Indexer des documents « du dedans » : comment répondre à la question du lieu de la donnée ?

Index Documents from the Inner Side: How Best to Respond to the where-is-the-data Problem?

Xavier-Laurent Salvador

Édition électronique

URL : <https://journals.openedition.org/questionsdecommunication/11057>

DOI : 10.4000/questionsdecommunication.11057

ISSN : 2259-8901

Éditeur

Presses universitaires de Lorraine

Édition imprimée

Date de publication : 1 septembre 2017

Pagination : 71-93

ISBN : 9782814303256

ISSN : 1633-5961

Référence électronique

Xavier-Laurent Salvador, « Indexer des documents « du dedans » : comment répondre à la question du lieu de la donnée ? », *Questions de communication* [En ligne], 31 | 2017, mis en ligne le 01 septembre 2019, consulté le 26 octobre 2021. URL : <http://journals.openedition.org/questionsdecommunication/11057> ; DOI : <https://doi.org/10.4000/questionsdecommunication.11057>

XAVIER-LAURENT SALVADOR

Théories: Langage Discours Digital Humanities

Université Paris 13

F-93430

xavier-laurent.salvador@univ-paris13.fr

INDEXER DES DOCUMENTS « DU DEDANS » : COMMENT RÉPONDRE À LA QUESTION DU LIEU DE LA DONNÉE ?

Résumé. — Dans la lignée des travaux de Dominique Cotte et Marie Després-Lonnet, l'article pose la question du lieu de la donnée au sein des ressources numériques et ambitionne d'y répondre en procédant à une description de ce que le mot *locator* de la ressource désigne, étant à la fois son adresse et le moyen d'y accéder. L'adresse du serveur, le chemin à l'intérieur du disque puis l'ensemble des moyens de dénomination et de description fine des *markup languages* permettent de construire une image fidèle de l'information documentaire au sein de collections qui autorisent de nouvelles stratégies pour la fouille de données. Quoi qu'il en soit, la fouille textuelle et l'analyse quantitative demeurent des positionnements idéologiques quand seuls l'intelligence et le discernement humains peuvent en dernier lieu en apprécier la richesse et en valider la pertinence.

Mots clés. — fouille textuelle, *Uniform Resource Name* (URN), *Uniform Resource Locator* (URL), *XML Path Language* (XPath), *Extensible Markup Language* (XML), *Resource Description Framework* (RDF), données

Dans une vision figée du web, un serveur est le dépôt d'un fichier-document comme une bibliothèque est celui de livres et autres ressources manuscrites ou imprimées¹. La granularité symbolique – c'est-à-dire la conception de la donnée comme un grain d'information au sein du silo internet – de l'archive documentaire repose sur la pensée du document, soit produit, soit calculé *ad hoc*, qui serait circonscrit par les bornes de l'intention du dépositaire de la source ou de l'algorithme. Comme l'explique pourtant Dominique Cotte (2011 : 59),

« le travail documentaire portait le plus souvent, comme son nom l'indique, sur ces objets achevés que sont les documents. On sait qu'une première brèche a été ouverte dès lors que le document, devenu numérique, évoluait vers une structure ouverte, granulaire et recomposable. L'idée du web de données est bien de repérer et de structurer des grains d'information à l'intérieur des structures documentaires pour les exploiter en les reliant entre elles de manière logique et standardisée ».

L'URL² (W3C, 1997) du document permet au lecteur d'identifier le lieu du dépôt – c'est pourquoi on en trouve dans les bibliographies – et de le consulter à travers une opération de téléchargement puis de reconstruction par le navigateur web (Berners-Lee, 1994a). Mais si l'on change le paradigme de l'élaboration de la recherche en considérant non plus les documents mais des données, il faut étendre la question de l'identification du lieu du dépôt (du document) à celle du lieu de recueil de l'information (des données). Ainsi, dans une logique qui associe une base de données à une architecture REST³ (Monnin, 2014), l'URL est-il associé à une fonction de programmation qui redimensionne le document demandé en fonction de la ressource accessible – la base de données – et de la requête du lecteur qui devient en quelque sorte co-constructeur de sa ressource. Autrement dit, l'information donnée à lire est élaborée avec le lecteur et les limites du document, traditionnellement les bornes d'une œuvre, sont fragmentées par l'URL dont la portée outrepassa les limites autrefois imposées par l'existence d'un fichier pour pénétrer jusqu'au cœur de l'information afin d'extraire la donnée précisément désignée par l'URL :

« Et au cours de ce travail se produit la transmutation de "données" en "documents" (en prenant bien évidemment en compte la façon dont l'affichage dynamique de ces éléments interroge à son tour la notion même de document). La volatilité de cette production, son caractère éphémère posent des problèmes redoutables dont doivent s'emparer professionnels de l'information, chercheurs en science de l'information, éditeurs et industriels » (Cotte, 2013 : 24).

Ce processus ramène à la problématique de la langue courante : « Dans le monde du web sémantique, *annoter* [souligné par l'auteur] c'est taguer; autrement dit indexer en langage documentaire. L'activité traditionnelle consistait à indexer des documents "du dehors" alors qu'ici, on indexe des données à l'intérieur d'un document ou d'une

¹ Dans le titre du présent article, l'expression « indexer [...] "du dedans" » réfère à D. Cotte (2011 : 59) qui parle d'une activité traditionnelle qui consistait à indexer « du dehors ».

² Les abréviations des noms relevant de l'informatique (formats, langages, protocoles, etc.) utilisées dans cet article sont rassemblées en annexe et accompagnées de leurs formes complètes, ainsi que de définitions succinctes.

³ Les applications *Idef(x)* (accès : <http://www.idefx.fr>), *Isilex* (accès : <http://www.isilex.fr>) et *Syllabator* (accès : <http://www.projetprada.fr/versification>) sont des applications qui toutes reposent sur une architecture XML/Xquery/REST.

*micro unité d'information*⁴ » (Cotte, 2011 : 59, nous soulignons). La question de la micro-unité d'information extraite d'une annotation de « l'intérieur » du document produit une fragmentation de l'objet documentaire et définit l'un des principaux enjeux des *small data*⁵ : de quels moyens disposons-nous alors pour répondre à la question « où se trouve la donnée textuelle ? » Quelle est l'adresse de la donnée sur le web au sein du document-texte lui-même ?

Pour cerner les enjeux de l'adressage des données et des conventions qui les identifient (on peut penser aux coordonnées d'un tableau, aux *offsets* – emplacements en mémoire – d'une chaîne binaire, aux balises d'un formalisme XML, aux conventions *xpath*), il faut peut être repartir de la logique documentaire issue de l'approche d'Andrea Crestadoro (1856), bibliothécaire en chef de la bibliothèque publique de Manchester entre 1856 et 1860, qui développa une réponse, connue aujourd'hui sous l'acronyme *kwic*⁶ (*keywords in context*, voir Salvador, 2016 : 110-113), à la problématique de l'identification des ressources en fonction de thèmes de recherches. Fondée sur la co-occurrence de termes dont l'interpolation dans le catalogue a pour résultat de produire un panorama complet des ressources sur un sujet donné sans analyse sémantique (et donc sans intervention humaine), son approche posa les fondements d'une réflexion sur la logique documentaire ; celle-ci aboutit très vite à l'introduction des thésaurus puis des ontologies (*kwac* pour *keywords and context* et *kwoc* pour *keywords out of context*), dont la fonction est de guider le lecteur en bibliothèque sur la base des mots de sa recherche et des titres – ou des résumés – des ouvrages disponibles. Par extension du domaine de recherche, on constate aisément que la logique de l'adressage du lieu de la donnée textuelle dans le document numérique poursuit et approfondit la logique *kw(.)*, en proposant une série de pistes pour transformer le document en lieu de dépôt d'une information accessible *via* des séries de références, qui prolongent l'*URN* du document lui-même pour permettre la désignation *via* l'*URL* d'une information dans le document.

⁴ Il faut ajouter qu'il serait encore nécessaire de lever l'ambiguïté de ce passage pour définir clairement ce que signifie la question de « l'intérieur du document ». L'ancienne vision d'un espace documentaire borné, vision synchronique, laisse progressivement place à une vision diachronique de la nature documentaire : ainsi un document dont on connaît la source ou *intention première* serait-il fermé.

⁵ Par *small data*, il faut entendre un volume de données facilement représentable par l'esprit humain, par opposition au *big data*.

