriser cette démarche. Enfin, il nous semble que toute la démarche décrite ci-dessus devrait être articulée avec une démarche permettant et favorisant l'indexation collaborative, le *tagging* social, et tous les autres outils du web 2.0 permettant d'enrichir les notices par les usagers. Une formalisation en sept étapes est proposée par Cécilia Fabry et ses coauteurs dans l'article *Publier des données liées et ouvertes en sept étapes*, qui insiste en premier lieu sur l'importance d'une étape que nous n'avons pas abordée mais qui est effectivement préliminaire à tout processus de publication : le choix de la licence. Comme le rappellent les auteurs, « les données publiées sur le Web nécessitent de clarifier les conditions et possibilités juridiques de leurs réutilisations. Il s'agit de garantir et sérier la sécurité juridique en déclarant une licence reconnue d'utilisation des jeux de données exposés » (47, Fabry *et al.*). Dans le cas qui nous occupe, cette question devrait évidemment faire l'enjeu d'une clarification au sein des instances interprofessionnelles.

Une manière beaucoup plus légère d'introduire dans le site web tel qu'il existe actuellement — donc en HTML — des élément sémantiques, peut être d'utiliser RDFa (*Resource Description Framework in attributes*), qui s'appuie sur la syntaxe HTML existante de la page web et ajoute des attributs qui donnent des informations lisibles par les machines. Thomas Francart décrit dans l'article *Des textes augmentés avec les données du Web* comment « les contenus textuels peuvent bénéficier de méthodes d'enrichissement dans les hypertextes depuis des bases de connaissances sémantisées » (48, Francart) en prenant comme exemple deux pages Web ainsi modifiées qui sont présentes sur son site[69] : « dans les deux articles, le texte a été annoté manuellement en marquant en HTML les fragments de texte correspondant à des entités (personnes, lieux, organisations, concepts de thésaurus) décrites dans les données. Ce marquage est réalisé à l'aide de balises « span » utilisant la syntaxe RDFa, et c'est l'URI de l'entité dans la source de données qui est utilisée comme annotation.

Au moment de l'affichage de la page, du code Javascript analyse le texte, récupère les URIs des annotations et interroge les données en SPARQL pour proposer les enrichissements. En particulier, les informations de latitude et de longitude des lieux sont extraites pour afficher la carte, et le titre, l'image et la description sont récupérés à la volée pour afficher l'encart d'information. Par ailleurs, dans l'exemple basé sur Isidore, on liste également les articles du même auteur. Tous les traitements sont réalisés dans le navigateur ; il n'y a pas de scripts sur le serveur » (ibid.). Le type d'encart ici décrit pourrait tout à fait trouver sa place sur le site de Dilicom, par exemple pour afficher les métadonnées relatives à un ouvrage en interrogeant à la volée de SPARQL Endpoint de la BnF — plutôt que de les afficher dans un onglet spécifique lors de la consultation par les professionnels du FEL sur le site de Dilicom, comme c'est le cas actuellement, ce qui suppose la récupération et l'intégration de ces données.

[69] Disponible sur : <http://labs.sparna.fr> (consulté le 20 août 2018)

Conclusion

Au cours de cette partie, nous avons pu montrer que ce qui apparaissait comme un horizon souhaitable de la standardisation des métadonnées, le paradigme des données liées, et ce qui apparaissait comme une source formidable d'enrichissement des métadonnées, des métadonnées sociales, présentent une convergence possible, à la fois théorique, par la notion de redocumentarisation, et pratique, les métadonnées sociales s'appuyant sur la structuration des métadonnées liées. Cela suppose toutefois un réaménagement de la conception d'un standard pour prendre en compte les besoins des usagers, notamment en termes de qualité des données et de « vérité » bibliographique, la collaboration d'experts d'un domaine ou la « sagesse des foules » n'étant pas forcément moins valable que l'expertise des catalogueurs de bibliothèque ou des éditeurs. Dès lors que ces possibilités d'enrichissement sont ouvertes, l'abondance informationnelle semble redoublée, comme nous l'avons craint dès notre introduction. Pourtant, nous avons pu voir que cette abondance informationnelle existe depuis longtemps, et qu'elle a seulement vu au cours du temps augmenter le périmètre de ceux qu'elle touche. Pour reprendre le titre d'une *keynote* de Clay Shirky, « Il n'y a pas de surcharge informationnelle. Il y a un manque de filtrage »[70]. Le modèle « enrichir et filtrer » d'Alemu et Stevens spécifie le type de filtrage pertinent pour le monde de bibliothèques, mais il nous a paru qu'il pouvait l'être également dans le monde du livre. Ce qui nous a amené à formuler une proposition pour le secteur commercial du livre, y compris dans ses interactions avec le secteur non marchand du livre, séparant des métadonnées bibliographiques liées et ouvertes, et des métadonnées commerciales liées également mais fermées, pour envisager finalement les linéaments de la mise en œuvre de cette proposition, dans laquelle Dilicom est appelée à jouer un rôle central.

[70] SHIRKY, Clay. *It's not information overload – It's filter failure*. [En ligne]. Web 2.0 Expo NY. 2008. Disponible sur : < https://www.youtube.com/watch?v=LabqeJEOQyI> (consulté le 20 août 2018)

Conclusion

En débutant ce travail, nous avons convoqué l'économie de l'attention pour dessiner ce qui nous apparaissait comme un cercle vicieux : l'abondance informationnelle entraîne la pénurie d'attention, qui pousse les acteurs du monde du livre à une recherche de visibilité, d'accessibilité, de « découvrabilité » des ouvrages qu'ils promeuvent, ce qu'ils cherchent à atteindre en enrichissant les métadonnées desdits ouvrages, ce qui ne fait qu'accroître l'abondance informationnelle. Nous avons tenté de montrer que ce cercle vicieux pouvait être rompu par la manière dont les métadonnées étaient produites et structurées, ce qui permettait de ne plus craindre d'accroître encore l'abondance de métadonnées, en mettant en place des interfaces de filtrage adéquates. Est-ce à dire que dans ce cas, l'abondance informationnelle n'entraînerait plus la raréfaction de l'attention ? Oui, dans la mesure où la contextualisation, la personnalisation de ces interfaces, l'usage de vocabulaires plus proches des besoins de l'usager, lui masquerait cette abondance et lui permettrait d'accéder plus aisément à ce qu'il recherche. Mais que recherche-t-il ? Au-delà d'un premier niveau de réponse : des produits culturels, il nous semble que ce qui est en jeu ici est la question de savoir comment il est possible de se forger une culture dans un monde d'information abondante. La question est d'importance et appellerait une étude en elle-même. Nous n'entendons ici poser que quelques éléments d'analyse, à partir de nos réflexions sur les métadonnées.

