humanités numériques

Humanités numériques

1 | 2020
Varia

Une épistémologie numérique des disciplines est-elle possible ? Étude d'un corpus de revues francophones en sciences humaines et sociales

Is a Disciplinary Digital Epistemology Possible? Studying a Corpus of French-Language Human and Social Sciences Journals

Max Beligné, Isabelle Lefort et Sabine Loudcher

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/revuehn/361

Éditeur
Humanistica

Référence électronique
Max Beligné, Isabelle Lefort et Sabine Loudcher, « Une épistémologie numérique des disciplines est-elle possible ? Étude d'un corpus de revues francophones en sciences humaines et sociales », *Humanités numériques* [En ligne], 1 | 2020, mis en ligne le 01 janvier 2020, consulté le 30 juillet 2020.
URL : http://journals.openedition.org/revuehn/361

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.

humanités
numériques

1 | 2020
Varia

Une épistémologie numérique des disciplines est-elle possible ? Étude d'un corpus de revues francophones en sciences humaines et sociales

Is a Disciplinary Digital Epistemology Possible? Studying a Corpus of French-Language Human and Social Sciences Journals

Max Beligné, Isabelle Lefort et Sabine Loudcher

Résumés

Les principales revues francophones de sciences humaines et sociales ont été numérisées et mises en ligne. Les outils des humanités numériques permettent de se saisir en théorie de cet ensemble important de données pour proposer de nouvelles lectures épistémologiques des disciplines. Les potentialités de ce champ de recherche sont détaillées afin de mieux comprendre les modalités et les intérêts multiples d'une épistémologie numérique, tout d'abord générale, puis, plus spécifiquement, appliquée aux revues. Malgré l'importance de ces potentialités, un état des lieux sur la question fait apparaître que les travaux existants et prenant appui sur les revues francophones sont encore très peu développés. Afin de comprendre cette situation, trois verrous majeurs sont explicités : la constitution d'un corpus représentatif, l'obtention et la préparation des données et, enfin, le choix précis de la méthode. Au-delà de l'objectif de renouvellement épistémologique disciplinaire, des dimensions politiques sous-jacentes de ce projet peuvent être mises en avant.

The most important French-language human and social sciences journals have been digitised and published online. Digital Humanities tools make it possible in theory to obtain this important dataset and propose new epistemological readings of the disciplines. The potential of this research field is detailed, firstly in order to better understand the methods and multiple interests of a general digital epistemology and

secondly to more specifically explore a digital epistemology applied to journals. Despite this considerable potential, a state of the art about digital epistemology based on French-language journals shows that there are only few research works. In order to understand this fact, three issues must be addressed: the choice of a representative corpus, the data acquisition and the data cleaning, and finally the selection of the method. Beyond the objective of a new epistemological reading of a discipline, the underlying political dimensions of this project can be highlighted.

Entrées d'index

MOTS-CLÉS : sciences de l'information et de la communication, constitution de corpus, discipline, épistémologie, revue, revue francophone

KEYWORDS: information and communication sciences, corpus building, discipline, epistemology, journal, francophone journal

Introduction

1 L'objectif de cet article est de montrer tout d'abord l'intérêt des outils des humanités numériques pour travailler l'épistémologie d'une discipline. Avant même de parler d'épistémologie numérique, définir ce que recouvre l'épistémologie est complexe. Dans un premier temps, cette dernière est ici entendue dans une perspective large et plutôt commune dans la sphère francophone : celle de l'étude critique des sciences. Le sujet de cet article étant centré sur les disciplines, les manuels d'épistémologie facilement disponibles pour chaque champ disciplinaire sont un point d'entrée commode pour se représenter pratiquement de quoi il est question. L'objectif de ces manuels est de proposer des lectures globales et critiques de l'évolution des disciplines. Or, les outils mobilisés par les humanités numériques peuvent traiter de larges quantités de données et, par leurs capacités computationnelles, produire rapidement des visualisations synthétiques. Plusieurs champs de recherche existent déjà comme la cartographie scientifique[1], la bibliométrie, la lexicométrie, l'analyse de réseaux, la fouille de textes... Par rapport à ce vaste ensemble, il est nécessaire d'essayer de mieux préciser ce qui peut être entendu par épistémologie numérique et de développer ce que peut apporter une telle voie de recherche.

