Introduction à l'écrilecture scientifique et aux modalités techniques de son augmentation

Evelyne Broudoux, Gérald Kembellec

Introduction à l'écrilecture scientifique et aux modalités techniques de son augmentation

Evelyne Broudoux et Gérald Kembellec (Dicen)

1.1. Préalable

Ce livre collectif est l'aboutissement d'un projet que nous avons initié en 2015, fruit d'une réflexion entre les membres du Laboratoire d'excellence haStec . Le projet avait commencé par un séminaire dont certains intervenants ont accepté de participer à la rédaction ou à l'évaluation de chapitres de cet ouvrage. Sont présentées ici des contributions originales sélectionnées et évaluées par au moins deux membres du comité scientifique, que nous tenons ici à remercier chaleureusement. Cette introduction propose des éléments de synthèse du séminaire et des clés de lecture destinées à faciliter la compréhension de l'ouvrage.
L'objectif de cette introduction est de placer l'écrilecture numérique dans le contexte des humanités numériques afin de mieux saisir l'ancrage de ce procédé comme partie prenante de cette mise en mouvement disciplinaire. L'écrilecture, y compris scientifique, est une pratique ancienne dont les procédés évoluent en même temps que les outils et qui n'a nullement eu besoin d'attendre les nouvelles théories de pensée du 20e siècle pour exister et structurer la réflexion de génération de penseurs. Cependant, dans la première moitié du 20e siècle, les écoles de pensée philosophiques récentes, à l'échelle de l'histoire de l'écriture scientifique, ont investi la pensée scientifique par des postures normatives dans lesquelles la pensée peut être décrite et catégorisée. La possibilité de relier les connaissances humaines de manière outillée a été ensuite pensée avec le Memex avec la fin de la deuxième guerre mondiale, même s'il a fallu attendre la dernière décennie du 20e siècle pour voir arriver un début de concrétisation : rappelons que le web était pensé initialement comme scientifique et inscriptible. Nous proposons ici de revenir brièvement sur les humanités numériques et leur lien avec le processus d'écrilecture qui s'externalise dans des formes dynamiques de réception. Nous en profiterons pour présenter les modèles de structuration de l'information, en particulier ceux des données liées du web sémantique qui permettront de mieux appréhender certains chapitres.

1.2. Humanités numériques

1.2.1. Un champ de pratiques

Depuis une dizaine d'années, les « humanités numériques » se sont progressivement imposées comme un champ interdisciplinaire de recherche auquel correspond un ensemble de pratiques en cours d'adoption par les sciences humaines. Une première phase a consisté à s'approprier les technologies informatiques offertes par la numérisation de documents. Plus particulièrement, les disciplines de l'Histoire, des Lettres, des Arts et des domaines muséographiques et archivistiques, qui ont vu leurs objets d'études numérisés, se sont vu offrir des possibilités de recherche inédites avec l'accès simplifié à des sources issues de la construction de nouvelles bases de données. Les représentations visuelles issues des calculs statistiques sur des données quantitatives ont été mises à la portée de tous grâce à des algorithmes encapsulés dans des interfaces graphiques. Le livre numérique matérialise son augmentation avec des sommaires à plusieurs entrées [TRE 14] et la représentation

cartographique facilite la recherche d'informations. Les modes de narration et d'illustration hypermédia viennent compléter cette offre comme l'indique l'exemple ci-dessous qui correspond à une image issue d'une base de connaissances prosopographiques en histoire de l'Art et représente une frise temporelle chronologique générée à partir de l'interrogation d'une base de données.

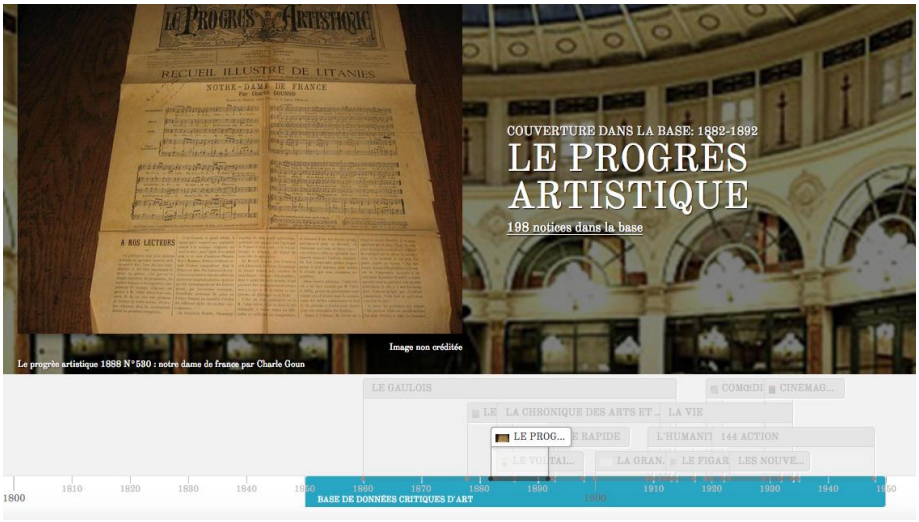

Figure 1.1. Frise chronologique permettant l'affichage et l'interrogation des contenus d'une base de données bibliographiques en liaison
avec des contenus extérieurs (LOD)

La première phase des « humanités numériques » a été donc de faire entrer les techniques numériques de plain-pied dans les méthodes d'analyse et d'interprétation de corpus, de langues, de terrains, d'archives, mais aussi d'encadrer principalement des projets éditoriaux de numérisation visant à mettre à disposition des chercheurs et du public, des œuvres, des auteurs, des documents historiques.
Une seconde phase plus réflexive — correspondant à l'arrivée d'objets de recherche numériquement natifs — ,a fait le constat du besoin de formation, correspondant, selon l'approche nord-américaine, à une alphabétisation ou literacy [LED 12] et qu'Emmanuel Souchier [SOU 13] préfère nommer « lettrure » numérique. Dans le même temps, était relevée la nécessité d'interroger ce qui apparaissait comme une transformation des méthodes d'analyse et de recherche [RIE 10]. Un déplacement des frontières se déroulerait à la fois au sein des disciplines et des métiers [DAC 15].