Nous devons d'abord préciser ce que nous entendons par « se forger une culture ». Nous avons placé en exergue de ce travail une citation de Goethe dans laquelle il explique qu'il attend, de ce qui l'instruit, que cela augmente ou anime son activité. Nietzsche place cette citation en tête de sa *Seconde considération intempestive*, sous-titrée : « De l'utilité et de l'inconvénient des études historiques pour la vie » (49, Nietzsche). Il y décrit une situation d'abondance de travaux historiques dans l'Allemagne de son époque qui peut être facilement mise en parallèle avec la situation d'abondance informationnelle qui nous occupe. Dans la conclusion de son ouvrage, il fait l'éloge d'une conception de la culture issue de l'antiquité grecque : « Les Grecs apprirent peu à peu à *organiser le Chaos*, en se souvenant, conformément à la doctrine delphique[71], d'eux-mêmes, c'est-à-dire de leurs besoins véritables, en laissant dépérir les besoins apparents. C'est ainsi qu'ils rentrèrent en possession d'eux-mêmes. […] Ceci est une parabole pour chacun de nous. Il faut que chacun organise le chaos qui est en lui, en faisant un retour sur lui-même pour se rappeler ses véritables besoins. Sa loyauté, son caractère sérieux et véridique s'opposeront à ce que l'on se contente de répéter, de réapprendre et d'imiter. Il apprendra alors à comprendre que la culture peut être autre chose encore que la

[71] Ce que Nietzsche appelle la « doctrine delphique » est la maxime, apparaissant à plusieurs reprises dans les Dialogues de Platon, « Connais-toi toi-même », qui aurait été inscrite à l'entrée du temple de Delphes.

décoration de la vie […] Ainsi se révélera à ses yeux la conception grecque de la culture — en opposition à la culture romaine — la conception de la culture, comme d'une nouvelle nature, d'une nature améliorée, sans intérieur et extérieur, sans simulation et sans convention, de la culture comme d'une harmonie entre la vie et la pensée, l'apparence et la volonté » (ibid., pp. 178-179). C'est de cette culture, qui ne peut être confondue avec l'accumulation érudite de savoirs, mais concerne l'être dans son ensemble, à la fois pensant et agissant, que nous voulons parler. Notons la correspondance entre l'appel de Nietzsche à « se souvenir de ses vrais besoins » et le modèle « *enrich and filter* » d'Alemu et Stevens, qui préconise l'enracinement du filtrage dans les besoins des usagers.

Sur un ton plus apaisé que celui de Nietzsche, il nous semble que le philosophe de l'éducation Olivier Reboul porte un message très similaire lorsqu'il explique que : « La culture n'est pourtant pas le savoir... une accumulation de savoirs ne fait pas une culture... mais une certaine qualité des savoirs, que l'on peut décrire ainsi. La culture, c'est le fait que les savoirs sont *disponibles*, qu'ils peuvent servir dans des circonstances toutes différentes de celles où on les a acquis, qu'en apprenant on "apprend à apprendre". Que les savoirs sont *assimilés* : je ne puis disposer d'un savoir que si je l'ai fait mien, intégré à ma personnalité ; si j'ai acquis mon propre style. Que les savoirs sont *communicables* : un savoir qu'on ne peut partager ni confronter avec ceux des autres est rejeté hors de la culture »[72]. L'auteur nous parle de savoirs, mais son lexique résonne étrangement avec celui que nous avons employé pour parler des ressources : la disponibilité qu'il évoque renvoie à l'accessibilité, et la communicabilité à l'interopérabilité, esquissant un parallèle entre le rapport d'un individu à ses propres savoirs et le rapport d'un groupe humain à ses ressources culturelles.

Nous avons déjà rencontré un tel lien au cours de notre travail, à travers les considérations de Richard Gartner sur le rôle des métadonnées dans la culture, en tant qu'elles interviennent comme liant entre les éléments constitutifs d'une culture, ce qui permet son entretien et sa transmission à travers les générations. Gartner parle ici de culture collective, mais établit une continuité anthropologique, précisément par l'usage des métadonnées, avec le processus de formation des connaissances dans le modèle Donnée-Information-Connaissance, la culture étant des « connaissances consolidées ». Dès lors, ce qui permettrait à un groupe de constituer et de transmettre une culture ne serait pas de nature différente de ce qui permettrait à un individu de se forger des savoirs, une culture. Observer, décrire, contextualiser : les qualités invoquées par Gartner ont une teneur commune, qui est de requérir de l'attention. Lorsque Georg Frank nous explique que « La formation n'est d'abord que de l'attention investie en elle-même. La connaissance, c'est de l'attention réifiée, sa vivacité créatrice est cristallisée, et en ce sens, elle est également de l'attention capitalisée » (17, Frank, p. 68), nous voyons en premier lieu de nouveaux liens se tisser entre l'économie de l'attention et le modèle Donnée-Information-Connaissance. Si nous

[72] REBOUL, Olivier. *Les valeurs de l'éducation*. Paris : PUF, 254 p.

acceptons l'idée que la connaissance est de l'attention cristallisée, c'est également le cas de l'information, et même des données. Il faudrait alors étudier les différents régimes d'attention engagés dans l'observation des phénomènes, la détection de structures dans ces phénomènes, la considération de relations entre ces phénomènes. En deuxième lieu, il nous semble que nous pouvons également tisser de nouveaux liens entre l'attention et les métadonnées sociales, qui sont également de l'attention cristallisée. Nous faisons dès lors face à un paysage très différent de celui que nous avions d'abord envisagé. Au consommateur dont l'attention raréfiée par l'abondance de l'offre est un bien disputé, et qui subit les diverses menées pour s'emparer de son attention, se substitue un *prosommateur* qui peut choisir de donner son attention (et éventuellement ses connaissances, donc de l'attention capitalisée) pour participer à la description de telle ressource, à la création voire à la structuration de métadonnées la concernant, et ce faisant participer à un édifice culturel qui lui permettra par ailleurs de se former, d'investir son attention en elle-même.

