INTRODUCTION

Philippe Caron

La Maison des sciences de l'homme et de la société de Poitiers, devenue USR (unité de service et recherche) en 2013, cherche à se doter non seulement des outils partagés dont ses équipes ont besoin mais aussi des savoir-faire qui leur sont communs. Toutes ou presque sont confrontées à la nécessité d'archiver, de valoriser, d'accroître leurs collections de données, de quelque nature qu'elles soient. Il faut donc accompagner cette mutation en essayant de deviner les chemins de l'avenir, « construire le présent sur les ruines de l'avenir, non pas sur celles du passé », c'est-à-dire discerner ce qui deviendra obsolète et encourager ce qui est vivace, même embryonnaire.

Il devient superflu d'insister sur la révolution scientifique que représentent la numérisation et les capacités de mémoire dans la création de bases de données de tous ordres. En peu de temps, tout cela est devenu routine, même si la compilation reste pavée de difficultés dont nous parlerons dans le présent volume. Le défi est désormais ailleurs : à la fois quantitatif et qualitatif. En effet la course est désormais à la constitution de très grosses ressources et les possibilités informatiques le permettent. Mais encore faut-il pouvoir les compiler d'une façon propre et intelligente. Le TGIR Huma-Num atteste en France la prise de conscience actuelle d'une sorte d'émulation constante vers les « big data » en sciences humaines. Ce changement d'échelle a une incidence automatique sur la vie académique : les recherches qui ne reposeront pas sur un fondement quantitatif appréciable risquent fort d'être frappées de caducité. Mais la constitution de ces gros ensembles nécessite un soin méticuleux, des moyens matériels, mais surtout humains qui ne sont pas à la portée de tous les chercheurs.

À supposer que ces conditions soient remplies, reste à s'entendre sur le contenu enregistré : aucun texte n'est numérisé et archivé de façon brute. Ce qui est à présent devant nous est donc autre : il s'agit de mettre en relation ces puits de savoir entre eux. Et nul ne sait comment les usagers, suivant des chemins que nous ignorons encore, exploreront ces interconnexions.

La rencontre scientifique placée sous le patronage de la Maison des Sciences de l'Homme et de la Société de Poitiers se place très exactement aux avant-postes de cette deuxième révolution que le numérique permet, couplé à internet : non seulement la croissance exponentielle des données disponibles mais la recherche de leur interconnexion.

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

Le défi est immense. Aussi ce volume s'est-il limité à en observer plus spécifiquement les effets sur le domaine des langues naturelles, de leur enregistrement et de leur annotation. Et les contributions offertes reflètent, dans leur diversité, les ingrédients de cette révolution. Elle n'est pas indolore, car tout changement de paradigme, ici à la fois quantitatif et qualitatif, requiert des efforts, de la lucidité, des moyens, parfois une conversion délicate.

Les questions qu'aborderont les contributions ci-dessous montrent différentes difficultés actuellement en co-présence :

– la mise à jour de données déjà enregistrées mais sous des formats désuets ;
– les difficultés de l'encodage « brut » proprement dit, par exemple lorsqu'un signal sonore est perturbé ou que le signe linguistique est inconnu du chercheur ;
– le choix de l'annotation cursive des données, chaque fois que de besoin, par exemple lorsqu'il s'agit de transcrire en continu un signal sonore à l'écrit et d'aligner ces deux types d'information hétérogènes. Ou d'accoler à chaque mot la catégorie grammaticale à laquelle il appartient ;
– l'adoption d'un système intégré de balises de contenu ou, au contraire, l'invention d'un réseau *ad hoc* ;
– la résistance du texte à l'étiquetage.

Nous avons réparti ces apports en trois parties successives.