1.2.2. Mise en mouvement disciplinaire

Une explication à ce déplacement, autre caractéristique des « humanités numériques », est qu'elles se décrivent comme une mise en mouvement disciplinaire. En effet, on peut imputer l'origine de ce « mouvement » à l'effort de désenclavement des « humanités » nord-américaines qui sont pointées du doigt comme non rentables et ne sont pas associées aux sciences sociales comme la sociologie ou l'anthropologie, comme c'est le cas en Europe où les Sciences Humaines et Sociales sont rassemblées dans le même bouquet. Ce « mouvement » ayant aujourd'hui largement essaimé hors des « humanités », il interroge fondamentalement le substrat théorique qui fonderait une interdiscipline.

Un mouvement est porté par des événements fédérateurs qui diffusent des points de vue partagés sans toutefois être dirigé par une entité spécifiquement désignée pour cette fonction. Les « digital humanities » ont gagné en notoriété grâce à des rassemblements dont l'organisation est particulière comme les « BarCamp » transformés en THATCamp disséminateurs d'idées et d'actions à entreprendre d'où sont issus des manifestes dont la raison d'exister est de décrire des situations et de prescrire des solutions.

Le premier manifeste publié le 15 décembre 2008 par Jeffrey Schnapp, Peter Lunenfeld, Johanna Drucker et Todd Pressner sur les serveurs de l'Université de Californie à Los Angeles (UCLA), a pour particularité d'être le résultat d'un séminaire (Mellon) et d'un travail collectif d'écriture intégrant 124 commentaires (triés sur invitation). Il utilise la plate-forme WordPress et son plugin dédié CommentPress , une entité scriptible par les lecteurs. Si la teneur de ce manifeste est volontairement subversive et radicale (par exemple : tout ce qui n'est pas « ouvert » est considéré comme « ennemi »), les commentaires en sont d'autant plus critiques. Son objectif majeur est ainsi de sortir les humanités hors-les-murs de l'université (bien que celle-ci ne soit aucunement nommée) ; les disciplines et départements étant considérés comme des systèmes de domination perpétuant des règles destinées à légitimer des rentes de situation et freinant d'autant plus les changements.
Un second manifeste « 2.0 » paraît en mai 2009 — traduit par [JUL 15] — entérinant le premier, en particulier dans son insertion dans la « wiki-économie » et sa lutte contre la « naturalisation » de la culture de l'imprimé. Les « digital humanities » y sont vues comme un faisceau de convergences plutôt que comme l'unification d'un champ. Une place de choix est attribuée à la curation en tant que « pratique savante augmentée » et à l'ouverture à des acteurs autres que scientifiques.
En France, un blog de veille Digital Humanities International , financé par un projet du TGE Adonis, a publié 568 billets sur cette thématique entre 2008 et 2012 ; puis un véritable élan a été impulsé par Open Edition avec le lancement du premier THATCamp européen sur la question des « digital humanities » en 2010. Une synthèse en est réalisée sous forme d'un manifeste rédigé en français se différenciant de ceux publiés sur le site de l'UCLA. Sont nommées :

– la « modification des conditions de production et de diffusion des savoirs » ;
– la formation du champ des humanités numériques issue de la « convergence d'intérêt des communautés » envers des pratiques, des outils ou des objets transversaux divers (encodage de sources textuelles, systèmes d'information géographique, lexicométrie, numérisation du patrimoine culturel, scientifique et technique, cartographie du web, fouille de données, 3D, archives orales, arts et littératures numériques et hypermédiatiques, etc.).

Les acteurs déclarent se constituer en « communauté de pratique solidaire, ouverte, accueillante et libre d'accès ». Des orientations privilégient l'accès libre aux données et aux métadonnées ainsi que le partage et le travail collectif.
Les projets d'humanités numériques sont aussi encouragés par les infrastructures publiques qui visent à soutenir techniquement les initiatives de numérisation en fournissant en France des équipements (TGE Adonis puis TGIR Huma-Num), une bibliothèque scientifique numérique (BSN) et pour l'Europe, l'infrastructure Dariah-EU .
Le dynamisme du mouvement se révèle à la consultation des informations publiées sur la liste DH, liste francophone de discussion autour des « digital humanities » ouverte en mars 2010 par Frédéric Clavert, Marin Dacos et Pierre Mounier, devenue depuis un service de l'Association francophone des humanités numériques/digitales, Humanistica.