⁶ L'idée du format d'index *kwic* est de réaliser un index alphabétique de tous les mots signifiants – tous sauf les mots grammaticaux – d'un catalogue et de réaliser une indexation de chacun en en faisant le mot thème de l'entrée. De ce point vue, la co-occurrence de deux termes est tourmente. Si je pense un thème sous la forme « Les (1) poissons des (2) océans », je dois pouvoir construire un catalogue où les ouvrages « Les (2) océans et les (1) poissons » et « les (1) poissons dans les (2) océans » sont recensés sous les deux entrées : *poisson(s)* (1) et *oc(é)an(s)* (2). Le format *kwic* a été introduit par A. Crestadoro (1856) dans la période 1856-1860 à propos du catalogage des *Uniform Resource Name* (*URN* – nom abstrait qui permet de nommer une ressource, mais qui ne permet pas de la consulter, contrairement aux *Uniform Resource Identifier* – *URI* – de type *Locator*, ou *URL* : voir *infra*) de la bibliothèque de Manchester. Le concept de développement de liste faisant abstraction des mots non pertinents dans leur ordre d'apparition a connu un fort développement dans les années 1960.

C'est pourquoi, comme l'a montré Tim Berners-Lee (1998), l'architecture logique du serveur de données qui héberge l'information est primordiale : si, par le biais d'un changement de paradigme tel qu'il est opéré par la logique REST (voir *infra*), on abolit la pensée d'un dépôt de document téléchargeable au profit d'un document calculé par un algorithme associé à l'URL, on achève alors de fragmenter la pensée documentaire au profit d'une représentation artificielle du lieu de la donnée dont l'empan et la profondeur sont indissociables de la requête du lecteur. Auquel cas, il s'agit d'une co-construction (« *design thinking* ») entre le lecteur et la fonction qui moissonne les bases de données. En effet, la base de données textuelle est au cœur de la dynamique réflexive qui cerne progressivement la question des *small data* et l'exploitation de leur dynamique interne. Elle devient une ressource outillée existant *via* la double articulation d'un portail accessible sur le web, avec un outil de gestion de bases de données comme *Basex*⁷ rendant accessible les informations construites. Dans cette optique, l'information textuelle se comprend comme un élément au sein d'un vaste ensemble discontinu, dont la cohérence interne est liée à la nature de l'information véhiculée, qui s'affranchit des limites pensées par les auteurs premiers des données du corpus, leurs éditeurs ou leurs commentateurs pour intégrer ce que la philologie appelle « un cycle ».

L'interrogation des données en base s'articule autour du mot qui en constitue l'unité première. Pourtant, l'entrée lexicale dans le moteur de recherche, qui semble la plus évidente dans la mesure où elle constitue désormais le point d'entrée dans la recherche encyclopédique (je cherche *poisson* dans un formulaire) conçue par les GAFA⁸, pose à elle seule une série de problèmes dont la résolution implique, de la part du concepteur, un ensemble de prises de positions idéologiques en amont de l'ouverture du portail. Ainsi, lorsqu'il recherche *poisson* dans un moteur, le chercheur s'attend-il à ce que l'on comprenne qu'il veut trouver à la fois les occurrences du mot au singulier et au pluriel, en même temps que toutes les occurrences d'un mot lié au « poisson » par sa famille (*poissonnier*, *poissonnerie*), son réseau synonymique (« animal sous-marin »), sa relation de parenté (*maquereau*, *morue*, *hareng*). Au-delà,

⁷ *Basex* est un logiciel sous licence libre développé par une équipe de chercheurs de l'Université de Constance (Allemagne) qui offre des performances inégalées (sur le sujet, voir notamment Lauret, Thomas, s.d.) en associant un serveur de bases de données XML à un serveur REST.

⁸ L'acronyme GAFA (Google, Apple, Facebook et Amazon) permet de désigner l'industrie qui gère, depuis les États-Unis, l'ensemble des données massives produites par; d'une part, les réseaux sociaux et, d'autre part, l'industrie marchande dématérialisée (que ce soit le *cloud* ou la vente de biens à distance par le biais des réseaux numériques). Au cœur des stratégies marchandes de cette industrie se trouve la question de l'identification des biens marchands et de leur représentation, quelle que soit la culture ou la langue du client potentiel. Cette stratégie commerciale implique la construction d'une ontologie encyclopédique mise au service des moteurs de recherches pour faciliter la découverte de nouvelles connaissances productrice du développement du marché. Cet encyclopédisme marchand est illustré par le traitement statistique de la polysémie : la recherche *bière* à travers les GAFA renvoie systématiquement à la boisson, jamais au cercueil. La raison s'explique par l'usage statistiquement massif qui est fait de ce mot en français pour désigner la boisson, mais également par la nécessité commerciale. Or, un moteur de recherche qui illustre, par ses articles, la diversité des choses dénotées est tantôt un catalogue, tantôt une encyclopédie.

finissent de se poser les questions de relations cognitives : par principe, le cétacé n'est pas un poisson, mais le moteur qui ne renverrait pas une occurrence de *baleine* serait considéré comme étant particulièrement imparfait. Le champ des asémantèmes connotés, comme *Moby Dick*, ne manque pas de poser le problème des listes finies accessibles en ressources (Ducos, Salvador, 2011). En regard de cette problématique passionnante en soi, se pose la question du champ en contrepartie des informations figurées portées par le corpus ; ainsi un énoncé tel *une grosse baleine m'empêchait de sortir de la rame du métro* et les métaphores filées qui en découlent devraient-ils être interprétés indéfiniment par le portail qui, faisant la charnière entre la donnée encodée et l'utilisateur final dans un processus de co-création, doit embarquer les moyens d'épurer la donnée publiée. Mais, une fois ce problème résolu, ce qui est faisable, se pose la question d'une recherche sur le mot *article*, entendu par le linguiste comme « l'ensemble des occurrences des parties du discours nommées article » là où le moteur précédemment décrit supposer : « Ensemble des occurrences du substantif, de ses synonymes, des mots de la famille »⁹ (Lin, 1998). Le regard posé par le terminologue sur le corpus, le prisme de la recherche documentaire, met en défaut toutes les stratégies d'élaboration de ressources documentaires exhaustives ou, toutefois, met en lumière la nécessité de redimensionner un corpus en fonction des attentes de l'observateur. Au cœur de cette problématique, se pose la question de notre capacité à penser la donnée et plus particulièrement son *adressage*¹⁰ (w3c, 1997). Dans l'exemple précédent, où commence la métaphore de la baleine et où se termine-t-elle ? Et quels sont les moyens que la ressource URL offre pour déterminer la profondeur de la donnée et résoudre les inférences ? Cette question n'est donc pas liée à la nature de l'hypertexte telle qu'on la redéfinit sans cesse depuis plus de 60 ans¹¹, mais aux questionnements littéraires et linguistiques de la résolution du sens, identifié avec le contenu.

Pour une définition du « plasmé numérique »

À l'heure actuelle, s'il semble envisageable de copier ou reproduire fidèlement une statue, une œuvre picturale voire sonore, on peine à se figurer la façon dont les spécialistes du texte se représentent la construction numérique fidèle d'un manuscrit du XIII^e siècle par exemple. En effet, la nature de la numérisation

⁹ La question du traitement statistique de la polysémie dérive de l'analyse des réseaux de similarités tels qu'analysés par D. Lin (1998 : 305) : « La mesure de la similarité n'est pas calculée comme dans les définitions précédentes, mais elle découle plutôt d'un ensemble de suppositions » (« *The similarity measure is not directly stated as in earlier definitions, rather, it is derived from a set of assumptions* »).

¹⁰ Par *adressage des données*, il faut entendre ici la question posée par la dématérialisation documentaire qui renouvelle nombre des problématiques issues de la construction des catalogues bibliothécaires. Quel formalisme adopter pour à la fois cibler et accéder au contenu du document (Jacquemin, Zweigenbaum, 2000) ?

¹¹ Il est sans doute important de rappeler que le mot *hypertext* est introduit dans la langue américaine vers 1965 par T. Nelson. Sur ce sujet, voir J.-P. Balpe (1990) et R. Laufer et D. Scavetta (1992).

change selon *ce que l'enquêteur attend*. Beaucoup de chercheurs considèrent que la numérisation est une opération de scanner, de production de texte pour ainsi dire noir sur blanc. Mais que fait-on de la matière de l'ouvrage, de sa résistance, des glus employées pour le fabriquer, du grammage de la feuille, des accidents qui le composent, des traces et des marques qui en altèrent la matière profonde ? Autant d'informations qui peuvent ne pas intéresser l'historien du texte, ou le linguiste, mais qui relèvent pourtant de l'opération de numérisation pour l'archéologue. La numérisation est donc une opération incomplète. Cela signifie que la photographie d'un texte, indépendamment des données de l'objet, satisfait, pour une partie du lectorat et de la communauté des chercheurs, aux critères de représentation de « l'objet texte » qui est le contenu du document. Cela renseigne sur le fait que le manuscrit numérisé est perçu par la majorité des lecteurs comme le recueil, la boîte qui abrite un seul trésor : un texte, noir et blanc.