Selon nous, cette posture individuelle (ici formulée en termes d'attention mais proche de la notion de « culture informationnelle », ou « *information literacy* » dans le monde anglo-saxon) est le complément nécessaire du filtrage qu'évoquent Alemu et Stevens, si l'on veut pouvoir se forger une culture individuelle dans une société d'abondance informationnelle. C'est à la condition de cette démarche active, de ce « filtrage cognitif », de cette politique de l'attention, de cette capacité à décider par soi-même ce qui correspond à ses véritables besoins, mais aussi de décider à quel moment on accepte de donner de l'attention et à quel moment on souhaite la capitaliser en se formant, que la structuration des métadonnées basée sur les standards et les ontologies, et l'enrichissement par les métadonnées socialement construites, peuvent véritablement être une opportunité pour se frayer son propre chemin dans une offre abondante, découvrir de nouvelles ressources à partir de celles que l'on connaît déjà, orienter sa curiosité, et se constituer ainsi un système de savoirs richement interconnectés, « assimilés », utiles à la vie.

Bibliographie

Introduction

(1) GARTNER, Richard. *Metadata : Shaping Knowledge from the Antiquity to the Semantic Web*. Londres : Springer, 2016, 114 p.

(2) SIMON H., *Designing Organizations for an Information-Rich World*. **In** : GREENBERGER M. éd. *Computers, communications, and the public interest*. Baltimore, MD : The John Hopkins Press, 1971, pp. 38-72.

Première partie

(3) MIKSA, Francis. *Metadata*. **[En ligne]**. Austin : The University of Texas at Austin. Cours *Organizing and Providing Access to Information*, Graduate School of Library and Information Science, 2000. Disponible sur : <https://www.ischool.utexas.edu/~l38613dw/readings/Miksa-Metadata-000918.PDF> (consulté le 20 août 2018)

(4) ACKOFF, Russell. *From Data to Wisdom*. Journal of Applied Systems Analysis, 1989, vol. 16, pp. 3-9.

(5) ZINS, Chaim. *Conceptual Approaches for Defining Data, Information and Knowledge*. Journal of the American Society for Information Science and Technology, 2007, vol. 58(4), pp. 479-483.

(6) PUIMATTO, Gérard. *Les métadonnées : pourquoi et pour quoi faire ?* **In** : Le blog Savoirs CDI, 2009. **[En ligne]**. Disponible sur : <https://www.reseau-canope.fr/savoirscdi/fileadmin/fichiers_auteurs/Societe_de_l_information/Tic_et_documentation/Les_metadonnees_Puimatto.pdf> (consulté le 20 août 2018)

(7) RILEY, Jenn. *Understanding metadata - What is metadata and what is it for?*. **[En ligne]**. NISO. 1er janvier 2017, 45 p. Disponible sur : <https://www.niso.org/publications/understanding-metadata-2017> (consulté le 20 août 2018)

(8) GILLILAND, Anne. *Setting the stage.* **In** BACA, Murtha éd., *Introduction to metadata* (2nd edition), Oxford University Press, 2008, pp. 1-19.

(9) BROUDOUX, Evelyne. *Indexation collaborative : traces de lecture et constitution de communautés.* **In** *Bibliothèques 2.0 à l'heure des médias sociaux*. Éditions du Cercle de la librairie, 2012, pp.125-134.

(10) PECCATTE, Patrick. *Une plate-forme sociale pour la redocumentarisation d'un fonds iconographique.* **In** BROUDOUX, Evelyne et CHARTRON, Ghislaine (dir.) *Traitements et pratiques documentaires : vers un changement de paradigme ?*, Actes de la deuxième conférence « Document numérique et Société », 2008, ADBS, Collection : Sciences et techniques de l'information, 2008.

(11) ZACKLAD, Manuel. *Réseaux et communautés d'imaginaire documédiatisées* **In** Skare, R., Lund, W. L., Varheim, A., *A Document (Re)turn*. Frankfurt am Main : Peter Lang, 2007, pp. 279-297.

(12) ALEMU Getaneh et STEVENS Brett, *An Emergent Theory of Digital Library Metadata*. Elsevier, 2015, 134 p.

(13) Direction Interministérielle du Numérique et du Système d'Information et de Communication de l'État (DINSIC), *Référentiel Général d'Interopérabilité – Standardiser, s'aligner et se focaliser pour échanger efficacement*. Version 2.0, Décembre 2015, 84 p.

(14) BLANDIN, Bernard. *Les normes sur les technologies de l'information pour l'éducation, la formation et l'apprentissage*. FFFOD – CESI. 23 juin 2003.

(15) CHAN, Lois Mai et ZENG, Marcia Lei, *La réalisation de l'interopérabilité entre vocabulaires d'accès matière et système d'organisation de la connaissance : une analyse méthodologique*. **[En ligne]**. 68th IFLA Council and General Conférence, 2002. Disponible sur : <https://archive.ifla.org/IV/ifla68/papers/008-122f.pdf> (consulté le 20 août 2018)

Deuxième partie

(16) École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB). *Fiche pratique n°1754*. **[En ligne]**. 2013. Disponible sur : <http://www.enssib.fr/bibliotheque-numerique/documents/1754-comprendre-et-connaitre-la-chaine-du-livre.pdf> (consulté le 20 août2018)

(17) FRANCK, Georg. *Économie de l'attention* **In** CITTON, Yves (dir.). *L'économie de l'attention*. Paris : La Découverte, 2014, pp. 55-72.

(18) WALTER, David. *Nielsen book UK study: the importance of metadata for discoverability and sales.* **[En ligne]**. 30 novembre 2016. Disponible sur : <http://www.nielsenbookdata.co.uk/uploads/10451_Nielsen_Book_UK_Study_The_Importan

ce_of_Metadata_for_Discoverability_and_Sales_Digital_D6%20(1).pdf> (consulté le 20 août 2018)

(19) ODEH, Souad et CHARTRON, Ghislaine. *Acteurs et économie des métadonnées du livre en France : analyse et avenir*. Documentation et bibliothèques. 2016, 62(1), pp. 21-32.