1.3. Spécificités de l'écrilecture

1.3.1. L'écrilecture scientifique

Les pratiques d'écrilecture numérique renvoient à la fois à la lecture savante et au web. Les pratiques liées aux usages de lecture « savante » se sont perpétuées au cours des siècles et les annotations sont elles-mêmes devenues objets d'études, comme plus-values des textes originaux ou documents à part entière.
Historiquement reconnues depuis le XIIe siècle, les premières techniques de lecture dites « savantes » réunissaient la lecture et l'écriture dans un processus de *lettrure*, mêlant lecture attentive et commentaire. Réservées à une élite restreinte de « lettrés » religieux, la lecture et l'écriture étaient pensées comme un seul processus, constitué par des actions liées et complémentaires dans lesquelles l'activité de lecture, hautement structurante, permettait au récepteur de la connaissance construite de devenir acteur par l'enrichissement des idées transmises. Ce procédé intervenait par capitalisation intellectuelle et agrégation dans une transformation scripturale et pouvait être matérialisé sur le support physique au moyen de *marginalia*, notes de bas de page et autres annotations.
La mise en réseau réalisée par le web a transformé cette activité en ajoutant des couches techniques qui concernent à la fois les processus d'écriture et de lecture, mais aussi la circulation des textes, leur augmentation potentielle et effectuée, leur diffusion et la captation des traces liées à leur réception. Le web et les technologies associées au lien hypertexte sont à l'origine d'environnements de lecture enrichie dont nous avons commencé à établir un état des lieux sous l'angle de l'innovation logicielle, mais aussi ceux des usages actuels et à venir.
Le néologisme écrilecture a permis de rendre compte de pratiques littéraires créatrices avec l'ordinateur, comme la génération automatique de textes. Sous le terme *ecrileitura* dès 1992, Pedro Barbosa a décrit ce phénomène de délégation au lecteur de la constitution de textes à lire, l'auteur se situant en amont dans un texte-programme générant de multiples variations dont il ne maîtrise ni les formes lisibles ni ses interprétations. Alain Vuillemin l'avait repris pour caractériser ce nouveau comportement du lecteur entraîné dans des manipulations créatrices face à l'écran. *« L'acte d'écrilecture, d'écriture et de lecture interactives, est alors conçu comme une action périphérique, faite par l'utilisateur d'un ordinateur autour d'un fragment de texte de référence »* [VUI 99, p. 103]. Le premier « système d'annotation dynamique » a été ainsi conceptualisé dès 1999 pour les lecteurs de la BNF dans un programme de numérisation :
« Il sera possible de constituer un corpus de texte à partir des collections, de l'organiser en y introduisant des signets ou des balises, puis d'y associer des annotations et des commentaires à propos de fragments qui auront été sélectionnés au préalable » [VUI 99, p. 103].
Malheureusement, ce projet a fait long feu et s'il a été question d'une *« seconde génération »* de postes de lecture, force est de constater que ceux-ci ont été remplacés par de simples postes de recherches de références sans fonctionnalité d'écrilecture ni de possibilités de partage des références cherchées entre lecteurs.
Celui-ci précisait sa pensée :
« L'idéal serait que […] la lecture puisse induire un acte d'écriture et les actes de récriture un acte de relecture qui puissent se faire non seulement autour d'un texte, mais aussi, en

quelque sorte, à l'intérieur de ce texte, dans son "épaisseur" intratextuelle et intertextuelle. […] Lorsque ce processus d'intégration s'affirmera, la lecture cessera d'être "assistée" par ordinateur pour devenir une forme de lecture active, voire interactive, jusqu'à se transformer en une action dynamique, sans doute indéfiniment renouvelée, bref en une véritable action d'"écrilecture" créatrice » [VUI 99, p. 102].

Dans leur chapitre *« L'écrilecture : une pratique révélatrice de la construction de connaissances au sein de communautés professionnelles »*, Viviane Clavier et Céline Paganelli reviennent sur l'écrilecture comme un processus intellectuel et instrumenté qui permet d'analyser comment s'élaborent les connaissances de communautés professionnelles à partir de l'observation de leurs pratiques documentaires. Elles distinguent l'écrilecture révélatrice de l'activité scientifique de la lettrure liée au livre et révélant l'érudition. Leur étude reprend trois terrains différents et propose une synthèse des points communs et des différences des pratiques observées s'appuyant sur des travaux déjà réalisés. Les trois communautés étudiées concernent des chercheurs en littérature, des médecins hospitaliers et des doctorants en sciences de l'information et de la communication.

Cette posture critique, qui est la condition de la transmission des connaissances, a été explorée par Thomas Bottini dans son chapitre *« "Les espaces de la critique" Une étude des conditions de possibilité d'une lecture savante et multimédia »*. En effet, le concept d'écrilecture part du principe que le travail d'écriture intérieure pendant la lecture peut s'externaliser dans différentes formes d'annotation soutenues par des procédures logicielles. Encore faut-il s'intéresser à la facette opératoire mettant en présence les opérations mentales spécifiques à la critique et les propriétés des supports des contenus « savants ». La conceptualisation d'un système d'écrilecture passe donc par la présentation des caractéristiques fondamentales auquel doit répondre un dispositif multimédia. D'une part, cet espace doit pouvoir accueillir des éléments sémiotiques variés (fragments textuels, graphiques ou sonores, etc.) sans entraver l'exploration critique et tout en préservant des gestes manipulatoires de base : *« rendre accessible la forme sémiotique d'appropriation, définir un point d'intérêt, délimiter une zone, extraire un fragment ».* D'autre part, les règles issues de la logique typodispositionnelle du document final ne doivent pas s'imposer au détriment des opérateurs critiques qui favorisent l'exploration d'un réseau émergent de signification.

L'annotation a pour particularité de concerner à la fois les humanités et les informaticiens s'intéressant à l'écriture et de nombreux travaux ont été réalisés depuis une trentaine d'années sur cette thématique. Récemment, des carnets de recherche rendent compte de l'évolution de thèses comme celui de Marc Jahjah (2014) sur *« Les marginalia de lecture dans les "réseaux sociaux" du livre »* ou de projets de recherche comme celui de Johanna Daniel qui a réalisé un benchmark d'outils d'annotations pour son mémoire intéressant les historiens des arts (2014).

L'annotation remplit plusieurs fonctions à tous les stades de la publication : fonctionnalité d'avancement d'un objet en cours d'écriture individuelle ou collective, fonctionnalité de commentarisation appuyant la constitution collaborative d'un appareil critique.

Rappelons la distinction entre métadonnées et annotations [PRI 06] : une métadonnée est attachée à une ressource identifiée en tant que telle alors que l'annotation est *« plus située au sein de cette ressource et écrite au cours d'un processus d'annotation-lecture ».*

L'annotation se réalise donc au sein de l'objet d'écriture, au cours d'un processus manuel d'écrilecture.

Un nouveau pas a été franchi lorsque l'on considère que les processus d'écrilecture, soutenus par de multiples fonctionnalités logicielles, pourraient avoir des prolongements automatisés réalisés par les raisonnements computationnels sur la sémantique, réalisant en cela une augmentation.