Il semble que la recherche et le monde de l'édition vivent aujourd'hui sur un consensus construit autour d'une grande métaphore, celle qui fait du document numérique une évolution naturelle du document autrefois manuscrit. Pour reprendre l'heureuse expression de Marie Després-Lonnet (2009 : 19), cette image culturelle « entretient de fait la confusion entre capacité technique de "rendre accessible" [sic] des données, par leur "mise en ligne" et accès aux savoirs, c'est-à-dire, la possibilité effective donnée à un individu de s'approprier de nouvelles connaissances ». Tout se passe comme si la notion d'auteur¹² n'avait pas évolué autrement que technologiquement, le clavier étant le comparé et la plume d'oie, le comparant. Or, la réalité est tout autre, mouvante et protéiforme. L'idée même de l'originalité de l'œuvre est battue en brèche par la matérialité des opérations de transferts et de téléchargements, voire par la réalité des états électroniques de la chimie du silicium qui est la véritable « encre » de production documentaire. Cette chimère intellectuelle partage des ressemblances avec les littératures orales en ce sens qu'elle se dissémine à des vitesses vertigineuses.

Il existe deux types de documents numériques, le document numérisé et le document nativement numérique. Le premier ne se définit pas uniquement comme une image du texte disponible, ce à quoi se résument trop souvent les campagnes de numérisation du patrimoine des bibliothèques organisées par les entreprises comme les collections nationales, ce qui conduira inévitablement, à plus ou moins longue échéance, à devoir recommencer en même temps que progresseront les techniques de numérisation et les formats de stockage. Il ne se résume pas non plus à un document texte issu de l'édition diplomatique d'un manuscrit ou d'un ouvrage ancien. En effet, la numérisation du contenu implique un ensemble de choix (le jeu de caractères, l'encodage, le format de sortie) qui, ne reflétant pas l'objet d'origine, en altèrent la qualité de reproduction. Ceci étant, il ne s'agit pas de dire que le texte

¹² Le domaine d'intérêt majeur « Sciences du Texte et Connaissances Nouvelles » (accès : <http://www.humanitesnumeriques.fr>) promu par la région Île-de-France cherche à encourager les méthodes d'investigation dans le domaine de la philologie numérique. Au cœur de ses préoccupations, se trouve la question de l'auctorialité (accès : <http://www.projetprada.fr>).

issu d'une campagne de numérisation est sans valeur ; mais, à l'heure de questionner l'objet « document numérique », il est bon de rappeler que l'on ne peut réduire une œuvre à un état de texte arbitrairement choisi. Ce dernier n'étant qu'un état parmi d'autres de l'édition d'un cycle du texte, qui a valeur documentaire mais qui n'épuise pas la puissance de l'objet numérique. De la même façon que le philologue est habitué à penser le cycle de vie d'un manuscrit médiéval au cours de la production duquel interviennent un auteur premier, des copistes et un éditeur moderne – qui finit souvent par apparaître comme l'auteur de l'œuvre médiévale –, l'investigateur du texte numérique doit s'habituer à penser l'infinie variation subtile du document au cours de la production duquel interviennent l'auteur, les constructeurs de la ressource (comme l'URL, les couches interprétatives de production automatique du document, les navigateurs qui modifient l'apparence – c'est-à-dire la mise en page – du texte lu), voire les lecteurs eux-mêmes.

Il faut questionner la notion de « document numérique ». Avant tout, il s'agit d'un fichier dont la fonction est de permettre l'exploitation d'un contenu informatif. Autrement dit, il existe de nombreux fichiers logiques (c'est-à-dire contenant des données logicielles) et, parmi ces derniers, certains ont pour « intention première » d'être exécutés par un *shell*¹³, d'autres de stocker un ensemble de données. Les documents font partie de ces derniers. Parmi eux, certains n'ont jamais d'existence autrement que par l'assistance d'un outil logique (un logiciel), que ce soit le traitement de texte ou le navigateur internet. Par document nativement numérique, nous désignons donc également une collection d'informations issues de la sérialisation¹⁴ pour le stockage puis la consultation. Les instructions qu'il contient sont destinées à être interprétées par des outils logiques et ce fichier, indépendamment de sa plateforme de création (traitement de texte, bloc-notes), n'existe pas sans assistance logicielle. Nous nous situons en quelque sorte en amont mais proche de ce que Marie-Anne Paveau (2015) désigne comme un « document numérisé », par opposition au « document numérique », mais sans intégrer la dimension de « fabrique numérique » créative du « texte » puisque l'appellation « document numérique » telle que nous l'entendons recouvre également les fichiers *swaps* ou temporaires produits par les serveurs de base de données, les fichiers binaires des index de ces mêmes serveurs dont le contenu est fragmentaire mais qui, une fois réassemblés, produisent un sens qui répond à l'interrogation formulée par une série de requête par et pour le lecteur. En circulant sur les réseaux, le document se fragmente au fur et à mesure qu'il est exploité : cela signifie que plus le temps passe, plus se multiplient des instances du document sur les postes clients qui ne ressemblent pas forcément à l'image conçue par son auteur. Comment se représente-t-on l'envergure d'un document numérique ? Si l'on prend le problème en extension, nous sommes bien en peine d'en déterminer le commencement et la fin. À l'idée qu'il commencerait au premier

¹³ Le « *shell* » est l'interpréteur de commande du système, à savoir la première couche interprétative d'un système d'exploitation accessible en ligne de commande.

¹⁴ Le terme *sérialisation* désigne l'« opération informatique par laquelle on traduit une information en suite élémentaire d'informations la composant en vue de la sauvegarder ou de l'acheminer *via* un réseau vers un destinataire » (Salvador, 2016 : 173).

offset du premier octet encodant le premier caractère et s'achèverait au dernier *offset* du dernier octet, s'oppose l'idée qu'un document numérique est sempiternellement éditable et que chaque possesseur d'un document peut, à n'importe quel moment, l'enrichir, l'annoter, le modifier, en changer la mise en forme *sans que l'unité de l'œuvre première déposée en un autre endroit ne soit jamais atteinte* : la question génétique de l'original premier, ou de l'authentique, est à ce compte impertinente. Le texte continue de vivre longtemps par-delà la volonté de l'auteur premier, et nulle analyse graphique, nulle enquête génétique ne permettra jamais d'en détecter les origines, ni d'en restituer l'intention première. Le travail de l'éditeur du texte ancien, qui posait couverture et quatrième, avait aussi pour mission de baliser le commencement et la fin de l'œuvre, d'en figer une version. Mais la discontinuité du travail numérique a depuis longtemps battu en brèche l'idée selon laquelle le document numérique « était comme » un document imprimé.

La solution envisagée par certains auteurs d'œuvres mouvantes, comme les contributeurs du *Dictionnaire du moyen français*¹⁵ numérique avant 2015, était intéressante car elle consistait à produire une version électronique d'un fichier au format *Portable Document Format* (PDF) du dictionnaire qui signalait en quelque sorte un état, une version de l'œuvre en cours. Mais, cette confiance dans le document numérique de type image, PostScript, ne résout en rien la question de la constante variance du document numérique : personne ne conteste l'idée qu'il y a une stratification des versions du document, mais la longue succession des versions n'existe jamais dans l'univers numérique où chaque version remplace la précédente. À cela s'ajoute l'idée suivante : de fait, chaque opération de consultation d'un document numérique est une opération de téléchargement. En effet, toute consultation d'un URL entraîne la duplication du document par le logiciel serveur (*Apache, Jetty, NGinx* par exemple) sur le poste client. Lorsque l'on « clique » sur le lien qui conduit à <http://www.crealscience.fr/DFSM/consult/ZIRBE>¹⁶, le document, la fiche intitulée « ZIRBE », fort heureusement, ne disparaît pas du serveur, mais est dupliquée sur le poste client pour être reconstruite par le navigateur web en fonction des normes et conventions établies entre les acteurs du web et les éditeurs de logiciels au sein du World Wide Web Consortium (w3c). Il n'y a pas de « prêt » du document en attente de restitution par le lecteur et cette opération conduit évidemment au clonage sur le poste client des données du serveur : autrement dit, un document original n'existe jamais ; les modifications apportées à un document ne sont répercutées sur les postes clients qu'à partir du moment où le client lui-même en fait la demande et deux versions absolument différentes du même document peuvent coexister sur différentes machines *en même temps*. En effet, une URL est à la fois la cote bibliothécaire d'un document et

¹⁵ Initié par Robert Martin, le *Dictionnaire du moyen français* est un dictionnaire numérique en ligne. Chaque année, l'unité de recherche Analyse et traitement informatique de la langue française (Atilf, CNRS, Université de Lorraine) produit une version XML téléchargeable qui marque l'état de l'existence des connaissances en ligne à une date donnée. Accès : <http://www.atilf.fr/dmf/>.

