Méta-données : de l'accessibilité des sources à l'élaboration des objets de la recherche (données et modèles)

Marie-Hélène Lay*[1]

[1]Formes et Représentations en Linguistique et Littérature (FORELL-EA3816) – Université de Poitiers – Maison des Sciences de l'Homme et la Société 99 avenue du Recteur-Pineau 86000 Poitiers - France, France

Résumé

La communauté des sciences humaines, et tout particulièrement celle qui se retrouve au sein de CAHIER, se réjouit, à n'en pas douter, du développement des *humanités numériques* et de la disponibilité croissante de données, qu'elles soient textuelles, orales, ou multimodales. Seuls les plus jeunes d'entre nous ne se souviennent pas d'une époque où nos objets de recherche n'existaient pas au format numérique, où les plus gros des laboratoires de recherche, dans les années 1990 disposaient de moins d'information textuelle au format numérique que tout un chacun sur son laptop en 2020. Beaucoup d'énergie est aujourd'hui encore mobilisée pour la mise à disposition de nos corpus (corpus d'auteurs ou autres), de façon plus ou moins co-ordonnée, ce qui montre bien la nécessité de lieux d'échanges comme CAHIER, dont l'extension constante au cours des 10 années écoulées a permis de sensibiliser et de fédérer une part toujours plus nombreuse de notre communauté : il nous est nécessaire de partager nos expériences et de consolider une communauté de pratiques, nécessaire de construire ensemble le paradigme numérique de nos pratiques scientifiques.

Car il faut bien le dire, l'enthousiasme éprouvé à l'idée de disposer de tant de sources et ressources se trouve tempéré par deux caractéristiques constitutives du paradigme numérique : d'une part la prise de conscience du fait qu'il ne suffit pas de disposer des textes " saisis ", par exemple en format .txt, .odt ou .doc pour en faire des données de la recherche au format numérique disponibles pour la communauté, d'autre part la prise de conscience d'un phénomène parfois appelé " infobésité " : la masse de données est telle qu'on ne sait pas toujours comment en faire bon usage, bon usage scientifique s'entend, qui suppose que les données de la recherche aient été patiemment élaborées pour cerner au plus près, de la façon la plus honnête et la plus exhaustive possible, les éléments permettant de nourrir réflexion et argumentation autour d'une question posée. Ce qui est nouveau dans l'environnement numérique, c'est que l'on peut apporter une réponse unifiée à toutes ces problématiques : la pratique de l'annotation, c'est-à-dire l'adjonction de métadonnées, permet de structurer les fonds documentaires, de localiser les ressources et de documenter les perspectives scientifiques, de transformer des données dites " brutes " en données de la recherche.

Mais ce point doit sans doute être mieux élucidé et la pratique de l'annotation mieux exposée. En effet, il est certain que l'annotation, et plus particulièrement l'annotation des contenus,

*Intervenant

est extrêmement coûteuse en moyens humains : il peut donc sembler vain de s'y contraindre, comme il est vain de chercher à utiliser des annotations " standard " pour aborder un point délicat de nos recherches. Notre réponse-réflexe est alors : " vu le temps et les ressources nécessaires, mieux vaut se concentrer sur notre recherche elle-même que sur toutes ses étapes bien trop coûteuses par rapport à notre objet d'étude ".

Présentées ainsi, comment ne pas souscrire à ces réticences ? Mais comment ne pas remarquer par ailleurs que ces données appelées des vœux des chercheurs à la fin du 20ème et au début du 21ème siècle ne semblent pas avoir massivement rénové la recherche de ceux qui sont d'abord des littéraires et des spécialistes de leurs auteurs. Il me semble raisonnable d'envisager que la raison en est assez simple : sauf à entrer dans des démarches d'analyse de type textométrique, le chercheur accédant aux ressources textuelles au format numérique reste dans une approche pré-numérique des textes.

Car de fait, le constat s'impose, les données patiemment élaborées et rendues disponibles sont largement sous exploitées. La simple disponibilité de ressources gigantesques, d'un *tas de données en tas* ne suffit pas à les rendre accessibles de façon pertinente, on risque de s'y " noyer ", même si elles sont proprement cataloguées et indexées, d'un point de vue de " bibliothèque ". Certes, le phénomène n'est pas nouveau, ce dont on trouve des formulations plaisantes, comme " Vous croulez sous vos données ? c'était déjà le cas du temps de Voltaire "1, phénomène dont nous avons tous fait l'expérience lorsque nous recherchions une citation précise dans un paquet de notes prises de façon trop peu organisée, sur des petites fiches cartonnées dispersées ça et là : un volume d'information assez restreint ne permet donc pas forcément d'échapper au sentiment de noyade. Aujourd'hui comme alors, il nécessaire d'organiser sources et ressources pour en permettre la localisation et le partage2, mais aussi (et surtout?) l'usage.