2 Plus spécifiquement, l'objet de recherche retenu pour cet article est celui d'un corpus de revues francophones de sciences humaines et sociales (SHS). Ce choix repose sur plusieurs raisons. Il y a sans conteste un aspect pratique qui provient du fait que de nombreuses revues sont directement disponibles au format numérique. Contrairement aux livres, il s'agit le plus souvent de données ouvertes et donc plus facilement accessibles. Le processus de mise en ligne des contenus numérisés ayant été largement soutenu depuis maintenant plus d'une dizaine d'années, il est possible de considérer que globalement[2] toutes les grandes

revues des disciplines ont été numérisées. Toutefois, il faut rappeler que ce ne sont pas les périodiques mais les ouvrages qui ont été considérés historiquement comme les références en SHS. Il est par conséquent nécessaire de détailler pourquoi les revues peuvent aussi être considérées comme un support légitime pour effectuer des relectures épistémologiques des disciplines à l'aide des outils numériques et cela au delà des aspects pratiques qui viennent d'être précédemment mentionnés.

3 Un point délicat du sujet est celui de la justification de la limitation aux revues francophones. En effet, on peut facilement admettre que penser la science à cette échelle n'est pas vraiment pertinent depuis déjà un moment. Pour mieux comprendre ce choix, il faut souligner que des outils comme VOSviewer [3], permettant directement de faire des requêtes dans les grandes bases internationales comme le Web of Science [4] ou Scopus [5], n'ont pas d'équivalent en France. Aucun outil n'offre la possibilité d'enquêter aussi directement et profondément sur les portails Persée [6], OpenEdition Journals [7], Cairn [8], Gallica [9] ou Érudit [10]. Certes, les métadonnées peuvent être moissonnées via diverses méthodes, mais toute étude d'ensemble est complexe, en particulier s'il y a la volonté d'analyser à partir de portails multiples le texte intégral ou les références bibliographiques. Par conséquent, il y a une certaine légitimité scientifique à vouloir s'attaquer à ces données qui ont été très peu exploitées. De plus, se restreindre à la sphère francophone permet d'éviter des difficultés liées aux traitements multilingues.

4 Il est nécessaire de mentionner qu'il y a très peu d'études qui ont réellement approfondi cette question d'une épistémologie numérique disciplinaire à partir des revues francophones en SHS. Bien qu'il soit possible de citer de nombreuses études abordant seulement une revue [11], il n'existe pas à notre connaissance de recherches abordant, de front et de manière coordonnée, la représentation d'une discipline par ses revues francophones et la relecture épistémologique par les outils numériques. Il est possible de citer quelques essais comme celui de Quoc-Tan Tran (2017) concernant l'émergence des humanités numériques, mais sur des disciplines ayant un ancrage historique plus ancien, aucune référence n'a été trouvée. Cette absence pose d'autant plus question que les potentialités de cette voie de recherche sont démontrées dans la première partie de cet article. Il convient donc d'expliciter plus précisément les difficultés d'une épistémologie numérique. Les trois principales difficultés développées dans la deuxième partie de cet article (la constitution d'un corpus représentatif, l'obtention et la préparation des données, le choix des méthodes) sont issues d'un travail de doctorat mené en épistémologie de la géographie [12] mais elles se présenteraient à tous ceux qui voudraient se lancer dans une aventure intellectuelle similaire sur un autre champ disciplinaire.

Les potentialités d'une épistémologie numérique disciplinaire

De l'épistémologie traditionnelle à l'épistémologie numérique

5 Avant d'aborder les potentialités d'une épistémologie numérique, il est nécessaire de définir ce que recouvre cette expression et ce qui la différencie d'une épistémologie plus traditionnelle. Historiquement, le travail de ceux qui se sont revendiqués « épistémologues » a surtout consisté en un examen critique des textes dans une perspective qualitative et érudite. Un premier stade d'utilisation de l'outil numérique peut se voir, par exemple, dans des travaux qui, après avoir fait de nombreux comptages à la main, synthétisent graphiquement leurs résultats grâce à un tableur. Bien que cette épistémologie utilise le numérique, ce n'est pas elle qui est ici visée par l'expression « épistémologie numérique ». Cette dernière désigne plutôt un deuxième stade d'utilisation du numérique où les recherches exploitent de manière plus évidente les capacités computationnelles des ordinateurs en travaillant sur des automatisations des recherches ou en faisant des mises en relation qui seraient très difficilement réalisables à la main. Il est malheureusement particulièrement complexe de définir un seuil précis qui permettrait de classer de manière évidente des travaux dans ce champ de l'épistémologie numérique.