¹⁶ Le terme *zirbe* (ou *zurbe*, sa variante allomorphe) désigne une partie de l'appareil digestif composée de l'intestin et du grand épiploon. Le projet « Crealscience » (financé par l'Agence nationale de la recherche) est un dictionnaire en ligne des sciences et techniques médiévales.

son *locator*, c'est-à-dire le moyen même d'en obtenir un duplicata. Et si l'on prend le problème en extension, il convient d'aborder la question des états électroniques qui trouvent une représentation à un moment donné sur les écrans ou les imprimantes des clients. Le traitement de texte qui, majoritairement, produit des documents au formalisme XML¹⁷ associés à une feuille de style et un navigateur d'un genre un peu particulier (*Word*, *OpenOffice*), masque l'existence de la réalité numérique du document textuel qui se résume à une représentation graphique (la police de caractère) d'une réalité électronique (les *bits*) encodée et décodée selon des règles d'affectation d'un groupe de *bits* (« octets ») à une table de glyphes (Unicode, Organisation internationale de normalisation – iso –, etc.). Ce que nous appelons « document numérique » n'est en fait qu'un fichier électronique trouvant son chemin jusqu'à notre intelligence par le biais de mécanismes fondés sur la chimie du silicium et dont nous tirons parti pour construire des représentations qui rejoignent des usages auxquels nous sommes habitués (le codex, l'image, le poster). Mais comment se figurer la stabilité de la représentation glyphique d'un objet « texte numérique » dans la mesure où cette illusion de donnée textuelle n'existe jamais véritablement qu'au moment où il est ouvert sur une machine cliente dont le texte est dépendant de la capacité de cette machine à « reconstruire » l'image la plus fidèle possible à l'original, étant entendu que cette image n'a aucune existence autrement que dans la description conventionnelle qu'en a produit l'auteur grâce à des formalismes qui en définissent le format (css, svg) ? Cette mouvance caractéristique du document numérique a modifié les habitudes de lecture. La démultiplication des documents numériques, leur peu de valeur aussi, ont conduit à rendre accessible non plus l'œuvre, mais le fragment *pertinent* en fonction de la demande du lecteur. On ne lit plus une œuvre, on en tire une donnée dont l'adresse est un paramètre du plume numérique, c'est-à-dire du recueil permanent d'informations accessibles, transformées et codées pour être assimilées par le client terminal. Ce dernier rendu accessible par la consultation de fichiers textes, de documents sonores ou vidéographiques perceptibles mis en ligne pour être questionnés.

Adresse macrostructurale

Qu'entend-on par « adressage » ? Il s'agit de l'ensemble des moyens de localiser et, en même temps, de consulter une donnée documentaire pertinente dont la consultation répond à la question posée par le lecteur. Cette problématique recouvre la notion d'*Uniform Resource Identifier* (URI)¹⁸ en quelque sorte, mais sortie de la

¹⁷ Les formats populaires de la suite Microsoft Office « docx » ou « odt » sont des documents respectant le formalisme XML au sein de l'espace de nommage OASIS (accès : <https://dtd.nlm.nih.gov/options/OASIS/tag-library/19990315/n-rk20.html>).

¹⁸ Le nom URI est une courte chaîne de caractères qui permet de cibler une ressource spécifique. Il existe deux types d'URI, celles de type *locator* (URL) et celles de type *name* (URN). L'URL permet d'obtenir la donnée souhaitée. C'est à la fois la donnée elle-même et son adresse ; elle permet donc d'identifier et de consulter la donnée. Lorsque vous tapez une adresse dans votre navigateur, vous consultez une ressource

logique documentaire (Fondin, 1991) pour être appliquée à l'analyse du corpus et à l'exploitation de ressources informatiques à travers le plasmé numérique du *big data*. Pourquoi parler d'adressage de la donnée ? La logique documentaire qui soutient la persistance de ce plasmé n'est pas éloignée du prêt documentaire des archives et bibliothèques que nous connaissons. À la différence que, lorsque nous ouvrons un livre, nous savons que nous débutons une histoire, un récit, un essai, une argumentation et il n'est pas rare que nous naviguions dans cette lecture de manière transversale pour parcourir la somme des informations contenues jusqu'à parvenir à la donnée recherchée qui contribue à élucider le problème que nous avons soulevé. Lorsque nous fouillons une donnée textuelle numérique, nous faisons abstraction de l'unité pour construire une représentation fragmentée des données que nous prétendons extraire indépendamment de la cohérence de l'œuvre que nous asservissons aux objectifs de notre recherche. Où se trouve donc l'information cherchée ?

Un premier élément de réponse revient donc à penser la problématique de l'URL dans la logique de la relation client-serveur. Lorsque l'on consulte une donnée *via* son *locator* (URL), on émet une requête qui repose sur la construction d'une adresse et le téléchargement d'un document. Ainsi, l'adresse <http://www.crealscience.fr/zirbe> (voir *supra*) doit-elle se lire de la façon suivante : je cherche, à travers un protocole de téléchargement (HTTP : *http://*) qui renvoie par défaut à la réalité électronique du port 80 de la communication Ethernet à obtenir une représentation documentaire d'une ressource nommée « zirbe » archivée sur une machine nommée *www* (mais ce pourrait être n'importe quel nom) au sein du domaine *crealscience* de l'espace reconnu comme *.fr* au sein du consortium du web. L'adresse de la machine hébergeant la ressource dont nous souhaitons obtenir une représentation (IP : 46.105.99.7) permet également d'accéder à la ressource, de sorte que *46.105.99.7:80/zirbe* est une expression synonyme de la précédente. Tout se passe comme si l'expression *46.105.99.7:80* désignait le lieu de stockage de l'information, la bibliothèque au sens courant du terme. L'archivage du document, et le chemin qui y mène à travers l'arborescence du disque dur relatif au répertoire du logiciel serveur de la machine 46.105.99.7 constitue un chemin vers le fichier-ressource, ce qui se traduit ainsi : « Sur la machine nommée 46.105.99.7, via le

accessible. C'est une URL, deux-en-un. Lorsque vous obtenez la référence d'un ouvrage en bibliothèque, vous obtenez une localisation où vous devez ensuite aller chercher le document. L'URN est un nom abstrait qui permet de nommer une ressource, mais pas de la consulter. Toutefois, l'URN est signifiante dans un espace de noms et renvoie à une réalité. Par exemple, la référence *International Standard Book Number* (ISBN) d'un ouvrage, ou la cote documentaire d'une archive sont des URN. Sur le design des URI, voir notamment T. Berners-Lee (1998) : « La date de création du document — la date à laquelle l'URI est republiée — est une chose qui ne changera jamais. C'est quelque chose de très utile pour distinguer les requêtes qui se conforment à un système récent de celles qui se conforment à un système plus ancien. C'est une bonne chose de commencer une URI avec cela présent à l'esprit. Si un document est daté de n'importe quelle manière, même si c'est un document d'un grand intérêt pour toutes les générations à venir, la date est un excellent point de départ » (nous traduisons, « *The creation date of the document — the date the URI is issued — is one thing which will not change. It is very useful for separating requests which use a new system from those which use an old system. That is one thing with which it is good to start a URI. If a document is in any way dated, even though it will be of interest for generations, then the date is a good starter* »).

port 80, je cherche à obtenir une représentation de la ressource accessible dans l'arborescence du disque dur à l'adresse `/var/www/vhosts/crealscience.fr/zirbe` ».

En l'occurrence, pourquoi parler de « représentation de la ressource nommée » et pas simplement de document numérique accessible en ligne ? La ressource est souvent un document stocké sur la machine elle-même, stocké ou généré au vol par des langages de *scripting*¹⁹ comme PHP ou Python²⁰ qui « brodent », sur un canevas préconstruit par l'auteur, un ensemble de données qui sont physiquement téléchargées par la machine cliente et reconstruites par le navigateur comme une image acceptable en fonction des paramètres du formulaire du document prototypique dont elles émanent. Dès lors, la notion de représentation semble contribuer à une complexification inutile. Il en va pourtant autrement, pour trois raisons :

- le téléchargement et la reconstruction d'une ressource par le navigateur impliquent sa multiplication. Ne consultant jamais le document original, toute requête web duplique la ressource pour la rendre accessible et construit donc une représentation de la source en la clonant ;
- la consultation du duplicata *via* l'URL entraîne la multiplication d'images de la source qui peuvent exister sous des formes différentes sur différents postes clients au même moment. Sur le poste client, chaque image, est donc une représentation donnée à voir au client qui ne vaut que pour lui pour une durée indéterminée ;
- la troisième raison est liée à l'existence d'une philosophie web, largement exploitée, dénommée *Representational State Transfer* (REST) qui pousse encore plus loin la démarche de fragmentation du lien URL-document. Pour reprendre les conclusions d'Alexandre Monnin (2014 : 38-39), nous dirons avec lui que, au problème de l'actualisation des URI soulevé par Tim Berners-Lee (1998),

« La solution proposée par REST (pour la résumer sans prétendre traiter toute la question) est très élégante. Elle part du principe que, au lieu de "fichier", "document", "objet", etc. tout ce que l'on place généralement sous le concept d'URI est une "ressource". Voilà la définition donnée par Fielding et Taylor : "une ressource R est une fonction d'appartenance temporaire²¹ $MR(t)$ qui, pour l'instant t , correspond à un ensemble d'entités ou de valeurs équivalentes. Les valeurs dans l'ensemble peuvent être des représentations de ressources et/ou des identificateurs de ressources. Une ressource peut renvoyer à un ensemble vide, ce qui autorise une référence à une ressource avant même qu'elle n'existe, chose qui était étrangère à toute logique de relation hypertextuelle avant l'avènement du web" (Fielding et Taylor 2002, 125) » (nous traduisons)²².