1.3.2. L'écrilecture comme concept majeur des « humanités numériques »

L'interrogation d'Olivier Le Deuff [LED 15] sur le rôle joué par l'indexation dans la fondation des « humanités numériques », en tant que pratique de lecture/écriture issue de la main, rejoint l'introduction du manifeste 2.0 des Humanités Numériques qui se définit comme « une main ouverte et tendue » [JUL 15].
Les évangélistes des « humanités numériques » diffusent l'idée que le numérique transforme profondément les activités liées à la construction des savoirs et rejoignent en cela d'autres précurseurs qui ont compris tôt que l'ordinateur était autant un outil d'écriture qu'un outil de calcul. Citons « Computers and writing – State of the art » [HOL 92], un ouvrage à caractère fondateur qui rassemble une compilation d'articles interdisciplinaires autour de l'analyse statistique de textes, l'indexation, la conception d'éditeurs de textes, la gestion de références, l'écriture collaborative, l'écriture hypertextuelle, les aspects cognitifs de l'écriture, etc.
Autant de directions qui n'auront cesse d'être creusées ultérieurement par ce que l'on pourrait appeler les précurseurs des « humanités numériques » comme Jay Bolter [BOL 90], co-auteur de l'outil d'écriture hypertextuelle Storyspace, pour lequel l'ordinateur représente une nouvelle phase de spatialisation de l'écriture, avec Illich [ILL 91] et Goody [GOO 79].
« Writing is always spatial, and each technology in the history of writing (e.g., the clay tablet, the papyrus roll, the codex, the printed book) has presented writers and readers with a different space to exploit. The computer is our newest technology of writing, and we are still learning how to use its space » [BOL 90].
Il est devenu commun pour un projet d'humanités numériques de mettre à disposition du lectorat une plate-forme pourvue de fonctionnalités d'annotation. Citons par exemple pour son côté réflexif, « The Debates in the Digital Humanities » , une plate-forme de publication hybride lancée en 2013 qui explore les débats du « champ » des humanités numériques au moment de leur émergence. Simultanément à l'édition imprimée, la publication en accès ouvert est disponible. Dans un deuxième temps, la plate-forme a inclus des fonctionnalités qui permettent aux lecteurs d'interagir avec le contenu en marquant explicitement des passages et en ajoutant des termes à un index réalisé de manière collective.
Aux confins des « humanités numériques » et des procédés d'écrilecture se trouvent les outils logiciels d'annotation censés soulager la charge cognitive en fournissant des espaces d'extériorisation de la pensée pour servir la posture critique. Marc Jahjah dans son chapitre *« "Annoter le monde et améliorer l'humanité" : imaginaires et fabrication d'un logiciel d'annotation »* met à nu les arguments fondés sur l'imaginaire dans la présentation du logiciel Hypothes.is qui vise essentiellement la recherche universitaire. L'analyse sémiotique des interfaces de cette extension, qui se branche sur un navigateur et crée une colonne supplémentaire autorisant des annotations sur les sites web visités, apparaît faciliter un processus d'écrilecture sans échange et conforter une vision surplombante.
Autres exemples d'annotations, celle de revues scientifiques s'ouvrant aux évaluations ouvertes et aux commentaires comme PeerJ, revue appartenant au champ des sciences biologiques et médicales ou — côté français — la récente expérience de la revue VertigO

qui s'est déroulée, pendant trois mois fin 2015, sur cinq textes pour des évaluations ouvertes au contraire d'un processus qui se déroule habituellement en double-aveugle et sur cinq autres ouverts à commentaires, dont l'objectif explicitement formulé visait une amélioration formelle de l'expression.
Cet appel à la communauté des pairs pour l'évaluation — consultable par tous — et les propositions de modifications formelles qui sont sollicitées sous la forme de commentaires — ouverts à tous —, est révélateur d'une tendance à élargir des processus de sélection scientifique au grand public.
Le chapitre de Lisa Chupin « La construction de normes d'écrilecture pour la transcription collaborative du patrimoine numérisé : entre algorithme, transmission et élaboration communautaire » présente ainsi un exemple de crowdsourcing scientifique de transcription d'étiquettes d'herbier. Les tâches d'écrilecture confiées au public sont décomposées en amont pour produire des écrits normalisés, à partir desquels les contributeurs sont en mesure de trier et de confronter leurs propositions, selon des formes d'interaction contrôlées. Les propositions seront ensuite traitées par des algorithmes aidant à la résolution de conflits d'interprétation et statistiques facilitant les choix établis selon des critères de validité scientifique. Le deuxième étage de l'écrilecture réside dans les commentaires dont la teneur n'est pas immédiatement utile au dispositif créé, mais dont l'exploitation ultérieure entraînera une augmentation des connaissances obtenue par les participants. Leur valeur informationnelle réside dans les liens susceptibles d'être créés au sein des collections internes ou dans les améliorations à apporter au design des interfaces.
Outre la numérisation et la participation, les concepts-clés des « humanités numériques » sont la « sémantique » et « l'interopérabilité » [BLA 15].
Si l'écriture hypertextuelle n'a pas tenu ses promesses dans sa remise en cause de la narrativité, force est de constater que la nouvelle instrumentation technique de la sémantisation de l'écriture ouvre des portes jusqu'ici invisibles.
En effet, les technologies du web sémantique composent une double-écriture et une double-lecture. Lecture par les humains et lecture par les machines. L'annotation dispose même de son propre vocabulaire de métadonnées descriptives . L'écriture automatisable par la machine en termes d'annotations et de métadonnées est à même de conditionner des lectures ultérieures par les humains. Ce conditionnement résulte d'indications d'indexation à destination des différents moteurs de recherche « horizontaux » comme celui de Google ou « verticaux » comme ceux moissonnant les archives ouvertes des publications scientifiques.