(20) Association des Librairies Informatisées et utilisatrice de Réseaux (ALIRE). *Présentation de Dilicom et du FEL*. **[En ligne]**. Disponible sur : <http://www.alire.asso.fr/index.php/fel-messages-edi/dilicom-et-electre> (consulté le 20 août 2018)

(21) BACKERT, Véronique. *ONIX : une norme pour communiquer entre familles professionnelles ?* **[En ligne]**. 27 juin 2014. Disponible sur : <http://www.bnf.fr/documents/afnor2014_onix.pdf> (consulté le 20 août 2018)

(22) EDItEUR. *FAQs*. **[En ligne]**. Disponible sur : <https://www.editeur.org/74/FAQs/> (consulté le 20 août 2018)

(23) EDItEUR. *Overview*. **[En ligne]**. Disponible sur : <https://www.editeur.org/83/Overview/> (consulté le 20 août 2018)

(24) Commission de Liaison Interprofessionnelle du Livre (CLIL). *Fiche produit du livre*. **[En ligne]**. Novembre 2013. Disponible sur : <https://clil.centprod.com/information/telechargementDoc.html?action=ouvrir&id=373> (consulté le 20 août 2018)

(25) EDItEUR. *ONIX Livres - Format Information Produit - Spécification*. **[En ligne]**. Version 3.0, avril 2009. Disponible sur : <http://www.cercledelalibrairie.org/Document/show.aspx?id=2> (consulté le 20 août 2018)

(26) EDItEUR. *Code-Lists*. **[En ligne]**. Disponible sur : <https://www.editeur.org/14/Code-Lists/> (consulté le 20 août 2018)

(27) BEKY, Endre. *ONIX: Is there a return on investment for all publishers?* Publishing Research Quaterly. Juin 2004, vol. 20, pp. 3-8.

(28) CLIL - Commission FEL. *Guide pratique ONIX*. **[En ligne]**. Disponible sur : <https://clil.centprod.com/information/detailDoc.html?docId=33> (consulté le 20 août 2018)

Troisième partie

(29) BIZER, Christian, HEATH T. et BERNERS-LEE, Tim, *Linked Data – The Story So Far*. Disponible sur : <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf> (Consulté le 20 août 2018)

(30) BUSH, Vannevar. *As We May Think*. *The Atlantic Monthly*, July 1945.

(31) OTLET P., *Traité de documentation : le livre sur le livre, théorie et pratique*, Bruxelles, Editions Mundaneum, 1934, p. 428.

(32) PEDAUQUE R.T., *La redocumentarisation du monde*, 2007, Toulouse, éditions Cépaduès, 213 p.

(33) BROUDOUX, Evelyne et SCOPSI, Claire (coord.), *Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus,* Études de Communication, 2011/1 (n°36), pp. 9-22.

(34) GUTTERIDGE, Christophe. *Linked Data vs Open Data vs RDF Data*. Southampton Web and Data Innovation Team, 17 juillet 2011. Disponible sur : <https://blog.soton.ac.uk/webteam/2011/07/17/linked-data-vs-open-data-vs-rdf-data/> (Consulté le 20 août 2018)

(35) BACHIMONT *et al*. Enjeux et technologies : des données au sens. Documentaliste-Sciences de l'Information. 2011/4, vol. 48, p. 24-41.

(36) CHARLET, Jean et KEMBELLEC, Gérald. *Du web sémantique au web de données, quels enjeux professionnels ?* I2D – Information, données et documents. 2016/2, vol. 53, pp. 54-55.

(37) DOCTOROW, Cory. *Métadaube : mettre le feu à sept bonshommes de paille de la méta-utopie*, 2001. Disponible sur : <https://people.well.com/user/doctorow/metacrap.htm> (Consulté le 20 août 2018)

(38) GILL, Tony. *Metadata ant the Web.* **In** BACA, Murtha (éd)., *Introduction to metadata* (2nd edition), Oxford University Press, 2008, pp. 20-37.

(39) PELLEGRINI, Tassilo. *Semantic metadata in the publishing industry – technological achievements and economic implications*. Electronic markets, Février 2017, vol. 27, Springer, pp. 9-20.

(40) JORDA, Jean-Paul. *L'enjeu du web de données pour les éditeurs*. I2D – Information, données & documents. 2016/2, vol. 53, p. 48.

(41) MERCANTI GUERIN, Maria. *Émergence des prospectivistes 2.0, le cas des planneurs stratégiques*. Management & Avenir. Mars 2008, numéro 17. pp. 126-141.

(42) KEMBELLEC, Gérald. *Que voit réllement Google de la sémantique des pages web ?* I2D – Information, données & documents. 2016/2, vol. 53, p. 65.

(43) POUPEAU, Gautier. *Les éditeurs et les métadonnées : ONIX*. Blog Les petites cases, 6 avril 2006. Disponible sur : <https://www.lespetitescases.net/les-editeurs-et-les-metadonnees-onix> (Consulté le 20 août 2018)

(44) VATANT, Bernard. *Vous lisez quoi cet été ?* Blog Leçons de choses, 7 août 2013. Disponible sur : <https://mondeca.wordpress.com/2013/08/07/vous-lisez-quoi-cet-ete/> (Consulté le 20 août 2018)

(45) BERNERS-LEE, Tim. *Design Issues – Linked Data*. Disponible sur : < https://www.w3.org/DesignIssues/LinkedData.html> (Consulté le 20 août 2018)

(46) HEATH, Tom et BIZER, Christian. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

(47) FABRY, Cécilia *et al*. *Publier des données liées et ouvertes en sept étapes*. I2D – Information, données & documents. 2017/1, vol. 54, pp. 12-14.

(48) FRANCART, Thomas. *Des textes augmentés avec les données du Web*. I2D – Information, données & documents. 2016/2, vol. 53, p. 45.

Conclusion

(49) NIETZSCHE, Friedrich. *Seconde considération intempestive*. Paris : Flammarion, 1988, 187 p.