¹⁹ Un langage de script est un langage de programmation permettant de manipuler les fonctionnalités d'un système informatique sans passer par son interface utilisateur (ce qui permet notamment d'utiliser des fonctionnalités de ce système non accessibles *via* son interface ou d'automatiser des tâches).

²⁰ Accès : <https://www.python.org/>.

²¹ Dans la logique floue de la théorie des ensembles, la « fonction d'appartenance » définit ou non l'appartenance d'un objet x à un ensemble E . $f(x) = 1$ si et seulement si $x \in E$. La fonction est dite « temporaire » quand elle reçoit comme paramètre un instant t . La fonction teste l'appartenance de x à E à l'instant t .

²² « The solution advocated in REST (in a nutshell, without exhausting REST's significance for the Web) is a very elegant one. It states that instead of files, documents, objects, and so forth, what is being referred to by a

L'idée forte est la suivante : au lieu d'être en quelque sorte attaché à un espace disque dont dépend sa gestion (droits d'écriture, stockage, place), si le serveur logiciel considèrerait chaque URL non comme une adresse physique, mais comme un appel à une fonction programmée. À chaque URL, on ferait correspondre non pas un fichier, mais une fonction logique d'un programme développé dans un langage choisi. Ceci correspond à un style d'architecture de service pour les réseaux qui a été développé par un étudiant, Roy Fielding (2000), dans un chapitre de sa thèse de doctorat. Assez innovante, l'idée permet de libérer le serveur des contraintes en droits d'écriture qui pesaient autrefois sur les logiciels serveurs. Dans cette logique, REST, l'URL renvoie à un nombre infini de représentations de ressources liées à la formation de l'URL elle-même. L'URL devient co-constructrice de la ressource, un peu comme si le bibliothécaire de la municipalité recopiait pour vous dans le livre cherché les pages qu'il estime que vous devez lire. Plus un livre, une œuvre, un fichier, un document : un extrait.

Chemin microstructural

La puissance de la logique REST ouvre des portes passionnantes pour le concepteur de ressources numériques, puisque toute adresse d'un portail web pointe désormais vers une fonction algorithmique qui altère ou modifie la représentation de la ressource. Voilà pourquoi il est nécessaire de parler de représentation de la ressource. Cela étant, au cœur du document numérique, érigé au rang de ressource, articulé *via* un serveur REST à un ensemble de fonctions publiées sur un portail pour être consultées, on en vient à poser la question de l'adresse microstructurale de la « micro unité » (Cotte, 2011 : 59). C'est probablement à ce niveau que se situe l'intervention des *markups languages* dans la chaîne de production de la représentation de la ressource initiale. Ces formalismes, au premier rang desquels XML, permettent de construire des représentations stratégiques des lieux d'information dans un flux textuel. XML, dont nous traiterons dorénavant spécifiquement, repose sur une écriture conventionnelle qui consiste à inscrire entre chevrons ce qui est de l'ordre de la communication métatextuelle (on parle de balises) et d'exclure des chevrons ce qui est de l'ordre de l'information principale (le contenu textuel informationnel du document lui-même). En se dotant d'un typage simple de l'entité « balise » (à des balises ouvrantes correspondent des balises fermantes) et en se fondant sur une description fine de cette entité nommée et de ses attributs, l'insertion de marqueurs dans le flux textuel autorise la limitation d'empans textuels nommés et caractérisés faisant surgir une ressource informative, un corpus au second degré, dont l'importance va de soi pour peu que la ressource respecte un cahier des charges dont l'uniformité est essentielle à sa consultation.

URI is a "resource." Here's the definition given in REST (by Fielding and Taylor): "A resource R is a temporally varying membership function $M_R(t)$, which for time t maps to a set of entities, or values, which are equivalent. The values in the set may be resource representations and/or resource identifiers. A resource can map to the empty set, which allows references to be made to a concept before any realization of that concept exists— notion that was foreign to most hypertext systems prior to the Web" (Fielding and Taylor 2002, 125) ».

Qui plus est, en matière de représentation de la ressource, XML s'impose comme un standard, malgré les nombreuses critiques qui lui sont adressées²³, pour la simple et bonne raison que le W3C a systématiquement développé les nouvelles ressources du web en se basant sur des stratégies de développement fondées sur la dialectalisation XML de certains secteurs de l'édition numérique. Par exemple, XHTML est un dialecte XML dans lequel certains noms réservés ont une portée instructive²⁴. Mais il en est de même pour SVG, par exemple, qui est un dialecte d'XML consacré à la représentation graphique sur *canvas*²⁵ en HTML5 où chaque nom de balise a une portée instructive (« *circle* » n'est pas un nom, mais une instruction dont la conséquence est le dessin d'un cercle de rayon égal à la valeur de l'attribut « *r* », etc.) ; MathML, un dialecte consacré à la représentation des équations mathématiques ; epub un format compressé (*zip*)²⁶ correspondant à une exploitation adroite d'XHTML.

La définition d'empan textuel répond en partie à la question du lieu de la donnée (Salvador, 2016 : 157) : elle se trouve, pourrait-on dire de manière tautologique, là où elle est désignée dans la ressource instruite par le serveur. Mais cette représentation du sens n'existe que dans la mesure où elle intègre un cahier des charges commun à un ensemble de ressources qui s'accordent pour nommer, désigner des lieux de données selon un formalisme et un objectif communs. En effet, la question de l'adresse est liée à la dénomination des lieux de données. Or, le caractère laxiste d'un formalisme comme XML²⁷, et la possibilité de laisser flotter la rédaction du métacorpus au fur et à mesure de l'ingéniosité des opérateurs de saisie laissent présager de grandes dérives dans la construction des répertoires d'adresses des lieux de données désignés par le balisage XML. C'est là qu'entre en jeu un double système de représentation de l'information qui concourt à la sauvegarde de l'accessibilité.

²³ Ici, il est fait allusion aux critiques récurrentes du caractère « bavard » (*verbose*) d'XML. Pour un aperçu du débat qui agite la communauté internet, voir J. Bone (2002).

²⁴ Par exemple, le mot *div* renvoie à une « division » d'un espace de texte. Il n'est pas question pour aucun créateur de contenus de décider arbitrairement que *div* désigne une « division » sportive ou militaire par exemple. Le sens est préconstruit et a une portée instructive, c'est-à-dire qu'il est « interprété » par le navigateur comme ayant un sens.

²⁵ Le terme *canvas* désigne un composant de HTML (voir *infra*) permettant d'afficher des images, graphes, animations ou jeux vidéo par le truchement de scripts.

²⁶ La compression de données est l'opération informatique consistant à transformer une suite de bits A en une suite de bits B plus courte pouvant restituer les mêmes informations, ou des informations voisines, après utilisation d'un algorithme de décompression. C'est une opération de codage qui diminue la taille (de transmission, de stockage) des données au prix d'un travail de décompression, cette dernière étant l'opération inverse de la compression. Zip est un format de fichier permettant l'archivage (utilisation d'un seul fichier pour en stocker plusieurs) et la compression de données.

²⁷ XML peut être dit laxiste, car le sens dénoté par le recours aux balises dans la structuration de la donnée repose sur l'interprétation, sur le « vouloir-dire », du constructeur. Sans contraintes extérieures appliquées par des systèmes de coercition comme les schémas XSD, la structuration du document peut varier infiniment au gré des changements de constructeurs sans que rien ni personne n'y puissent rien. XML n'est que symbolisme.

Dans un premier temps, la pensée du cahier des charges à travers des constructeurs de prototypes (DTD, XSD, RELAX NG, etc.) autorise la mise en place d'outils de coercition qui exercent sur la saisie des données une contrainte à partir d'un schéma prototypique embarquant de manière programmatique le typage des données représentables, ce dernier consistant en la description de la nature des informations portées par un élément XML (date, prix, lieu, chaîne de caractères de longueur N, nombre entier, irréal...). Plus le typage des données est fin, plus il est possible d'exercer une contrainte forte sur la constitution d'une ressource. Les types de données sont portés par des descripteurs dont la liste est finie, et longue. En appliquant la logique de la structuration informative et du typage à l'élaboration du métacorpus informatif, on parvient progressivement à concevoir une topographie fine de la donnée au sein du document.