6 L'usage de cette expression se retrouve de manière assez avant-gardiste chez Matthieu Valette avec une signification particulière :

> L'épistémologie numérique consiste en un réinvestissement linguistique de l'épistémologie à la française. Il s'agit d'étudier les textes scientifiques au moyen des outils d'analyse documentaire et statistique développés ces trente dernières années dans le domaine du traitement automatique du langage et de la linguistique de corpus et d'une méthodologie proprement linguistique. (Valette 2006)

7 Il est important de préciser que l'étude de M. Valette se concentre sur un seul auteur. Il est difficilement envisageable dans le cas d'une recherche impliquant plusieurs auteurs de ne pas prendre en compte ce paramètre dans l'analyse. Cela est d'autant plus vrai que l'on peut s'appuyer sur le travail de Camille Roth (2008) qui a popularisé le concept de « communauté épistémique » reliant les réseaux d'auteurs et de concepts. Ainsi, l'épistémologie numérique peut rejoindre aussi bien les épistémologies internalistes soucieuses des procédures cognitives de production des connaissances que celles plutôt externalistes attentives aux contextes socio-politiques.

8 De manière générale, il est possible de prendre en compte n'importe quel indicateur permettant de revenir à la définition large de l'étude critique des sciences. Le plus connu est sans doute les citations par l'intermédiaire de la bibliométrie, mais de nombreux autres indicateurs sont possibles : les plans d'article (Abbott et Barman 1997), les figures, les lieux… Dans cet article, l'approche très orientée sur le texte définie par

M. Valette sert de point d'entrée pour ne pas se perdre dans la diversité des approches possibles, mais les autres indicateurs (les auteurs, les citations…) ne sont pas exclus.

Les intérêts d'une épistémologie numérique

9 Comme l'exprime Damon Mayaffre (2006), « la lecture empathique ou intuitive des textes […] demeure encore aujourd'hui majoritaire en SHS ». Que peut apporter par conséquent une épistémologie numérique ? Tout d'abord, une rupture par rapport à la quantité de textes qui peut être prise en charge dans les analyses. Il n'est pas question de remettre en cause les larges connaissances des érudits, mais il est facile de comprendre qu'au-delà d'une centaine d'articles, il est difficile pour un humain de mettre en relation de manière assez fine et systématique les textes. L'ordinateur permet de changer d'échelle en prenant en considération facilement des milliers de textes. Alors que la tradition épistémologique a tendance à se concentrer sur quelques textes de référence, le numérique peut ouvrir le champ d'investigation.

10 Il y a le plus souvent dans le processus informatique une égalité de traitement des textes. Les relations ne sont plus faites de manière intuitive mais deviennent systématiques. Les processus automatiques (comme les calculs de proximités, les spécificités, les analyses factorielles…) mettent partiellement de côté les cadres *a priori* de l'analyste et permettent de découvrir le corpus sous une forme pas forcément attendue. Cette mise de côté de la grille de lecture n'est que partielle car elle reste en fait présente au niveau de la constitution du corpus, du choix des outils, de l'interprétation. Mais le processus de calcul est fondamentalement ascendant. Il part des données et peut donc produire de l'inattendu, ce qui rend ce type d'analyse particulièrement intéressant. Les résultats « bruts » sont bien souvent mal nommés car il y a une construction sous-jacente, mais quelque chose dans cette expression renvoie à un petit pas de côté essentiel par rapport à l'omnipotence de nos filtres théoriques dans la lecture traditionnelle d'un texte.

11 Un autre aspect fondamental est que le numérique permet de s'affranchir de la linéarité qui prédomine largement dans toute lecture oculaire de textes par les chercheurs. L'ensemble des praticiens en analyse de données textuelles montre bien la complémentarité des lectures tabulaires et réticulaires par rapport à ce mode initial de lecture [13]. La méthodologie permise par la plupart des logiciels d'analyse de données textuelles est d'alterner dans les analyses les visions synthétiques avec les retours dans le texte pour mieux comprendre les résultats obtenus. Au-delà de l'analyse textuelle, ce retour aux données réalisé après leurs agrégations est toujours profitable pour confirmer les résultats obtenus et limiter les fausses interprétations.