Annexes

Annexes de la deuxième partie

Annexe 1 - Exemple de notice Onix écrite en XML[73]


```
<?xml version="1.0" encoding="UTF-8"?>
<ONIXMessage release="3.0" xmlns="http://ns.editeur.org/onix/3.0/reference">
	<Header>
		<Sender>
			<SenderName>normes et standards</SenderName>
			<ContactName>normes et standards</ContactName>
			<EmailAddress>normesetstandards@normesetstandards.fr</EmailAddress>
		</Sender>
		<Addressee>
			<AddresseeName>normesetstandards.fr</AddresseeName>
		</Addressee>
		<MessageNumber>1</MessageNumber>
		<SentDateTime>20180129</SentDateTime>
	</Header>
	<Product>
		<RecordReference>notice1</RecordReference>
		<NotificationType>03</NotificationType>
		<ProductIdentifier>
			<ProductIDType>15</ProductIDType>
			<IDValue>9782123456780</IDValue>
		</ProductIdentifier>
		<DescriptiveDetail>
			<ProductComposition>00</ProductComposition>
			<ProductForm>ED</ProductForm>
			<ProductFormDetail>E101</ProductFormDetail>
			<ProductFormDetail>E200</ProductFormDetail>
			<ProductFormDetail>E202</ProductFormDetail>
			<ProductFormFeature>
				<ProductFormFeatureType>15</ProductFormFeatureType>
				<ProductFormFeatureValue>101C</ProductFormFeatureValue>
				<ProductFormFeatureDescription>ePub 3.0</ProductFormFeatureDescription>
			</ProductFormFeature>
			<ProductFormFeature>
				<ProductFormFeatureType>01</ProductFormFeatureType>
				<ProductFormFeatureValue>SLV</ProductFormFeatureValue>
				<ProductFormFeatureDescription>Couverture argentée</ProductFormFeatureDescription>
			</ProductFormFeature>
			<ProductFormFeature>
				<ProductFormFeatureType>02</ProductFormFeatureType>
				<ProductFormFeatureValue>BLK</ProductFormFeatureValue>
				<ProductFormFeatureDescription>Tranche noire</ProductFormFeatureDescription>
			</ProductFormFeature>
			<ProductFormFeature>
				<ProductFormFeatureType>03</ProductFormFeatureType>
				<ProductFormFeatureValue>Arial</ProductFormFeatureValue>
				<ProductFormFeatureDescription>Police majoritaire : Arial</ProductFormFeatureDescription>
			</ProductFormFeature>
			<!-- XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX --
>
			<!-- CARACTÉRISTIQUES DÉTAILLÉES DÉDIÉES ACCESSIBILITÉ -->
			<!-- Résumé d'accessibilité -->
```