La contribution de **Marie-Hélène Lay** constitue la première. Elle ouvre opportunément le volume en présentant de façon convaincante le paysage dans lequel nous évoluons dans son entier. Elle interroge ce qu'elle appelle les résistances du milieu des sciences humaines et sociales (désormais SHS) à l'utilisation et à l'annotation de grosses données informatisées. Elle indique pourtant que c'est une activité constante et ancienne de la pensée scientifique que l'annotation, probablement une des activités prototypiques de toute recherche puisqu'elle consiste à créer des catégories d'observation. Puis elle décompose les phases d'un travail de ce genre, pointant à chaque fois les enjeux : d'une part les « grosses données » qui, par leur ampleur, bouleversent notre champ et obligent, pour certaines opérations, à recourir au traitement automatique pour des annotations vraiment répétitives. Dans le cadre de cette opération, elle plaide évidemment pour l'adoption de standards qui garantissent l'interchangeabilité. Toutes ces données n'arrivent pas forcément au statut patrimonial mais la standardisation de leur annotation est un vecteur intéressant si l'on vise ce stade ultime d'une ouverture au plus vaste public. Elle distingue ensuite ces métadonnées « génériques » qui sont les accompagnateurs presque obligés des grosses données et les annotations « sélectionnées », celles que l'utilisateur va constituer lorsqu'il transforme tout ou partie de ces grosses données en corpus, c'est-à-dire lorsqu'il constitue un sous-ensemble à des fins spécifiques. Cette couche de marquage *ad hoc* représente une sorte de préemption, ce qui peut bloquer partiellement une recherche qui ne serait pas instrumentée dans le même sens mais les outils logiciels actuels permettent largement de réorienter le marquage ou de le compléter. Et Lay de plaider pour finir pour

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

cette activité modélisante assistée par ordinateur qui conduit par ses possibilités à un changement de paradigme.

La deuxième section du volume est consacrée aux aides existantes, c'est-à-dire à l'outillage actuellement mis à la disposition du chercheur.

Lou Burnard, architecte majeur de la TEI (Text Encoding Initiative), ouvre largement son chapitre en montrant que la question des normes est au cœur de l'échange de biens, de quelque nature qu'ils soient. Les grosses ressources textuelles sont des biens un peu particuliers mais elles sont soumises à cette nécessité croissante de standard qui permet de savoir au juste ce qu'on échange ou ce qu'on compare. La problématique de la TEI est en effet, dès son élaboration, la généralisation d'un standard d'encodage textuel pour tous les types de textes. Elle vise à étiqueter de façon vaste ou précise les composants d'un texte, quels qu'en soient le support ou la complexité. Elle est par ailleurs indépendante d'un système d'exploitation, d'une application quelconque, ce qui lui garantit une longévité appréciable. D'autre part cette entreprise vise à faciliter la relation entre les chercheurs et les informaticiens. À ceux qui objectent qu'un tel réseau ne peut pas s'adapter à la singularité de leur matériau d'observation, il montre bien que la TEI est utilisable à des niveaux fort différents de balisage, des étiquettes génériques pouvant être remplacées, selon le degré de précision que souhaite l'usager, par des étiquettes de type « espèce », plus proches de l'actualisation précise. Enfin il indique que l'utilisation de ce système d'encodage doit être déclarée comme telle mais qu'un utilisateur peut toujours apporter à la communauté de la TEI de nouveaux modules en fonction du balisage de nouveaux types de texte, d'où les éditions successives du manuel en ligne qui inclut à mesure des propositions venues des chercheurs qui se réclament de cette architecture. Évidemment l'adoption suppose une connaissance méticuleuse de l'objet à baliser et la comparaison terme à terme de ses composants avec l'étiquetage adéquat dans les chapitres de la TEI réservés au type de texte envisagé. Enfin, reste que « tout projet de numérisation doit forcément trouver un bel équilibre entre le faisable et l'utile car la liste des notions balisables, chacune d'importance pour un type d'analyse quelconque, risque de devenir ingérable ». L'encodage TEI apparaît donc comme un « buffet », comme le compare Burnard, dans lequel il faut savoir prélever ce dont le projet a besoin. Et il faut, dit-il, s'armer de patience pour examiner les textes à baliser afin que l'appareillage de l'annotation soit vraiment exact. Enfin il apporte une brève démonstration des possibilités offertes à l'adaptation des attributs d'une balise et il conclut que c'est sans doute cette adaptabilité qui fait la longévité de la TEI.