Parallèlement à la disponibilité de ressources se pose donc la question de leur exploitation. Leur volume est devenu tel, que la contrainte de l'outillage informatique s'impose d'elle-même : la consultation séquentielle, ou plus précisément, dans le cas présenté ici, la lecture linéaire par un humain n'est certainement pas la bonne méthode. Outre la transposition à l'environnement numérique des pratiques documentaires traditionnelles, permettant l'organisation des ressources et l'accès aux documents3 sont apparus des outils assurant l'accès direct au *contenu* des documents et le traitement de l'information localisée : c'est le cas d'outils comme TXM, bien représenté au cours de cette journée.

C'est un point de vue un peu différent qui sera adopté ici (complémentaire des autres éléments évoqués ci-dessus). L'*annotation de contenu* ne sera pas d'abord présentée comme une pratique unifiée permettant la réutilisabilité des données, ou une étape utile avant le recours à des outils de traitement statistique. L'*annotation manuelle de contenu* sera réinscrite dans la lignée des pratiques pré-numériques de l'enrichissement des textes, enrichissement au sens ou une information nouvelle est associée à une information initiale : c'est ce que nous faisons communément en préparant une lecture critique, en produisant diverses versions d'un texte, en corrigeant des épreuves pour un éditeur, en corrigeant des travaux d'étudiants. Dans les pratiques d'annotation manuelle dont il est question ici, les métadonnées sont généralement insérées dans le document lui-même4.

Annoter, c'est, formulé de la façon la plus simple, ajouter des " notes " à un document, ajouter une information B à une information initiale A. C'est enrichir des données (information A) par d'autres données (information B). Ces autres données sont appelées métadonnées. En des termes plus techniques, on peut emprunter à Fort (2012, p. 175) la définition suivante : " l'annotation recouvre à la fois le processus consistant à apposer (ad-) une note sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger a priori de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif) ".

Notre objectif sera ici de montrer en quoi la pratique de l'annotation manuelle n'est pas une contrainte supplémentaire imposée par la mise à disposition des informations au format numérique, mais en quoi elle est un outil de façonnage des données de la recherche, un outil d'expérimentation des modèles que nous élaborons, un laboratoire de formulation et de test, un outil de pertinent venant à l'appui de nos heuristiques6.

Références Bibliographiques

André V. & Canut E. (2010), " Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français) ", *Pratiques*, n○ 147-148, p. 35-51.

Baude O. (2006), *Corpus oraux, guide des bonnes pratiques*,

http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Corpus_Oraux_GBP%202006_version_imprimee.pdf

Baude O. (2007), " Aspects juridiques et éthiques de la conservation et de la diffusion des corpus oraux ", *RFLA XII-1*, p. 85-97.

Bilger M. (2000), *Corpus - Méthodologie et applications linguistiques*, Paris, Champion.

Broudoux E. & Scopsi C. (2011), " Métadonnées sur le web : les enjeux autour des techniques d'enrichissement des contenus ", *Études de communication*, n○ 36, p. 9-22.

Burnard L. (1995), " Text Encoding for Information Interchange. An Introduction to the Text Encoding Initiative ", *TEI Document. Proceedings of the Second Language Engineering Conference*, TEI J31.

Burnard L. (2005), " Metadata for corpus work ", *Developing linguistic corpora: a guide to good practice,* Wynne, M. (ed.), Oxford, Oxbow Books, p. 30-46.

http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter3.htm

Burnard L. (2011), " Qu'est-ce que l'annotation et pourquoi en parle-t-on de manière si inquiétante ? ", *École thematique annotation de données langagières,*

http://www.lattice.cnrs.fr/IMG/pdf/burnard-annotation.pdf

Burnard L. (2012), " Encoder l'oral en TEI : démarches, avantages, défis.... ", *Séminaire de l'institut du numérique,*

Mots-Clés: exploitation des ressources textuelles, annotation manuelle, élaboration des données de la science, heuristiques