[73] Syndicat National de l'Édition (SNE). *Exemple de notice Onix écrite en xml*. **[En ligne]**. Disponible sur : <https://www.sne.fr/document/exemple-de-notice-onix-ecrite-en-xml/> (consulté le 20 août 2018)

```
<ProductFormFeature>
<!-- Caractéristique = Détail d’accessibilité de publication numérique (cf. code list 79) -->
<ProductFormFeatureType>09</ProductFormFeatureType>
<!-- Valeur de Détail d’accessibilité (cf. code list 196) -->
<ProductFormFeatureValue>00</ProductFormFeatureValue>
<ProductFormFeatureDescription>Cet ePub est nativement accessible. Pour son adaptation aux publics empêchés de lire, l'éditeur a réalisé les actions suivantes : indication de langue, alternatives textuelles pour tous les éléments non textuels (dont formules scientifiques), respect de l'ordre de lecture, table des matières, index.</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Label du niveau d'accessibilité -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>03</ProductFormFeatureValue>
<ProductFormFeatureDescription>Accessibility Specification 1.0 AA</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Désactivitation des options d'accessibilité -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>10</ProductFormFeatureValue>
<ProductFormFeatureDescription>Aucune option d’accessibilité au système désactivée (exception)</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Table des matières -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>11</ProductFormFeatureValue>
<ProductFormFeatureDescription>Navigation dans la table des matières</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Index -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>12</ProductFormFeatureValue>
<ProductFormFeatureDescription>Navigation dans l’index</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Ordre de lecture -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>13</ProductFormFeatureValue>
<ProductFormFeatureDescription>Ordre de lecture</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Descriptions courtes -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>14</ProductFormFeatureValue>
<ProductFormFeatureDescription>Brèves descriptions alternatives</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Données visuelles aussi fournies sous forme non visuelle -->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>16</ProductFormFeatureValue>
<ProductFormFeatureDescription>Représentations graphiques de données également accessibles comme données non graphiques</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Contenu mathématique accessible-->
<ProductFormFeature>
<ProductFormFeatureType>09</ProductFormFeatureType>
<ProductFormFeatureValue>17</ProductFormFeatureValue>
<ProductFormFeatureDescription>Contenu mathématique accessible</ProductFormFeatureDescription>
</ProductFormFeature>
```

```
<!-- Préservation du foliotage papier -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>19</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Numérotation de pages équivalente au document imprimé</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Sychronisation texte-son -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>20</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Audio synchronisé préenregistré</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Indication de langue -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>22</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Balisage de la langue fourni</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Indication par rapport à  la dyslexie -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>24</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Lisibilité adaptée à la dyslexie</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- URL de description détaillée -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>94</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Page web pour les informations détaillées d'accessibilité : https://enseignants.nathan.fr/enseignants/accessibilite</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Page de l'organisme certificateur tiers -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>95</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Page web de l'intermédiaire approuvé : http://www.braillenet.org/nous-contacter</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Tests d'accessibilité-->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>97</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Compatibilité testée par l'outil ACE + contrôle humain sur iPad</ProductFormFeatureDescription>
</ProductFormFeature>
<!-- Contact de l'organisme certificateur tiers -->
<ProductFormFeature>
	<ProductFormFeatureType>09</ProductFormFeatureType>
	<ProductFormFeatureValue>98</ProductFormFeatureValue>
	<ProductFormFeatureDescription>Contact intermédiaire approuvé : contact@braillenet.org</ProductFormFeatureDescription>
	</ProductFormFeature>
<!-- XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX -->
<ProductFormDescription>ePub3 reflowable, contenant du texte, des illustrations en couleur, de l'audio pour du media-overlay, de la vidéo, du media-overlay et des formules mathématiques</ProductFormDescription>
<!--  Catégorie commerciale (cf. code list 12) - Valeur = Livre scolaire, déclaré par l'éditeur -->
<TradeCategory>08</TradeCategory>
<!-- Contenu principal et secondaire -->
<!-- Type de contenu produit (cf. code list 81) - Valeur = Texte (lisible par l'homme) -->
```

```
                        <PrimaryContentType>10</PrimaryContentType>
                        <!-- Images statiques / graphiques -->
                        <ProductContentType>07</ProductContentType>
                        <!-- Autre enregistrement audio -->
                        <ProductContentType>04</ProductContentType>
                        <!-- Vidéo -->
                        <ProductContentType>06</ProductContentType>
                        <!-- Script -->
                        <ProductContentType>40</ProductContentType>
                        <!-- Pays de fabrication (cf. code list 91) -->
                        <CountryOfManufacture>FR</CountryOfManufacture>
                        <!-- Protection technique de publication numérique (cf. code list 144) - Valeur = Filigrane numérique (alias
tatouage) -->
                        <EpubTechnicalProtection>02</EpubTechnicalProtection>
                        <!-- Restrictions d'utilisation : -->
                        <!-- Autorisation de lecture par un moteur de synthèse vocale -->
                        <EpubUsageConstraint>
                                <!-- Type d'utilisation (cf. code list 145) - Valeur = Synthèse vocale -->
                                <EpubUsageType>05</EpubUsageType>
                                <!-- Statut d'utilisation (cf. code list 146) - Valeur = Autorisation sans limite -->
                                <EpubUsageStatus>01</EpubUsageStatus>
                        </EpubUsageConstraint>
                        <!-- Autorisation de 5 copies personnelles -->
                        <EpubUsageConstraint>
                                <!-- Type d'utilisation (cf. code list 145) - Valeur = Partage par différents périphériques de lecture
pour un même utilisateur -->
                                <EpubUsageType>04</EpubUsageType>
                                <!-- Statut d'utilisation (cf. code list 146) - Valeur = Autorisation soumise à  limite -->
                                <EpubUsageStatus>02</EpubUsageStatus>
                                <EpubUsageLimit>
                                        <Quantity>5</Quantity>
                                        <!-- Unité de la quantité limite (cf. code list 147) - Valeur = Périphériques -->
                                        <EpubUsageUnit>06</EpubUsageUnit>
                                </EpubUsageLimit>
                        </EpubUsageConstraint>
                        <ProductClassification>
                                <ProductClassificationType>05</ProductClassificationType>
                                <ProductClassificationCode>49019900</ProductClassificationCode>
                        </ProductClassification>
                        <Collection>
                                <CollectionType>11</CollectionType>
                                <TitleDetail>
                                        <TitleType>01</TitleType>
                                        <TitleElement>
                                                <TitleElementLevel>02</TitleElementLevel>
                                                <TitleText>Nom de la collection</TitleText>
                                        </TitleElement>
                                </TitleDetail>
                        </Collection>
                        <TitleDetail>
                                <TitleType>01</TitleType>
                                <TitleElement>
                                        <TitleElementLevel>01</TitleElementLevel>
                                        <TitleText>Titre du livre</TitleText>
                                </TitleElement>
                        </TitleDetail>
                        <Contributor>
                                <SequenceNumber>1</SequenceNumber>
                                <ContributorRole>A01</ContributorRole>
                                <NameIdentifier>
                                        <NameIDType>16</NameIDType>
                                        <IDTypeName>ISNI</IDTypeName>
                                        <IDValue>Numéro ISNI</IDValue>
                                </NameIdentifier>
```

```
                        <NamesBeforeKey>Normes</NamesBeforeKey>
                        <KeyNames>Standards</KeyNames>
                        <BiographicalNote>Note bibliographique</BiographicalNote>
                    </Contributor>
                    <Language>
                        <LanguageRole>01</LanguageRole>
                        <LanguageCode>fre</LanguageCode>
                    </Language>
                    <Extent>
                        <ExtentType>00</ExtentType>
                        <ExtentValue>152</ExtentValue>
                        <ExtentUnit>03</ExtentUnit>
                    </Extent>
                    <!-- ILLUSTRATIONS -->
                    <!-- Illustrations oui/non(cf. code list 152) - Valeur = oui -->
                    <Illustrated>02</Illustrated>
                    <NumberOfIllustrations>185</NumberOfIllustrations>
                    <!-- Contenu secondaire -->
                    <AncillaryContent>
                        <!-- Illustrations et autres types de contenu (cf. code list 25) -->
                        <AncillaryContentType>02</AncillaryContentType>
                        <AncillaryContentDescription>Illustrations en couleur</AncillaryContentDescription>
                        <Number>187</Number>
                    </AncillaryContent>
                    <!-- INDEXATION THÉMATIQUE -->
                    <Subject>
                        <!-- Identifiant de classification thématique (cf. code list 27) - Valeur = CLIL -->
                        <SubjectSchemeIdentifier>29</SubjectSchemeIdentifier>
                        <SubjectCode>3007</SubjectCode>
                        <SubjectHeadingText>Manuels scolaires Secondaire Technique et
professionnel</SubjectHeadingText>
                    </Subject>
                    <Subject>
                        <!-- Identifiant de classification thématique (cf. code list 27) - Valeur = Classification thématique
Thema -->
                        <SubjectSchemeIdentifier>93</SubjectSchemeIdentifier>
                        <SubjectCode>YP</SubjectCode>
                        <SubjectHeadingText>Matériel didactique</SubjectHeadingText>
                    </Subject>
                    <Subject>
                        <!-- Identifiant de classification thématique (cf. code list 27) - Valeur = Qualificatif de niveau
scolaire Thema -->
                        <SubjectSchemeIdentifier>97</SubjectSchemeIdentifier>
                        <SubjectCode>4Z-FR</SubjectCode>
                        <SubjectHeadingText>France : programmes d'enseignement</SubjectHeadingText>
                    </Subject>
                    <!-- PUBLIC CIBLE -->
                    <Audience>
                        <!-- Type de codification du public cible (cf. code list 29) - Valeur = Public ONIX -->
                        <AudienceCodeType>01</AudienceCodeType>
                        <!-- Public visé (cf. code list 28) - Valeur = Enseignants -->
                        <AudienceCodeValue>05</AudienceCodeValue>
                    </Audience>
                </DescriptiveDetail>
                <CollateralDetail>
                    <TextContent>
                        <TextType>03</TextType>
                        <ContentAudience>00</ContentAudience>
                        <Text>Texte de présentation de l'éditeur</Text>
                    </TextContent>
                    <SupportingResource>
                        <ResourceContentType>01</ResourceContentType>
                        <ContentAudience>00</ContentAudience>
                        <ResourceMode>03</ResourceMode>
```