Dans un second temps, la logique du chemin vers la donnée (xpath) permet de nommer concrètement l'adresse de la donnée au sein d'un corpus XML. Pour n'en dire qu'un mot, l'adresse (ou chemin) xpath d'un lieu de donnée repose sur un symbolisme simple où chaque branche d'une arborescence du corpus est séparée de sa descendance par un signe « / » qui symbolise un nœud, un embranchement. Il existe de nombreuses nuances dans l'écriture au sein de ce système, amplement décrite par ailleurs. Ainsi est-il possible d'écrire que, dans le fichier /var/www/index.xml, il existe une information dont le chemin est /bdd/texte/p[1]/text(), c'est-à-dire qu'une entité de nature « texte » contenue dans le premier élément <p/> de tout élément <texte/> (il peut y en avoir plusieurs) de la racine nommée <bdd/>. Le fonctionnement conjoint du prototypage de la donnée et des moyens de l'adressage collaborent à l'édification d'un mécanisme d'abstraction qui est adapté à la désignation de l'adresse d'une donnée au sein d'un texte plus ample. On constate aisément ici que la chaîne doc('/var/www/index.xml')/bdd/texte/p[1]/text() traverse en quelque sorte la frontière documentaire (l'œuvre) pour accéder au lieu d'une donnée précise (un mot, une expression, un caractère) en son sein. Ce lieu présente la particularité d'être nommée, c'est-à-dire de faire sens en entrant dans une démarche réflexive autour de « la langue de la métalangue » (Portine, 1997 : 36). La stratification que nous venons de décrire dresse également un portrait de la circulation de la représentation de la ressource :

```
[client -[URN/URL]-> serveur (REST)
-> ressource XML
<- contrainte XSD/DTD
<- adressage xpath]
```

Mais cette relation dynamique de la consultation vers un fragment appelé « textes » conçu comme une « chaîne de caractères » ne rend pas compte de l'implantation de toute ressource au sein du plasmé numérique qui est une forêt d'arbres très denses au sein de laquelle cet objet variable « Texte » peut exister un nombre incalculable de fois selon les langues, les dates, le genre, etc. Pour que la représentation de la donnée puisse devenir un objet pensé, il faut admettre que la standardisation de la production de données est une étape cruciale vers la possibilité de l'interopérabilité des données linguistiques. Cette pensée de l'uniformisation de la conceptualisation du texte est au

cœur de la problématique des opérations de mise en commun de lexiques standards pour la conception des prototypes globalisés. Ces initiatives (comme la TEI), nées à la suite des principes de Poughkeepsie²⁸, ne sont en fait que des lexiques visant à uniformiser les mots de la dénomination des lieux de données hors de tout texte. Leur existence ouvre la porte vers la synthèse des ressources textuelles, dans la majorité des cas²⁹.

Macrocollection de métadonnées

Le chercheur contemporain doit non seulement produire de la connaissance pertinente (Després-Lonnet, 2006), mais aussi s'assurer de l'intégration de cette connaissance dans une économie des échanges. Imaginons qu'il s'agisse tout simplement de déposer soi-même un ouvrage rédigé dans une bibliothèque nationale. La ressource numérique est l'ouvrage ; le web en est forcément la Bibliothèque numérique. Si une chaîne de caractères telle `doc('lvar/www/index.xml')/bdd/texte/p[1]/text()` symbolise, dans le contexte d'une arborescence connue, un élément « de l'intérieur » du document, et si l'on admet dans le même temps qu'une fonction REST du type `mySpaceName:function($text){ for $x in doc('lvar/www/index.xml') return <html>{$x/bdd/texte/p[1]/text()}</html>}` (c'est-à-dire une fonction rattachée à un espace de nommage XML du nom de `mySpaceName`, qui a pour nom `function` et qui reçoit un ensemble de données variables appelé `$text` produit une extraction au format HTML d'un sous-ensemble de `$texte`) est associée à un chemin du type `%rest :path('showText[$text]')` (c'est-à-dire, dans la logique REST, une URL variable du type `http://monserveur.fr/showText('monCorpus')` liée à la fonction précédente) qui peut être exprimé par l'URL `http://myDomain/showText['myRequest']`, alors on comprend que la donnée textuelle s'affranchit pleinement des limites imposées non seulement par le cadre du document, mais encore par le serveur lui-même et par l'ensemble des contraintes physiques de stockage et de représentation qui s'exercent sur toute donnée textuelle. Il est possible d'intégrer le fragment textuel dans une économie d'échanges plus vaste qui laisse entrevoir une fusion des sources de connaissances. Une façon d'envisager cette solution consisterait à proposer un modèle de métadonnées susceptibles de favoriser ces échanges et de construire des passerelles entre les sources et les données. Le w3c a établi une série de préconisations pour construire un modèle de graphes permettant de décrire les ressources attachées

²⁸ Moment fondateur de la *Text Encoding Initiative* (TEI), en « 1987, près de New York dans la ville de Poughkeepsie, une conférence internationale réunissant l'agence fédérale américaine et l'UE, produit un manifeste définissant les règles [de description d'un document]. Depuis cette date, le Steering Committee and Editors [de la TEI] est une société qui s'est constituée en différents comités (pour la documentation, pour la représentation du texte, pour l'analyse du texte, pour les métadéscripteurs auxquels beaucoup de projets s'agrègent, souvent sans vraiment en cerner les enjeux). Il perçoit ses financements de la Fondation Mellon, de la Commission Européenne et de l'agence fédérale américaine (dont les représentants composent le Committee) et a obtenu des partenariats avec des associations savantes en Humanités Numériques » (Salvador, 2016 : 69).

²⁹ Nous avons ailleurs évoqué les ouvrages et toutes les entreprises littéraires qui résistent à ces tentatives de codification (Salvador, 2016 : 70-72).

³⁰ L'exemple emprunte son formalisme à la syntaxe xquery.

à une URI, et leurs métadonnées, pour en permettre la collecte par des robots. Le formalisme de sérialisation de ces données le plus exploité est un protocole commun de description de ressources : *Resource Description Framework* (RDF)³¹. Chaque nœud du graphe RDF est une valeur. Les nœuds sont liés par des arcs, qui prédisent une information sur les nœuds. L'URN d'une fiche de dictionnaire (*dict*) pour l'identifiant (*id*) d'un mot comme *champignon* est `[[urn:dict:15963]]`—>*champignon*.

Pour un rédacteur, cette relation est signifiante, mais elle ne veut rien dire en dehors de l'espace de relation (l'arc) au sein duquel cette relation est vraie. Pour compléter l'information, on doit préciser que cette assertion est vraie à l'adresse (le *locator* de l'URL) `http://www.dict.fr/champignon`. La matrice d'information qui permet de lier l'identifiant précédent à l'URL est donc `[[urn:id:15963]]`, `http://www.crealscience.fr/xylocassie,XYLOCASSIE`. En combinant la série d'informations liées aux identifiants d'un dictionnaire, on constitue une collection qui modélise en quelque sorte la profondeur de la ressource accessible sur le web. On obtient un tableau d'entrées à trois colonnes d'informations descriptives (plus une colonne numériques correspondant au rang) où chaque information correspond à un modèle de données ordonné. Ce catalogue peut à son tour donner lieu à l'écriture d'une URN, que nous pourrions appeler `[[urn:dict]]` basé sur l'identifiant. Pour créer cette ressource, on pourrait alors, en exploitant le mot clé *Bag*, élaborer une ressource `<Bag RDF:about="[[urn:creal]]"/>` qui contiendrait par exemple une liste de relations telles celles que nous avons décrites sous forme d'une part d'une liste d'éléments significatifs :

```
<li RDF:ressource="[[urn:dict:vedette]]"/> (un élément de liste li dont l'attribut ressource au sens que ce mot peut avoir au sein du dictionnaire RDF a pour valeur élément identifiant URN).
```

À cette liste, on liera, d'autre part, une description associée à chaque URN et reprenant l'identifiant, le *locator*, et la vedette :

```
<Description RDF:about="[[urn:dict:1]]" dict:id="1" dict:url="http://www.dict.fr/champignon" dict:texte="champignon"/>
```

Si l'on associe maintenant les en-têtes du document descriptif et le document RDF entier au formalisme XML, cela donnerait :

```
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  [xmlns:RDF="http://www.w3.org/1999/02/22-rdf-syntax-ns#"]
  [[xmlns:creal="http://www.dict.fr/#"]]>
  <Bag RDF:about="[[urn :dict]]">
    <li RDF:ressource="[[urn:dict:1]]"/>
  </Bag>
  <Description RDF:about="[[urn:dict:1]]" dict:id="1" dict:url="http://www.dict.fr/champignon" dict:texte="champignon"/>
</RDF>
```

³¹ Pour une définition d'RDF, voir notamment F. Gandon et al. (2002) : « RDF nous permet d'annoter sémantiquement les ressources de la mémoire : il propose un modèle de données simple fournissant un langage de représentation des propriétés et des relations des ressources du Web ». Voir également P. Hayes (2004).