Une fois le travail effectué, reste à lui offrir la niche souhaitable afin qu'il trouve là sa pleine utilité et puisse entrer, éventuellement, en interconnexion avec d'autres ressources susceptibles de l'éclairer. L'une de ces « niches », ORTOLANG, est présentée par **Jean-Marie Pierrel**. ORTOLANG offre pour les ressources linguistiques, une plateforme de dépôt, d'amélioration, d'interconnexion de données. Cette proposition répond largement, de façon généreuse et ingénieuse, aux besoins d'équipes qui cherchent à la fois des outils de traitement et l'arène commune au sein de laquelle leurs ressources trouveront à la fois

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

la visibilité et l'éventuelle interconnexion. Ce qui caractérise aussi cette entreprise, c'est qu'elle sait s'appuyer sur une diversité de laboratoires et montre ainsi l'exemple du partenariat qui, souvent, permet de régler ce qu'une seule institution ne pourrait assumer par manque de moyens. Un tel consortium attire également les aides par les gages qu'il donne. Enfin il permet aussi la pérennisation des ressources, problème que l'obsolescence des matériels fait régulièrement surgir au grand dam des chercheurs. En effet une infrastructure solide, fortement sécurisée lui offre la garantie que son travail ne sera pas piraté aisément et qu'il migrera, le moment venu, vers les mises à jour logicielles requises. De plus cette entreprise se présente comme un « service spécialisé » directement en prise avec le TGIR Huma-Num. Chaque fois que celui-ci reçoit des demandes relevant des sciences du langage, ces dernières sont systématiquement transmises à ORTOLANG. Cet exemple est une source d'inspiration féconde pour la largeur de sa vision, son sens de la coopération et de l'anticipation des besoins à venir. Jean-Marie Pierrel détaille enfin les étapes de la préparation des données et les outils mis à la disposition des usagers. On voit dans ce dispositif un effort très conséquent dans l'aire francophone pour éviter la dispersion des ressources et/ou leur hétérogénéité et assurer leur pérennité.

La troisième contribution de cette deuxième partie, celle de **Nicolas Ballier**, revêt le caractère d'une introduction à R, logiciel encore assez confidentiel dans le domaine des humanités numériques car il est un monde à lui tout seul, et un monde constamment en évolution. L'auteur s'emploie à convaincre que, dans le train de traitements qu'autorise R se profile beaucoup plus qu'un outil sophistiqué : une révolution épistémologique. R est en effet à la fois un langage de programmation et un environnement de travail. Il se décline, selon les orientations, en programmes spécifiques. Ballier ne nie pas que cet outil aux multiples ramifications ne soit difficile. Mais il pointe le fait que R peut permettre de co-traiter, et donc de mettre en relation, des données très souvent traitées séparément (« R permet donc de manipuler des jeux de données, structures tabulaires hybrides comportant potentiellement des métadonnées, des extractions de données annotées, et des ré-annotations générées automatiquement »). Ce logiciel est, par nature mais aussi par son âge, dans une phase évolutive très forte, entre forces centrifuges (la diversité constamment en évolution des packages et les passerelles constamment possibles vers d'autres logiciels) et tendances centralisatrices que symbolise assez bien l'interface graphique RStudio. Il est aussi un logiciel transparent qui permet un retour critique sur son modèle statistique. Enfin c'est une architecture qui n'est pas « téléonomiquement fermée ». C'est un intégrateur dont l'auteur signale plaisamment qu'il peut à peu près tout faire de nos tâches quotidiennes à l'exception de la tasse de café de la pause ! Cette polyvalence a un coût et Ballier d'ajouter qu'il faut s'attendre au déplacement de la fracture numérique, entendons par là que ses utilisateurs distanceront à l'avenir ceux qui n'y auront pas accédé, c'est-à-dire ceux qui ne sauront pas écrire une ligne de commande pour se constituer des scripts plus spécifiques à leur recherche.

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

Une troisième partie de ce volume est consacrée au terrain et à ses difficultés. Trois entreprises, à des degrés divers d'achèvement, sont révélatrices des problèmes rencontrés.