<ResourceVersion>
<ResourceForm>01</ResourceForm>
<ResourceVersionFeature>
<ResourceVersionFeatureType>01</ResourceVersionFeatureType>
<FeatureValue>D502</FeatureValue>
</ResourceVersionFeature>
<ResourceVersionFeature>
<ResourceVersionFeatureType>02</ResourceVersionFeatureType>
<FeatureValue>800</FeatureValue>
</ResourceVersionFeature>
<ResourceVersionFeature>
<ResourceVersionFeatureType>03</ResourceVersionFeatureType>
<FeatureValue>521</FeatureValue>
</ResourceVersionFeature>
<ResourceVersionFeature>
<ResourceVersionFeatureType>07</ResourceVersionFeatureType>
<FeatureValue>53000</FeatureValue>
</ResourceVersionFeature>
<!-- url vers le visuel dela 1èere de couverture -->
<ResourceLink>http://www.nathan.fr/catalogue/fiche-produit.asp?ean13=9782092558232</ResourceLink>
<ContentDate>
<ContentDateRole>17</ContentDateRole>
<DateFormat>00</DateFormat>
<Date>20171205</Date>
</ContentDate>
</ResourceVersion>
</SupportingResource>
<SupportingResource>
<ResourceContentType>16</ResourceContentType>
<ContentAudience>00</ContentAudience>
<ResourceMode>06</ResourceMode>
<ResourceVersion>
<ResourceForm>01</ResourceForm>
<!-- Lien vers le feuilleteur -->
<ResourceLink>http://www.nathan.fr/catalogue/fiche-produit.asp?ean13=9782092558232</ResourceLink>
</ResourceVersion>
</SupportingResource>
</CollateralDetail>
<PublishingDetail>
<Imprint>
<ImprintIdentifier>
<ImprintIDType>06</ImprintIDType>
<IDValue>3052452430017</IDValue>
</ImprintIdentifier>
<ImprintName>NORMES ET SANDARDS</ImprintName>
</Imprint>
<Publisher>
<PublishingRole>01</PublishingRole>
<PublisherName>Normes et Standards</PublisherName>
</Publisher>
<PublishingDate>
<PublishingDateRole>01</PublishingDateRole>
<DateFormat>00</DateFormat>
<Date>20180301</Date>
</PublishingDate>
<PublishingDate>
<PublishingDateRole>27</PublishingDateRole>
<DateFormat>00</DateFormat>
<Date>20171223</Date>
</PublishingDate>
<PublishingDate>
<PublishingDateRole>02</PublishingDateRole>

```
                        <DateFormat>00</DateFormat>
                        <Date>20180301</Date>
                    </PublishingDate>
                    <PublishingDate>
                        <PublishingDateRole>09</PublishingDateRole>
                        <DateFormat>00</DateFormat>
                        <Date>20171223</Date>
                    </PublishingDate>
                </PublishingDetail>
                <RelatedMaterial>
                    <RelatedProduct>
                        <ProductRelationCode>13</ProductRelationCode>
                        <ProductIdentifier>
                            <ProductIDType>03</ProductIDType>
                            <IDValue>ISBN 13 de la version papier (sans les tirets)</IDValue>
                        </ProductIdentifier>
                        <ProductForm>BC</ProductForm>
                    </RelatedProduct>
                </RelatedMaterial>
                <ProductSupply>
                    <Market>
                        <Territory>
                            <RegionsIncluded>WORLD</RegionsIncluded>
                            <CountriesExcluded>BE CA CH</CountriesExcluded>
                        </Territory>
                    </Market>
                    <SupplyDetail>
                        <Supplier>
                            <SupplierRole>02</SupplierRole>
                            <SupplierIdentifier>
                                <SupplierIDType>06</SupplierIDType>
                                <IDValue>GLN du distributeur</IDValue>
                            </SupplierIdentifier>
                            <SupplierName>Nom du distributeur</SupplierName>
                        </Supplier>
                        <ProductAvailability>20</ProductAvailability>
                        <Price>
                            <PriceType>04</PriceType>
                            <DiscountCoded>
                                <DiscountCodeType>02</DiscountCodeType>
                                <DiscountCodeTypeName>Nom de la table de
remise</DiscountCodeTypeName>
                                <DiscountCode>Code remise</DiscountCode>
                            </DiscountCoded>
                            <PriceAmount>12.00</PriceAmount>
                            <Tax>
                                <TaxType>01</TaxType>
                                <TaxRatePercent>5.5</TaxRatePercent>
                                <TaxableAmount>11.37</TaxableAmount>
                            </Tax>
                            <CurrencyCode>EUR</CurrencyCode>
                            <Territory>
                                <CountriesIncluded>FR</CountriesIncluded>
                            </Territory>
                        </Price>
                        <Price>
                            <PriceType>01</PriceType>
                            <DiscountCoded>
                                <DiscountCodeType>02</DiscountCodeType>
                                <DiscountCodeTypeName>Nom de la table de
remise</DiscountCodeTypeName>
                                <DiscountCode>Code remise</DiscountCode>
                            </DiscountCoded>
                            <PriceAmount>11.37</PriceAmount>
```

```
                                        <CurrencyCode>EUR</CurrencyCode>
                                        <Territory>
                                                <RegionsIncluded>WORLD</RegionsIncluded>
                                                <CountriesExcluded>FR BE CA CH</CountriesExcluded>
                                        </Territory>
                                </Price>
                        </SupplyDetail>
                </ProductSupply>
        </Product>
</ONIXMessage>
```

Annexe 2 - Guide pratique ONIX - Commission FEL de la CLIL : les données vitales

Désignation	Liste des éléments concernés	Précisions et recommandations CLIL
Auteur	Cf contributeur	
Barème de remises	Cf remise	
Classification ScoLOM.FR	Cf Sujet	
Code de collection sérielle	Cf Collection	
Code ISBN	Cf Identifiant produit	
Collection	P.5.1 Type Collection (Collection type code) P.5.3 (Collection identifier type code) P.5.5 (Identifier value) P.5.7 (Title element level) P.5.10 (Title text)	Collection
Composants	P.4.2 (Product identifier type code) P.4.4 (Identifier value) P.4.5 (Product form code (product part)) P.4.6 Présentation détaillée du produit (Product form detail)	
	P.4.13 Nombre d'exemplaires (Number of copies (product part))	
Conditions de retour	Cf Retour	
Contributeur	P.7.1 (Contributor sequence number) P.7.2 (Contributor role) P.7.6 (Name identifier type) P.7.8 (Identifier value) P.7.12 (Person name part 2 : names before key names) P.7.14 (Person name part 4 : key names)	Contributeurs
Date d'application (concerne le prix uniquement)	P. 26.83 (Price date role code) P. 26.84 (Date format) P. 26.85 (Date)	
Date de parution	P.20.3 (Publishing date role code) P.20.4 (Date format) P.20.5 (Date)	Date de parution
Date fin commercialisation	P.20.3 (Publishing date role code) P.20.4 (Date format) P.20.5 (Date)	Date de fin de commercialisation
Date limite de retour	Cf Retours	
DEEE (Montant Eco-taxe)	Créé dans la version 3.0.4	
Diffuseur	P.25.1 (Agent role) P.25.2 (Agent identifier type code) P.25.4 (Identifier value) P.25.5 (Agent name)	Diffuseur
Disponibilité	P.26.17 (Product availability)	Disponibilité
Distributeur	P.26.1 (Supplier role) P.26.2 (Supplier identifier type code) P.26.4 (Identifier value)	Distributeur