1.4. Technologies hypertextes actuelles

1.4.1. De l'hypertexte au web de données

Dans un premier temps, à l'heure du premier web, les éléments purement textuels forment le contenu du document, ils sont illustrés, argumentés, par des hypermédias proposés directement par le rédacteur qui peut en être également l'auteur ou agrégés par l'auteur-éditeur qui va les sélectionner depuis des sources externes pour les charger dans son document en hyperliant éventuellement la source. Ensuite, avec la révolution du web social (web 2.0), les hyperdocuments sont proposés au sein de dispositifs qui permettent une interaction entre le lectorat et l'auteur par des techniques de fils de commentaires qui enrichissent le document d'une métaréflexion commune qui peut parfaitement s'appliquer à un processus scientifique (voir le chapitre « Les herbiers »). Avec la troisième révolution,

dite sémantique, il devient possible de segmenter et documenter finement les composantes de sa production hypertextuelle dans une optique de normalisation et surtout de partage : les données inscrites dans les documents deviennent des entités contextualisables, mais autonomes comme de petites unités de sens (microdata en anglais), qui peuvent librement être liées à d'autres données porteuses d'un sens similaire dans un autre contexte : le principe du linked open data peut ainsi être approché étymologiquement d'un certain « tissage » pour faire référence à la théorie du texte de Roland Barthes [BAR 74].
Le principe de cette interaction et les bienfaits en termes de désambiguïsation conceptuelle, de sérendipité, de découvertes d'informations connexes se perçoit aisément dans l'optique du berrypicking, qui consiste à « rebondir » d'un document vers un autre et de requalifier son besoin d'information au fur et à mesure de la découverte de contenus selon le modèle présenté par Marcia Bates [BAT 93] . Si l'on comprend facilement l'intérêt du « pourquoi » de cette éditorialisation des métadonnées autour des contenus en termes d'interconnexion de données et de construction du savoir, les aspects conceptuels du « comment » sont peut-être moins triviaux. Nous proposons de revenir dans la partie suivante sur la définition du formalisme des descriptions et liaisons de contenus dans le cadre précis des hyperdocuments tout en nous attachant à en préciser le bénéfice dans un cadre d'écrilecture contextualisé scientifiquement.
Avec les nouvelles méthodes de liaisons de données intra et inter documentaire ou tout du moins leur vulgarisation récente, le champ des possibles en termes de production scientifique s'élargit surtout avec la mise à disposition de jeux de données, de vocabulaires descripteurs dédiés, les plates-formes d'écriture collaboratives, les bases de médias scientifiques et bien sûr, les entrepôts d'articles scientifiques.
Au coeur de ces principes, se trouve donc l'interopérabilité. Camille Claverie et Annaïg Mahé reviennent sur le principe de l'interopérabilité conceptuelle et technique, présupposée à l'heure du web des données liées, entre les fragments d'information, mais dont la mise en oeuvre est loin d'être triviale avec une gouvernance mondialisée [BOU 16]. Par certains côtés, au sein de la communauté scientifique, les dépôts d'informations constituent une tour de babel de normes, standards et protocoles qui ralentissent plus qu'ils accélèrent la possibilité de lier d'annoter, d'augmenter des références dans un processus d'écrilecture. Rosemonde Letricot et Francesco Beretta, par un retour d'expérience probant sur un projet d'humanités numériques en Histoire, reviennent sur ces problématiques d'« ensilotage » de données en présentant une méthodologie de modélisation des fragments informationnels qui non seulement permet l'interopérabilité, mais l'encourage par une description fine des contenus et des liens qui les associent.
Cette problématique peut sembler purement documentaire, voire datée avec la description de documents finis dans un contexte hypermédié, au sein duquel l'expérience de lecture ne se limite pas à un document, qui lui-même voit fondre ses limites [BRO 15] avec ses inclusions, ses liens entrants et sortants, ses versions possibles comme dans le cas des wikis. Pour sortir de ce travers d'obsolescence, il faut s'inscrire dans un contexte plus restreint : celui des contenus, de la donnée atomisée aussi nommée fragment informationnel [PRI 04]. L'information a été longtemps définie, dans une logique computationnelle, comme la réception d'un contenu contextualisé, inscrit, avec une éditorialisation, dans un contexte de réception dépendant d'un lectorat. Cela posait le problème de la réception, qui bien sûr est fondamental dans un processus d'écrilecture. Shannon posait le problème très clairement dans son article fondateur de 1948 [SHA 48] :

« The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem ».

Le web des données déplace ce paradigme avec des contenus qui peuvent être numériques — ou pas — mais dont la collection, finement décrite et désambiguïsée, permet de former un corpus dont la matérialité est nécessairement numérique avec un mode unique de description : le triplet sémantique, mais dont les modes d'inscription et de réception peuvent être multiples sur la forme, mais absolument pas sur le fond. Stéphane Crozat va jusqu'à avancer que les nouvelles chaînes éditoriales outillées par la sémantique permettent de « rendre calculables des productions originellement sémiotiques » [CRO 16]. Nous reviendrons sur ce point après avoir explicité, pour rappel, ce qu'est un triplet et sous quelles formes il peut être rencontré dans divers contextes liés à la recherche. Nous définissons le triplet sémantique ou Resource Description Framework (RDF) comme le formalisme de description d'un contenu basé sur un principe très simple assez similaire à la construction grammaticale d'une phrase, l'association d'un « sujet », d'un « verbe » et d'un « complément » respectivement nommés « sujet », « prédicat » et « objet ».

– Le « sujet » est la ressource présentée dans l'association, cette ressource peut être représentée par une adresse sur Internet, on parle d'Unified Ressource Index (URI), une chaine de caractère on parle de « littéral » ou encore, un identifieur unique dans une base de connaissances : une Unified Ressource Number (URN). Par exemple : si l'on souhaite décrire un article scientifique, on peut par exemple y faire référence grâce à son adresse pérenne sur un entrepôt scientifique comme HAL ou ArXiv (URI) ou encore, son Digital Object Identifier (DOI), un identifiant unique attribué par les autorités scientifiques qui est donc une forme d'URN dédiée aux articles de recherche.

– Le « prédicat » est la propriété attribuée au « sujet ». Cette propriété se rattache à une catégorie prédéfinie par des règles exposées au sein d'un ensemble de règles adoptées par les communautés d'usages et stockées de manière pérenne sur des serveurs dédiés dont l'adresse sur Internet est statique dans le temps. Le « prédicat » est ainsi présenté sous la forme d'une adresse sur le web qui spécifie en préfixe l'adresse du vocabulaire descriptif choisi (ou schéma), en radical, le concept descripteur concerné, enfin, le suffixe est l'un des attributs descripteurs du concept.

– Le dernier élément du triplet, l'« objet » est la valeur de la propriété ou « prédicat » attribuée au « sujet », ce peut être comme le sujet, un littéral, une URI ou une URN.