Gabriel Bergounioux, avec toute l'expérience qu'ESLO (Enquête Sociolinguistique à Orléans) lui confère, consacre sa contribution à une opération cardinale dans l'accompagnement des données orales : la transcription. Il présente tout d'abord d'une façon partiellement historique l'archivage de données orales avec son outillage d'enregistrement puis il s'installe précisément à cet endroit-charnière : la transcription qui retransforme une deuxième fois le signal acoustique en une représentation graphique. Cette opération, dit-il, « n'est pas toujours considérée comme une variété d'annotation ». Celle-ci peut aller pourtant d'un rendu phonétique fin par toutes les ressources de l'API, par exemple, jusqu'à la construction d'un équivalent orthographique de plus en plus éloigné de la matérialité exacte du signal sonore. C'est là que l'opération de transcription s'apparente, dans une langue comme le français, à une première couche d'annotation dans la mesure où le code orthographique est déjà une première interprétation lexicale et morpho-syntaxique. Elle restitue par exemple des marques flexionnelles depuis longtemps amuïes. Toutefois les deux opérations restent distinctes dans la mesure où la transcription est le plus souvent continue, alignée sur le signal sonore tandis que l'annotation peut être discontinue, sélective, orientée vers un type d'investigation particulier. Et Bergounioux détaille alors ce qui, selon les cas, peut apparaître dans la transcription : certaines propriétés phonétiques, certains appuis phatiques du discours, certaines prononciations erronées, certaines propriétés prosodiques comme les pauses par exemple. Bref la transcription est une pratique utile qui par moments rejoint tangentiellement l'annotation. Elle reste problématique par ce qu'elle ajoute au message et soustrait parfois quelque chose. L'équipe ESLO adopte pourtant cette couche de transcription qui facilite l'accès de l'usager au matériel sonore tout en masquant une partie des spécificités phonétiques de l'original. Il en va de l'interopérabilité des données. Ainsi le chercheur est-il présent au cœur de la donnée « brute », préemptant en quelque sorte ce qu'il estime être la juste élucidation du signal.

Philippe Caron et Louise Dagenais, quant à eux, récapitulent le travail qui, sur une quinzaine d'années, a abouti à la publication du *Dictionnaire Critique de la langue française informatisé* de Jean-François Féraud, le mieux à jour des dictionnaires portatifs de la langue pré-révolutionnaire. Ils montrent que la complexité d'un dictionnaire ancien, très textualisé d'une part, très souple également dans la syntaxe et l'expression de ses composants d'autre part, leur a donné du fil à retordre. D'où une première livraison HTML de ce chef-d'œuvre de la lexicographie en mode « plein texte » géré par le moteur de recherche Philologic de Mark Olsen et placé sur les deux sites miroirs de l'ARTFL de Chicago et de l'ATILF de Nancy. Ce choix était, au cours de la dernière décennie du xx^{e} siècle, le format le plus recommandé pour les dictionnaires anciens rétroconvertis. Puis, prenant à bras-le-corps la question de l'œuvre dans son entier, ils offrent une deuxième version sur le CNRTL, version plus aboutie qui adjoint au chef-d'œuvre imprimé son supplément manuscrit d'une façon qui permet de naviguer ingénieusement de l'un à l'autre sans proje-

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

ter sur ce dernier une quelconque grille de lecture. D'autre part ils automatisent un balisage modéré qui autorise alors des recherches plus fines parce qu'elles permettent de les conditionner par plusieurs critères de choix. Les requêtes obtiennent alors moins de bruit parce qu'elles sont ciblées sur des zones plus déterminées du dictionnaire. Cette entreprise de longue haleine est un cas de partenariat réussi entre des laboratoires, des chercheurs et des sources de financement multiples.