Désignation	Liste des éléments concernés	Précisions et recommandations CLIL
DRM, logiciels de verrouillage et droits associés		DRM, logiciels de verrouillage et droits associés
Droits d'usage		Droits d'usage (livre numérique)
Distribution unique ou multiple	N'est pas géré en entrée Cf Distributeur	
Eco-taxe	Cf DEEE	
Eco-participation	Créé dans la version 3.0.4	
EAN13	Cf identifiant produit	
Epaisseur	Cf Mesures	
Extrait	Groupe P.16 xxx (Links to supporting resources)	Extrait
Forme du produit (anciennement présentation éditeur)	P.3.2 Présentation produit (Product form code)	Forme du produit
Hauteur	Cf Mesures	
Identifiant Collection	Cf Collection	
Identifiant Contributeur	Cf contributeur	
Identifiant Diffuseur	Cf Diffuseur	
Identifiant Distributeur	Cf Distributeur	
Identifiant Editeur/Marque éditoriale	Cf Editeur/Marque éditoriale	
Identifiant produit	P.2.1 Type de code de l'identifiant produit (Product identifier type code) P.2.3 Valeur de l'identifiant produit (Identifier value)	Utiliser des identifiants normalisés pour désigner les produits Identifiant produit
Impression à la demande	P.26.17 (Product availability)	
ISNI	Cf Contributeur	
Largeur	Cf Mesures	
Libellé	P.6.1 (Title type code) P.6.2 (Title element level)	

Désignation	Liste des éléments concernés	Précisions et recommandations CLIL
	P.6.5 (Title text)	
Libellé caisse	Cf Libellé	
Libellé étendu	Cf Libellé	
Libellé standard	Cf Libellé	
Lien entre les codes	Cf produit lié	
Marque éditoriale	P.19.1 (Imprint identifier type) P.19.3 (Identifier value) P.19.4 (Imprint or brand name)	Editeur et marque éditoriale
Mesures	P.3.12 Type de mesure (Measure type code) P.3.13 Mesures (Measurement) P.3.14 Unité de mesure (Measure unit code)	Mesures
Millésime	P.6.4 (Year of annual)	
Minimum de commande (nouveau)	Créé dans la version 3.0.3	Minimum de commande
Montant hors taxe	P.26.69 (Amount of price taxable)	
Motif de suppression	P.1.3 Motif de suppression (Reason for deletion)	Limiter les suppressions de notices à des cas spécifiques et exceptionnels
Multiple de commande	Créé dans la version 3.0.3	Multiple de commande
Nombre de pages	P.11.1 (Extent type code) P.11.2 (Extent value) P.11.4 (Extent unit)	Nombre de pages
Nombre de références	Calculé automatiquement par DILICOM en fonction des éléments décrits dans P.4 <ProductPart> Cf également Composants.	
Numéro de la notice BnF	Cf Identifiant produit N'est pas géré en entrée	
Pays de production	P.3.15 Pays de production (Country of manufacture)	Pays de production
Peut être commandé par le revendeur	P.26.17 (Product availability)	

Désignation	Liste des éléments concernés	Précisions et recommandations CLIL
Poids	Cf Mesures	
Présentation magasin	Cf Forme du produit (anciennement présentation éditeur)	
Prix TTC	P.26.62 (Price amount)	Prix
Produit lié	P.23.1 (Product relation code) P.23.2 (Product identifier type code) P.23.4 (Identifier value)	Produits liés
Produits spécifiques (qui sont destinées à un segment de clients : ex : présentoir garni pour une enseigne, produit présenté par les représentants)	P.24.5 (Sales restriction type code) P.24.6 (Sales outlet identifier type) P.24.8 (Identifier value)	
Produits spécifiques PNB (réservés aux collectivités)		Produits spécifiques pour un marché donné : le marché des collectivités (PNB)
Publics	P.13.2 (AudienceCodeType) P.13.4 (AudienceCodeValue) P.13.5 <AudienceRangeQualifier>	
Publics – « Réservé aux enseignants »	P.9.1 (Edition type code)	Publics – Réservé aux enseignants
Référence interne fournisseur	Cf Identifiant produit	
Remise	P.26.54 (Discount code type code) P.26.55 (Discount code type name) P.26.56 (Discount code value)	
Retour	P.26.14 (Returns code type) P.26.16 (Returns conditions code) P.26.18 (Supply date role code) P.26.19 (Date format) P.26.20 Date (Date)	
Désignation	**Liste des éléments concernés**	**Précisions et recommandations CLIL**
Retours (instructions)	Composé <ReturnsConditions> P.26.14 (Returns code type) P.26.15 (Returns Code type name) P.26.16 (Returns conditions code) P.26.16.a (Returns conditions note) (3.0.3 uniquement)	
Scolaire	P.3.9 Catégorie (Trade category code)	Scolaire
Sujet	P.12.1 (Main subject flag) P.12.2 (Subject scheme identifier) P.12.5 (Subject code)	Sujet
Symbolisation	P.2.4 Type de code-à-barres (Barcode type) P.2.5 Position sur le produit (Position on product)	
TARIC (code de nomenclature douanière) (nouveau)	P.3.22 (Product classification type code)	TARIC
Taux de TVA	P.26.66 (Tax type) P.26.68 (Tax rate percent)	
Texte de présentation par l'éditeur	P.14.1 Type de texte P.14.2 Audience P.14.3 Texte	Texte de présentation par l'éditeur
Thema (nouveau)	Cf Sujet	
Thème CLIL	Cf Sujet	
Type de lot	P.3.1 Composition du produit (Product composition)	
Type de mise à jour	P.1.2 Type de mise-à-jour (Notification or update type code)	
Type de prix	P.26.42 (Unpriced item type) P.26.43 (Price type code)	Prix
Type de produit	P.3.1 (Product composition)	Type de produit
Vente promotionnelle (nouveau)		
Désignation	**Liste des éléments concernés**	**Précisions et recommandations CLIL**
Visuels	Groupe P.16 xxx (Links to supporting resources)	