Un exemple simple de triplet peut être un article scientifique hébergé en ligne par l'équipement d'archives du CNRS. Le « sujet » sera dans ce cas au choix l'URI à laquelle cet article est accessible ou son numéro unique dans l'archive. Pour le prédicat, il faut choisir l'adresse web d'un langage de description faisant consensus, comme celui du Dublin Core Metadata Initiative avec ses quinze descripteurs basiques, puis spécifier le descripteur sélectionné.

Ainsi, la phrase « L'article hébergé par Hal avec l'URI https://hal.archives-ouvertes.fr/hal-00628355 a pour titre "Ontologie franco/anglaise du domaine informatique comme accès à un corpus de textes scientifiques" » se traduira sous forme du triplet suivant :

<https://hal.archives-ouvertes.fr/hal-00628355>,
<http://purl.org/dc/elements/1.1/title>,

<‘Ontologie franco / anglaise du domaine informatique comme accès à un corpus de textes scientifiques’>

Pour continuer la description de cette ressource, il est possible de créer d’autres triplets avec le même « sujet », ici l’article, mais des prédicats différents, comme son (ou ses) auteur(s), sa date de publication, le sujet traité, la langue de diffusion, etc., et bien évidemment, les « objets » correspondants au prédicat pour ce même objet.
La métaphore de la grammaire évoquée précédemment peut se poursuivre avec une phrase que l’on exprime à la voix active ou à la voix passive en inversant le sujet et le complément, avec des notions de collection transitives : « les chapitres composent un ouvrage » équivaut à « l’ouvrage est composé de ses chapitres ». Cette réflexivité est absolue sur le fond, car sur la forme, comme dans la langue française l’accent est mis dans un cas sur l’objet « chapitre », dans le second sur la collection : « l’ouvrage ».
Il existe techniquement plusieurs méthodes pour décrire les fragments informationnels au sein des pages web, les plus utilisées comme les microdonnées et RDFa bénéficiant d’un réel enthousiasme dans les communautés de pratiques et scientifiques sont normalisées. Des vocabulaires et modèles de données sont créés et mis à disposition sur le site schema.org y compris pour modéliser des objets scientifiques. Certains gestionnaires de contenus commencent à les intégrer de manière fine pour les rendre « découvrables » et exploitables par l’utilisateur final.
Ainsi, une simple page web peut être la somme augmentée de mentions à des contenus matériels, eux-mêmes référencés dans des catalogues numériques et présentés dans des langages de description consensuels disponibles en ligne. Les parties textuelles proposées par l’auteur pour décrire et/ou les critiquer peuvent également être jalonnées d’éléments descripteurs.
Pour illustrer ce propos, prenons l’exemple de la notice biographique d’un artiste qui pourra intégrer la biographie du peintre de manière textuelle, un portrait illustrant l’auteur, une catalographie partielle ou complète de ses œuvres – localisées et identifiées –, de la littérature secondaire et éventuellement une analyse critique de sa production incluant ses influences, ses thématiques de prédilection, ses liens de co-création éventuels, les contextes de production comme les revues ou les galeries fréquentées. Si l’on se penche davantage sur cet exemple, nous pouvons analyser de manière plus fine les méthodes de description pour les contenus composant cette notice biographique. La transposition du texte vers l’hypertexte a déjà largement été traitée dans la littérature sous un aspect sémiotique avec l’analyse de la segmentation textuelle et le balisage associé. La « calculabilité numérique » entre réellement en jeu avec la dernière version du langage HyperText Markup Language (HTML) qui offre la possibilité de composer des hyperdocuments dont les parties peuvent toutes être explicitement identifiées selon une typologie orientée plus sur la sémantique que la présentation. Il devient ainsi possible d’affiner la granularité de la segmentation des documents avec des balises déclarant leurs contenus comme des « articles », « dates », « définitions ». Ainsi, les dates et lieux de naissance et de mort seront des éléments factuels qui seront balisés dans le texte par des « bornes » sémiotiques qui peuvent être présentées ou mises en exergues, par des outils de lecture hypertextuelle. Ces outils sont le plus souvent des additifs logiciels pour les navigateurs qui peuvent être facilement activés ou désactivés selon les besoins.
Des travaux plus récents ont analysé les contextes métalangagiers des liens hypertextes, mettant en valeur d’autres enjeux invisibles à l’œil humain : les contextes d’éditorialisation

de l'hypertexte ne concernaient plus seulement un destinataire humain, mais tenaient compte des « besoins » des machines, mais aussi des « algorithmes » d'indexation. Les algorithmes d'indexation, Google en tête, recevront des informations complémentaires à ce qui est affiché, grâce à des métadonnées inscrites dans le code hypertextuel [KEM 16a] : pour des raisons d'ordre éthique, éditoriale ou économique par exemple, il est possible de citer et d'hyperlier une ressource pour l'œil humain tout en « interdisant » aux moteurs de recherche de suivre le lien sortant et de le comptabiliser dans l'algorithme de popularité [SAE 15a, SAE 15b, SIR 13]. Le contraire est également envisageable : il est également possible de lier une ressource ou de la méta-information à destination exclusive des outils d'analyse [KEM 16b] tout en estimant que le bénéfice de l'affichage de cette information n'en justifierait pas le coût cognitif, selon le principe de la surcharge informationnelle. Dans tous ces cas de figure, les enjeux de la description et de la liaison des données en contexte traditionnel d'optimisation de classement des pages web par la qualité structurelle intrinsèque SEO, va aussi s'appliquer au cadre de la recherche, y compris en sciences humaines, car les nouvelles plates-formes scientifiques d'accès aux documents de la recherche utilisent ces nouvelles méthodes pour proposer l'accès et la liaison entres documents de la recherche.

1.4.2. Spécificités de l'augmentation scientifique

Quelques exemples.
Grâce à ces nouvelles méthodes de liaison sémantique des fragments du web, de nouvelles possibilités d'enrichissement émergent. La TGI Adonis par exemple, offre une extension logicielle pour les carnets de recherche qui permet de proposer contextuellement de manière automatique des recommandations de contenus de la recherche (articles, thèses, chapitres, illustrations, etc.) tirés de la plate-forme Isidore et en lien direct avec le billet en cours [POU 16].