Appartenant au même axe de recherche du FORELL que Philippe Caron (« Diachronie et variation »), **Nicolas Videau**, **Nicolas Trapateau**, **Jean-Louis Duchet** et **Sylvie Hanote** présentent l'état évolutif de deux ressources en cours d'élaboration à la MSHS de Poitiers sous l'égide du FORELL : une chaîne historique de dictionnaires de prononciation pour la langue anglaise d'une part, un corpus d'anglais oral, américain et britannique, extrait de radios. Le chapitre navigue donc entre ces deux sources, l'une indirecte, celle écrite des dictionnaires, l'autre plus directe, celle de l'oral enregistré in situ. Depuis plusieurs années, en effet, le FORELL héberge des recherches sur la phonologie accentuelle de l'anglais, qu'elles soient diachroniques ou actuelles. Le chapitre passe en revue les problèmes successivement rencontrés dans la constitution et le balisage de cette ressource et présente en parallèle les éléments de résolutions propres à chaque *medium* d'information, l'écrit lexicographique et l'oral enregistré. Du côté de l'écrit, les difficultés de l'océrisation, du côté de l'oral les innombrables bruits générés par l'enregistrement de conversations ou d'émission in situ. Les auteurs posent ensuite la question délicate de la fiabilité des données et de la représentativité du corpus recueilli au regard des phénomènes que l'on ambitionne d'étudier. Ils abordent enfin la question des conventions et de l'annotation. Certes « le choix d'une forme d'annotation standard permet d'assurer à l'avenir une meilleure transmission du corpus » mais elle demande parfois une sorte de révolution lorsque les premiers travaux ont déjà largement pris une autre direction depuis 1970. Le chapitre aborde donc ici la question de la mise à jour parfois délicate des premiers inventaires, des archivages effectués selon des besoins alors très limités. C'est ainsi qu'il faut parfois passer d'une logique « bottom-up » à une logique de type « top-down », avec toutes les difficultés qu'une telle inversion de polarité comporte. Le chapitre aborde également la question de la transcription, du stockage, de l'interrogation et pour finir de la mise à disposition de cet ensemble composite de données venues de deux *mediums* différents.

Au total, ce tour d'horizon met les chercheurs en demeure de prendre la mesure de ce qui se joue sous nos yeux et qui fait déjà passer aujourd'hui les linguistes et les autres spécialistes du texte à une pratique autrement plus exigeante de l'investigation dans leurs disciplines. Croiser des sources d'informations jusqu'alors disjointes, le faire à grande échelle, c'est éviter de prendre l'arbre pour la forêt bien sûr, mais c'est aussi faire apparaître dans l'enchaînement causal d'un phénomène complexe des indices, des réseaux d'informations qui peuvent, sur certains points, modifier l'image jusqu'alors prépondérante d'une évolution. Nous changeons d'échelle quantitative mais les multiples voies de la comparaison, de la superposition de données, finissent par donner de ce que nous regar-

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr

dons une « radiographie » beaucoup plus fine que les outils dont nous disposions avant. On pourrait risquer une comparaison : nous sommes passés de la radioscopie de papa à des outils qu'on peut utiliser simultanément comme le scanner, l'échographe, le tout couplé à l'imagerie par résonance magnétique. Le corps apparaît alors en trois dimensions.

Des remerciements spécifiques sont adressés aux membres du comité d'organisation des journées : « Données, Métadonnées des Corpus et Catalogage des objets en Sciences Humaines et Sociales » qui ont eu lieu à la Maison des sciences de l'homme et de la société de Poitiers les 6 et 7 juin 2016. Cet ouvrage est issu d'une partie des travaux qui y furent présentés. Les membres de ce comité étaient : Lydie Bodiou, Anne de Cools, Françoise Puthon-El Quassimi, François Rigalleau et Geneviève Robert [1].

Enfin nos remerciements vont particulièrement à **François Rigalleau** qui fut constamment l'aiguillon positif de l'entreprise de sa conception à son achèvement et qui a participé à l'édition autant que sa charge absorbante le lui a permis. Sa modestie seule nous a empêchés de l'inscrire au rang des coéditeurs.

1. Sans oublier le comité scientifique composé de Nicolas Ballier, Lydie Bodiou, Lou Burnard, Valentina Carbone, Frédéric Chauvaud, David Chesnet, Anne de Cools, Véronique Dasen, Callisto, Manuel Gimenes, Simos Grammenidis, Sylvie Hanote, François Lecellier, William Marx, Véronique Mehl, Martine Mespoulet, Dominique Moncond'huy, Pascale Piolino, Noël Richard, François Rigalleau, Nelly Robin, Juan David Sempere, Philippe Vendrix, Inès de La Ville, Pierre Volle, Florian Waszak.

L'enjeu des métadonnées dans les corpus textuels – Philippe Caron, Rodolphe Defiolle et Marie-Hélène Lay (dir.)
ISBN 978-2-7535-7640-7 — Presses universitaires de Rennes, 2019, www.pur-editions.fr