Le dictionnaire n'existe plus en tant que nomenclature ou collection : au contraire, c'est l'existence d'une infinie description de micro-unités liées par un dénominateur commun RDF qui donne sa légitimité à l'objet dictionnaire et, partant, au lexicographe. Le principe fondamental de RDF est de décrire sous forme de matrice le sujet, le prédicat et l'objet d'une ressource, généralement une URI de type *Locator* (ou URL). Il propose une série de vocabulaires qui permettent de décrire des ensembles, eux-mêmes pouvant servir de bases à une série de schémas descriptifs. Ces formalismes, comme OWL (pour la linguistique), SKOS³² (pour les thésaurus) ou encore FOAF³³, s'appliquent à des données qui expriment une organisation du vocabulaire en lexique. Nous entrons alors dans un typage des données documentaires qui, à nouveau, repose sur une série de techniques issues des formalismes d'adressage visant à projeter sur des données textuelles une série de classifications, de nomenclatures ontologiques renouant en quelque sorte avec les logiques KW(.):C mais en poussant la « granularité » (Cotte : 2011) au raffinement de la donnée au cœur même du document ciblé. Il apparaît qu'il est possible d'exploiter les mécanismes d'adressage pour cibler non seulement un document, mais plus précisément un lieu de données (un *topos* comme le pensait la rhétorique) précisément défini dans le corps du document.

Conclusion

Quel est le lieu de la donnée dans les espaces numériques ? De fait, la stratification des réponses recouvre la technologie de la mise à disposition du document. L'URL permet d'identifier et de dupliquer l'objet de la recherche documentaire. Elle opère selon un protocole, sur une machine nommée, dans un répertoire du serveur logiciel, avec l'aide et le conseil de couches interprétatives comme XQuery, SPARKL, PHP ou Python, etc. La combinaison de ces technologies permet la production, le téléchargement et la recomposition de l'original sur le poste client. La représentation numérique des réalités électroniques du document texte se voit doublée d'un ensemble de stratégies qui visent à construire une image du document fidèle à l'image de l'original. Mais le flux discontinu dans lequel s'inscrit la pseudoréalité du document numérique a conduit à inventer des outils pour la structuration fragmentaire des textes numériques. Au cœur de ces enjeux, les *markup languages*, au premier rang desquels XML. Ces langages ont permis de mettre en place une série de mécaniques d'adressage qui a peu à peu rendu possible de produire des éléments efficaces de cotation, de recensement et d'identification des lieux de données. Cette démultiplication des technologies liées au web et à l'édition numérique a progressivement mis au jour la possibilité d'isoler, de reconfigurer et de redimensionner des corpus de grande ampleur en fonction des attentes des consultants par le biais d'exploitations fonctionnelles

³² Accès : <http://www.w3.org/TR/skos-reference/>.

³³ Accès : <http://www.foaf-project.org/>.

des URL qui co-constructrices de la donnée. C'est là le plus grand paradoxe : la donnée n'est accessible qu'à travers l'expression qui la construit et le monde « se redocumentarise³⁴ » (Pédaque, 2006).

Image 1. Le Travail intellectuel (Otlet, 1935)³⁵

³⁴ Selon E. Broudoux et C. Scospi (2011 : 9), « en 2006, l'auteur collectif Roger T. Pédaque traduisait par le terme "redocumentarisation" le phénomène d'une transformation des fonctions documentaires par le processus de numérisation provoquant la révision d'un ordre documentaire antérieur et l'hypothèse d'une entrée dans une "seconde modernité" (Beck, 2001), la "première modernité" étant le résultat du développement de l'imprimé. Cette reconfiguration comporte des enjeux considérables puisqu'elle ne touche "rien de moins en effet, que les artefacts qui nous permettent d'affirmer notre identité, de participer à une communauté, de régler nos transactions ou encore de maintenir notre mémoire sociale" (Pédaque, 2006 : 10) ».

³⁵ « Le Travail intellectuel: Les éléments du monde (A) conçues par les sens (B) et leurs instruments B' et donnant lieu au système des notions ou sciences (C) sont aménagées par la main droite (D) dans la Documentation (E) au moyen duquel il est agi (F) par clavier (G) à distance avec ou sans fil (H) sur les mêmes choses (A) [Cycle] ».

Reste que toute donnée publiée n'est pas forcément consultée, ni justement interprétée : il faut un œil expert pour la déchiffrer. Qui peut comprendre et saisir ce qui se cache derrière la lourde sentence d'un empereur : « Songe aux fleuves de sang où ton bras s'est baigné » (Corneille, 1643, *Cinna*, acte IV, scène 2) ? Pourtant, le texte entier de la tragédie peut être consulté sur le web. Le lieu de la donnée, le lieu de l'inférence reste un enjeu majeur de la description documentaire. C'est sans doute pour cela que nous avons toujours été touché par un dessin de Paul Otlet³⁶ intitulé *Le Travail intellectuel*. Les doigts du personnage représenté sont de longs fils, en même temps des flèches qui le lient à son clavier et, sur la gauche, une série de listes décrit les opérations mentales qui conduisent à la mise en texte de la donnée dans la machine. Cette image représente et symbolise les enjeux de la fouille textuelle lorsqu'elle cherche à présenter au chercheur les lieux d'informations les plus pertinents pour sa recherche. Que l'on pardonne au bibliste-médiéviste cette expression empruntée : le moteur de recherches moissonne finement le grain, mais en définitive seul l'esprit est capable de discriminer le bon grain de l'ivraie interprétative.

Bibliographie

- Balpe J.-P., 1990, *Hyperdocuments, hypertextes, hypermédias*, Paris, Eyrolles.
- Berners-Lee T., 1994a, « Universal Resource Identifiers in www. A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web », memorandum, june. Accès : www.w3.org/Addressing/rfc1630.txt.
- Berners-Lee T., 1994b, « Plenary at www Geneva 94. W3 future directions. Plenary Talk », diapositives. Accès : <http://www.w3.org/Talks/WWW94Tim/>.
- Berners-Lee T., 1998, « Cool URIs don't change ». Accès : <http://www.w3.org/Provider/Style/URI>.
- Bone J., 2002, « Does XML Suck? Revisited ». Accès : http://archive.oreilly.com/pub/post/does_xml_suck_revisited.html.
- Broudoux É., Scopsi C., 2011, « L'enjeu des métadonnées dans un contexte de 'redocumentarisation' », *Études de communication*, 36, pp. 9-22.
- Cotte D., 2011, « Les nouvelles perspectives du web sémantique pour les professionnels de l'information », *Documentaliste-Sciences de l'information*, 48 (4), pp. 58-59.
- Cotte D., 2013, « Qui (re)travaillera la donnée ? », *Documentaliste-Sciences de l'information*, 50 (3), pp. 24-25.
- Cotte D., Després-Lonnet M., 2006, « Information et document numérique : entre métaphore et matérialité », *Sciences de la société*, 68, pp. 80-92.
- Crestadoro J.-B., 1856, *The Art Of Making Catalogues Of Libraries; Or a method to obtain in a short time A most perfect, complete, and satisfactory printed catalogue of the British Museum Library*, Londres, Literary Scientific & Artistic Reference Office.

³⁶ Bibliographe belge, P. Otlet (1868-1944) est, avec H. La Fontaine, à l'origine de la classification décimale universelle et de l'Office international de bibliographie.