Figure 1.2. Preuve de concept d'enrichissement d'une page hypertexte

Dans un même esprit, Thomas Francart propose d'intégrer des contenus hypermédia, affichables à la demande, tirés des entrepôts de données scientifiques comme ceux tirés de médiHal ou de bases de données encyclopédiques sémantisées (voir la figure 1.2), comme la base de connaissances dbPédia, à des articles ou des carnets de recherche en marge des

contenus [FRA 15]. Ces contenus peuvent être sélectionnés manuellement par l'auteur, par un écrilecteur physique ou encore, un algorithme en fonction de l'observation de traces.

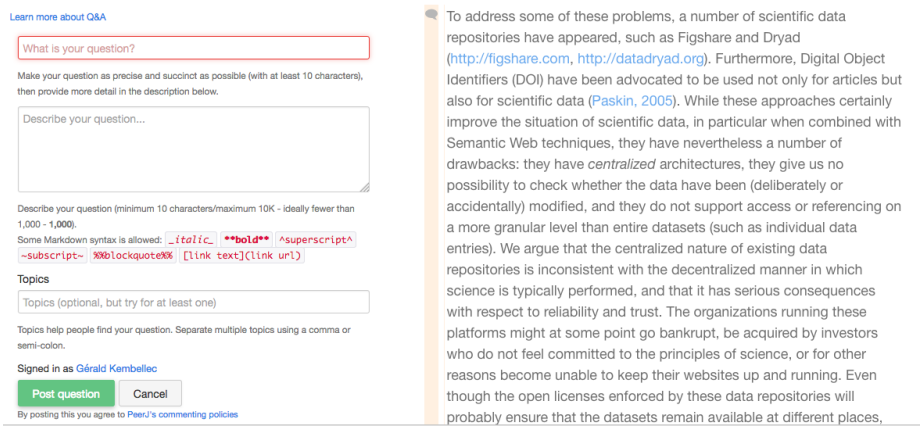

Figure 1.3. Exemple d'annotation sur un article dans la revue PeerJ

Dans le cadre de l'édition scientifique, Hans Dillaerts et Lise Verlaet reprennent le concept de linked data en lui préférant le terme de semantic publishing (éditorialisation sémantique) proposé par Shotton [SHO 09] qu'ils associent à l'écrilecture [VER 16] : « De nouvelles formes de revue scientifique permettent désormais aux lecteurs de participer aux processus d'éditorialisation sémantique, notamment via des outils d'"écrilecture" ». Le semantic publishing offre de nombreux avantages pour l'auteur comme pour le lecteur : l'enrichissement sémantique des publications scientifiques avec des données interactives ; un renforcement du sens des articles par de balisage sémantique ; des liens directs vers des ressources externes ou vers les références citées ce qui encourage la découverte et la réutilisation des nouvelles connaissances en publiant l'article et les données de recherche dans des formats lisibles par les machines et les humains. Les enjeux et défis de ces nouvelles modalités d'écrilecture seront présentés dans le chapitre 8 écrit par Hans Dillaerts et Lise Verlaet.
La revue annotable en ligne PeerJ permet, grâce à des métadonnées éditorialisées en POSH , RDFa, Microdata, d'annoter contextuellement, de poser des questions situées à l'auteur, de débattre donc (voir la figure 1.3, avec le texte à droite l'outil d'annotation sur la gauche) au sein d'une revue peer reviewed dont l'historique des versions/reviews est accessible et rend observable, en asynchrone, la construction définitive du document avec les points de vue sur la science en train de se faire. Sur un modèle similaire, Alexandre Monnin a proposé une version annotable de sa thèse en 2013 sur le site Philoweb.org , dans la logique d'un web scientifique sémantique augmenté, en segmentant ses contenus sur le modèle du HTML5. Le dispositif d'interaction s'équipait du plugin cité plus haut « CommentPress » qui permettait aux lecteurs critiques d'adresser à l'auteur des commentaires situés. Sur la durée de l'expérience, soit deux ans, une première couche de péritexte s'est greffée pour que l'auteur puisse présenter des informations connexes, puis les lecteurs ont commencé à participer à une annotation du tapuscrit. Après un début qualifié d'intéressant par l'auteur, le débat s'est mené en parallèle en épitexte sur le média social lié au projet : la page Facebook associée à l'expérimentation est également devenue un véritable lieu d'échanges. De son côté, Johanna Daniel, lors de son passage en Master « Technologies numériques appliquées à l'Histoire » à

l'Ecole des Chartes, rejoint de manière très pragmatique le fond et la forme en proposant un suivi de l'écriture de son mémoire sur « les outils d'annotation au sein d'un outil d'annotation et leur emploi dans le cadre de l'édition scientifique de corpus textuels », précisément en utilisant un outil d'annotation. Son expérience n'était pas désintéressée, car elle lui a permis d'exposer son travail et d'avoir de nombreux retours tant méthodologiques que de fond, ainsi qu'une correction orthographique collaborative .

1.5. Conclusion

L'introduction de ce livre avait pour objet de présenter le cadre de l'écrilecture numérique, ainsi que les principaux éléments historiques, conceptuels et techniques qui permettent son existence au sein des humanités numériques.
Les chapitres sélectionnés approfondissent ces différents aspects en les illustrant et ouvrent des pistes de réflexion qui concernent par exemple, la reproduction des modèles éditoriaux classiques ou leur transformation par les formes collaboratives de l'écrilecture. De même, l'automatisation de la référence ne suppose pas forcément une modification du modèle éditorial traditionnel, mais plutôt son intégration dans un parcours encadré. En ce qui concerne l'écrilecture, la technicité de sa pratique n'est-elle pas le signe d'une transformation de l'acte d'écriture ? Enfin, une réflexion sur l'augmentation réalisée devrait s'attacher non seulement aux contenus, mais aussi aux différentes formes des autorités concernées. Des questions restent ouvertes, comme la construction tangible du sens, la création de nouvelles connaissances par la liaison des données, permise par la sémantique du web, dans l'esprit de la réalisation d'un « Memex » tel qu'idéalisé par Bush ou au contraire, un appauvrissement sur le modèle des travers observés dans les média sociaux et qui avaient déjà été critiqués dans la presse télévisée par Bourdieu : peu de contenus originaux, dupliqués à l'infini dans une « circulation circulaire » de l'information. L'apport et la qualité des annotations sociales seront des aspects à discuter dans un proche avenir.

1.6. Bibliographie

[BAT 93] BATES M., « The design of browsing and berrypicking techniques for the online search interface », Online Information Review, vol. 13, no 57, 128, p. 407-424, 1993.
[BLA 15] BLANCHARD A., SABUNCU E., « Les humanités numériques, une science "plug and play" ? », dans V. CARAYOL et F. MORANDI (DIR.), Le tournant numérique des sciences humaines et sociales, Maison des sciences de l'homme d'Aquitaine, Pessac, 2015.
[BOL 90] BOLTER J., Writing Space: the computer, hypertext, and the history of writing, Lawrence Erlbaum Associates, Mahwah, 1990.
[BOU 16] BOULET V., « De la SDN à la Nuit debout : les métadonnées et les enjeux de gouvernance internationale », I2D – Information, données & documents, vol. 53, p. 35-36, 2016.
[BRO 16] BROUDOUX E., « Contours du document numérique connecté », dans C. PAGANELLI, S. CHAUDIRON et K. ZREIK (DIR.), Documents et dispositifs à l'ère post-numérique, Conférence Cide 18, Europia 2016, Paris, France, p. 7-16, 2016.
[CRO 16] CROZAT S., « Ecrire avec une machine à calculer, écrire pour une machine à calculer », I2D – Information, données & documents, vol. 53, p. 62-64, 2016.

[DAC 15] DACOS M., MOUNIER P., Humanités numériques : état des lieux et positionnement de la recherche française dans le contexte international, Rapport de recherche, Institut français, 2015.
[FRA 15] FRANCART T., « L'apport de la sémantique dans l'écriture scientifique augmentée », Quatrième séance du séminaire écrilecture augmentée sur le web pour les communautés scientifiques, Cnam, Labex Hastec, Paris, France, mars 2015.
[GOO 79] GOODY J., La Raison graphique. La domestication de la pensée sauvage, Editions de Minuit, Paris, 1979.
[HOL 92] HOLT P., WILLIAMS N. (DIR.), Computers and writing: state of the art, Intellect books, Kluwer Academic Publishers, Dordrecht, 1992.
[ILL 91] ILLICH I., Du lisible au visible. Sur l'art de lire de Hugues de Saint-Victor, Les éditions du Cerf, Paris, 1991.
[JUL 15] JULIEN Q., CITTON Y., « Manifeste pour des humanités numériques 2.0 », Multitudes, disponible à l'adresse : http://www.cairn.info/revue-multitudes-2015-2-page-181.htm, vol. 59, p. 181-195, 2015.
[KEM 16a] KEMBELLEC G., « Que voit réellement Google de la sémantique des pages web ? », I2D – Information, données & documents, vol. 53, p. 65, 2016.
[KEM 16b] KEMBELLEC G., « Le web de données en contexte bibliothécaire », I2D – Information, données & documents, vol. 53, p. 30-31, 2016.
[LED 12] LE DEUFF O., « Humanisme numérique et littératies », Semen, n° 34, p. 117-134, 2012.
[LED 15] LE DEUFF O., « Les humanités digitales précèdent-elle le numérique ? », dans I. SALEH (DIR.), H2PTM15, ISTE Editions, Londres, 2015.
[POU 16] POUYLLAU S., « Isidore Suggestion, des recommandations de lecture pour les blogs de science », I2D – Information, données & documents, vol. 53, p. 44, 2016.
[PRI 04] PRIE Y., GARLATTI S., « Méta-données et annotations dans le web sémantique », Revue I3 Information-Interaction-Intelligence, vol. 4, p. 45-68, 2004.
[RIE 12] RIEDER B., RÖHLE T., « Digital Methods. Five Challenges », dans D.-M. BERRY (DIR.), Understanding Digital Humanities, p. 67-84, Palgrave Macmillan, Basingstoke, 2012.
[SAE 15a] SAEMMER A., Rhétorique du texte numérique : figures de la lecture, anticipations de pratiques : essai, p. 61 et 163, Presses de l'Enssib, Villeurbanne, 2015.
[SAE 15b] SAEMMER A., « Pour une sémiotique critique de l'hyperlien », Quatrième séance du séminaire écrilecture augmentée sur le web pour les communautés scientifiques, disponible à l'adresse : http://www.dicen-idf.org/evenement/quatrieme-seance-du-seminaire-ecrilecture, Paris, France, juin 2015.
[SHA 48] SHANNON C., « A mathematical theory of communication », The Bell System Technical Journal, vol. 27, p. 379-423 et 623-656, juillet et octobre 1948.
[SHO 09] SHOTTON D., « Semantic publishing: the coming revolution in scientific journal publishing », Learned Publishing, vol. 22, n° 2, p. 85-94, 2009.
[SIR 13] SIRE G, La production journalistique et Google : chercher à ce que l'information soit trouvée, Thèse de doctorat, p. 339, Université Panthéon-Assas, novembre 2013.
[SOU 13] SOUCHIER E., « La "lettrure" à l'écran », Communication & langages, vol. 2012, n° 174, p. 85-108, janvier 2013.
[TRE 14] TREHONDART N., « Le livre numérique "augmenté" au regard du livre imprimé : positions d'acteurs et modélisations de pratiques », Les Enjeux de l'information et de la communication, n° 15/2, p. 23-37, 2014.

[VER 16] VERLAET L., DILLAERTS H., « L'enjeu du web de données pour l'édition scientifique », I2D – Information, données & documents, vol. 53, p. 49, 2016.
[VUI 99] VUILLEMIN A., « La lecture interactive et l'écrilecture », dans A. VUILLEMIN et M. LENOBLE (DIR.), Littérature, informatique, lecture, Presses Universitaires de Limoges, Limoges, p. 101-110, 1999.