- Després-Lonnet M., 2009, « L'écriture numérique du patrimoine, de l'inventaire à l'exposition : les parcours de la base Joconde », *Culture & Musées*, 14, pp. 19-38.
- Després-Lonnet M., 2014, *Temps et lieux de la documentation : transformation des contextes interprétatifs à l'ère d'internet*, mémoire d'habilitation à diriger des recherches en sciences de l'information et de la communication, Université de Lille.
- Doueih M., 2008, *La Grande Conversion numérique*, Paris, Éd. Le Seuil.
- Ducos J., Salvador X.-L., 2011, « Pour un dictionnaire de français scientifique médiéval : le projet Crealscience », *Langages*, 183 (3), pp. 63-74.
- Fielding R.T., 2000, *Architectural Styles and the Design of Network-based Software Architectures*, thèse de doctorat en *information and computer science*, University of California.
- Fielding R.T., 2008, « REST APIs must be hypertext-driven », *Untangled*. Accès : <http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven>.
- Fondin H., éd., 1991, *Bibliographie sur l'informatique documentaire : outils, méthodes, applications*, Talence, Centre de documentation de l'École internationale de Bordeaux.
- Gandon F., Faron-Zucker C., Corby O., 2012, *Le Web sémantique. Comment lier les données et les schémas sur le web ?*, Paris, Dunod.
- Gandon F., Dieng-Kuntz R., Corby O., Giboin A., 2002, « Web sémantique et approche multi-agents pour la gestion d'une mémoire organisationnelle distribuée », *13^e Conférence Ingénierie des connaissances*, Rouen, 28 mai.
- Hayes P., 2004, « RDF semantics. w3C Recommendation ». Accès : <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/>.
- Jacquemin C., Zweigenbaum P., 2000, « Traitement automatique des langues pour l'accès au contenu des documents », pp. 71-109, in : Le Maître J., Charlet J., Garbay C., édés, *Le Document en sciences du traitement de l'information*, Toulouse, Cépadués.
- Lauffer R., Scavetta D., 1992, *Texte, hypertexte, hypermédia*, Paris, Presses universitaires de France.
- Lauret W., Thomas A., s.d., « A study on BaseX and Benchmarking ». Accès : <https://mailman.uni-konstanz.de/pipermail/basex-talk/attachments/20100822/5b4d114d/attachment-0001.pdf>.
- Lin D., 1998, « An Information-Theoretic Definition of Similarity », pp. 296-304, in: *Proceedings of the 15th International Conference on Machine Learning*, San Francisco, M. Kaufmann.
- Monnin A., 2013, *Vers une philosophie du Web. Le Web comme devenir-artefact de la philosophie (entre URIS, tags, ontologie(s) et ressources)*, thèse de doctorat en philosophie, Université Paris I Panthéon-Sorbonne.
- Monnin A., 2014, « The Web as Ontology: Web Architecture Between REST, Resources, and Rules », *Philosophical Engineering: Toward a Philosophy of the Web*, pp. 31-51.
- Otlet P., 1935, *Monde. Essai d'Universalisme. Connaissances du Monde. Sentiments du Monde. Action organisée et Plan du Monde*, Bruxelles, Mundaneum.
- Paveau M.-A., 2015, « Ce qui s'écrit dans les univers numériques. Matières technolangagières et formes technodiscursives », *Itinéraires*, 2014-1 | 2015. DOI : 10.4000/itinéraires.2313.
- Pédauque R.T., 2006, *Le Document à la lumière du numérique*, Caen, C&F Éd.
- Portine H., 1997, « D'où vient le métalangage ? », *Linx*, 36, pp. 25-39.

Indexer des documents « du dedans » : comment répondre à la question du lieu de la donnée ?

Salaün J.-M., 2007, « La redocumentarisation, un défi pour les sciences de l'information », *Études de communication*, 30, pp. 13-23.

Salton G., 1989, *Automatic Text Processing: The transformation, analysis, and retrieval of information by computer*, Boston, Addison-Wesley.

Salvador X.-L., 2016, *XML pour les linguistes*, Paris, Éd. L'Harmattan.

w3c, 1997, « Naming and Addressing: URIs, URLs, ... ». Accès : www.w3.org/Addressing.

Annexe. Glossaire des abréviations

CSS	<i>Cascading Style Sheets</i> (feuilles de styles en cascade)	Formalisme de description de mise en page des ressources sur le web couramment utilisé pour la conception de sites internet, des livres numériques (epub). Les navigateurs web interprètent la relation entre des entités nommées grâce au dialecte HTML et des instructions de mise en page afin de construire un affichage sur écran.
DTD	<i>Document type definition</i> (définition de type de document)	Fichier ou fragment d'un document SGML ou XML décrivant la grammaire – liste des éléments (ou balises), des attributs, leur contenu et leur agencement – et le vocabulaire d'un document ou d'une classe de documents XML sous la forme d'une liste d'entités. Ce fragment permet d'exercer une coercion forte sur la saisie de données et d'homogénéiser un ensemble.
Epub	<i>Electronic Publication</i>	Structures de fichiers de sites web (HTML et CSS au minimum) compressés exploitée par des outils d'assistance à la lecture courante, comme les tablettes, liseuses...
FOAF	<i>Friend Of A Friend</i>	Vocabulaire spécialisé RDF consacré au catalogage descriptif d'entités nommées.
HTML	<i>Hypertext Markup Language</i>	Formalisme de balisage utilisé pour concevoir des documents web.
HTML5	<i>Hypertext Markup Language 5</i>	Dernière version (2007) de HTML.
HTTP	<i>Hypertext Transfer Protocol</i>	Protocole de téléchargement.
IP	<i>Internet Protocol</i>	Protocole permettant d'attribuer des identifiants uniques (des adresses, c'est-à-dire l'équivalent d'un numéro de téléphone) pour tous les terminaux connectés à un réseau (domestique, professionnel, mondial).
KWAC	<i>Keywords and Context</i>	Format d'index alphabétique représentant une extraction lexicale dans un environnement dont la dimension est fenêtrée. C'est l'usage des concordanciers linguistiques.
KWIC	<i>Keywords in Context</i>	Format d'index alphabétique utilisant des mots clés introduit par A. Crestadoro (1856) représentant des extractions lexicales classées au sein de la phrase ou du paragraphe, voire du syntagme où on le rencontre.
KWOC	<i>Keywords out of Context</i>	Format d'index alphabétique représentant des extractions lexicales en relation avec d'autres unités senties comme pertinentes entretenant un lien de cooccurrence en corpus, mais indépendamment des autres mots du contexte.

MathML	<i>Mathematical Markup Language</i>	Schéma de données ou DTD dotant un ensemble de noms d'entités XML d'une valeur sémantique signifiante dans l'objectif de conceptualiser des symboles mathématiques.
OWL	<i>Ontology Web Language</i>	Vocabulaire spécialisé RDF pour la description d'ontologies web.
PDF	<i>Portable Document Format</i>	Langage de description de pages à destination des matrices d'impression convertissant la représentation écran des données en points réels sur une page formatée. Il dérive de PostScript et permet de représenter en acte la mise en forme d'un document en puissance (polices de caractère, images, objets graphiques, etc.).
PHP	<i>Hypertext Preprocessor</i>	Couche interprétative d'un logiciel serveur statique permettant d'interpréter des requêtes web variables en fonction de formulaires de pages web dynamiquement remplis.
RDF	<i>Resource Description Framework</i>	Modèle de graphe permettant la description formelle des ressources web et de leurs métadonnées.
RELAX NG	<i>Regular Language for XML Next Generation</i>	Au même titre que DTD, formalisme de méta-description de documents XML... au format XML.
REST	<i>Representational State Transfer</i>	Philosophie d'architecture informatique pensée par R. T. Fielding (2000 : 76-106) permettant de construire des applications hypermédia distribuées.
SGML	<i>Standard Generalized Markup Language</i>	Formalisme de description par balises.
SKOS	<i>Simple Knowledge Organization System</i>	Au même titre qu'OWL, vocabulaire RDF destiné à la description de thésaurus.
SPARKL		Langage de requête adapté aux données RDF, qui se présente comme une forme dérivée de SQL mais adossé à une base de données allégorique qui serait le web.
SVG	<i>Scalable Vector Graphics</i> (graphique vectoriel adaptable)	Format de fichier XML décrivant des images vectorielles, c'est-à-dire des images composées d'objets géométriques individuels définis chacun par différentes entités XML (forme, position, couleur...) auxquelles on peut appliquer plusieurs transformations (homothéties, rotations...). Ces images sont donc différentes des images matricielles (ou <i>bitmap</i>), constituées de pixels.
TEI	<i>Text Encoding Initiative</i>	Communauté internationale visant à établir des recommandations pour l'encodage de documents textuels. Le sigle TEI désigne aussi le schéma XML élaboré et recommandé par cette communauté
URI	<i>Uniform Resource Identifier</i>	Courte chaîne de caractères permettant de cibler une ressource spécifique. Il existe deux types d'URI, celles de type <i>locator</i> (URL) et celles de type <i>name</i> (URN).
URL	<i>Uniform Resource Locator</i>	URI de type <i>locator</i> . Permet d'obtenir la donnée ciblée.
URN	<i>Uniform Resource Name</i>	URI de type <i>name</i> . Permet de décrire la ressource elle-même, indépendamment du dépôt.

Indexer des documents « du dedans » : comment répondre à la question du lieu de la donnée ?

w3c	<i>World Wid Web Consortium</i>	Consortium de standardisation chargé de promouvoir la compatibilité des technologies du web.
xhtml	<i>Extensible Hypertext Markup Language</i>	Formalisme de balisage dérivé d'xml utilisé pour concevoir des pages à destination du web. Il a été conçu comme le successeur propre et rigoureux de html.
xml	<i>Extensible Markup Language</i>	Formalisme de description de données au format textuel fondé sur l'exploitation d'entités définissant des fragments d'informations nommés au moyen de marqueurs eux-mêmes textuels, appelés balises, dont l'appellation est insérée dans le corps du document entre chevrons.
xpath	<i>xml Path Language</i>	Technologie informatique permettant de représenter symboliquement des informations au sein d'un document xml.
xquery		Langage de programmation de haut niveau permettant d'extraire des informations d'une base de données xml et de reconstruire de nouveaux documents ou fragments xml.
xsd	<i>xml Schema Definition</i>	Au même titre que rng qui en dérive, formalisme de méta-description de documents xml...au format xml.