12 La quantification permet également de préciser l'importance ou non de certains phénomènes ou évolutions. En effet, traditionnellement, chaque auteur, en fonction de sa grille de lecture, peut sélectionner dans un ensemble de textes les passages les plus significatifs pour sa démonstration. Si cette pratique permet de construire des cadres interprétatifs précieux, il est difficile après coup d'estimer ce qui a été surévalué. Le passage par le quantitatif n'apporte pas forcément de réponses car il y a souvent des questions de seuils : à partir de quand, par exemple, une

rupture devient-elle une révolution ? En revanche, la quantification couplée à une saisie plus globale permet d'ouvrir le champ de la discussion en fournissant de nouveaux éclairages.

13 Enfin, l'aspect quantitatif oblige à justifier le mieux possible les choix réalisés. C'est une pratique loin d'être toujours facile car, dans la réalité, certains choix sont parfois faits pragmatiquement pour avancer avec des justifications qui peuvent se faire *a posteriori*. Malgré ces cas, il reste une exigence fondamentale de clarification et de justification pour que les résultats soient les plus reproductibles et pertinents possible. C'est cette dimension construite qui est au final importante. Les résultats, avant même toute interprétation, dépendent de la constitution du corpus et des outils employés. Ils ne peuvent se comprendre qu'en fonction des hypothèses de travail et du construit scientifique réalisé pour répondre à ces questionnements. C'est par conséquent dans la justification des constructions que se trouvent le nerf de l'épistémologie numérique et le vecteur de discussions argumentées potentiellement nombreuses.

Un objet spécifique mais porteur : une discipline par ses revues

14 En prenant comme support les revues, l'histoire relue est loin d'être secondaire car les quatre fonctions traditionnelles des périodiques scientifiques – l'enregistrement, la diffusion, l'archivage et la certification [14] – expliquent que les revues sont « en 350 ans d'histoire [...] un média structurant de la communication scientifique entre pairs » (Boukacem 2014). D'après Vincent Larivière, les revues ont actuellement encore deux fonctions importantes : fédérer les communautés et fournir une hiérarchie [15]. Elles sont par conséquent des points d'entrée pertinents pour étudier les disciplines. Cela est d'autant plus intéressant qu'elles permettent d'aborder les interactions entre les dimensions scientifiques (les problématiques, les objets, les méthodes) et les dimensions sociales (les institutions, les collectifs, les personnes) qui constituent les disciplines.

15 Une des caractéristiques fondamentales de la revue par rapport au livre est sa périodicité. Cet aspect temporel permet de mettre au jour des continuités et des ruptures dans des séries temporelles. Les permanences reflètent des lignes éditoriales plus ou moins explicites. La revue peut être un véritable outil pour faire durer, voire imposer un programme de recherche dans le temps long [16]. Par exemple, le périodique l'*Espace géographique* a été le promoteur de l'analyse spatiale dans la géographie française. Les ruptures, quant à elles, reflètent l'évolution des idées et des acteurs. Des phénomènes de changements de lignes éditoriales ou des processus d'innovations plus ou moins durables peuvent être mis au jour. L'étude d'Isabelle Lefort (2011) sur *Géocarrefour* montre bien ces évolutions à l'échelle d'une revue. De plus, par leurs dimensions périodiques, « les revues scientifiques et savantes bruissent des polémiques du moment » (Voisenat 2013), comme le rappelle le projet *Bérose* [17] qui s'est attelé à la construction d'une base de données pour l'anthropologie et l'ethnographie. Si cet aspect est très variable d'une revue à l'autre, il est indéniable que ce média peut permettre d'approcher

ce qui a constitué l'actualité scientifique des disciplines, et plus particulièrement les objets, les méthodes, les débats qui ont animé une ou plusieurs communautés de chercheurs sur une période donnée.

16 Historiquement, les revues sont plutôt des dispositifs qui sont fortement localisés. Elles permettent d'aborder ainsi une dimension de l'épistémologie mise en avant par Donna Haraway (1988), celles des savoirs situés. Par exemple, Numa Broc (2001) montre bien l'opposition qui a existé entre l'école de Grenoble et l'école de Paris entre 1910 et 1940. Il n'est pas étonnant que la *Revue de géographie alpine* soit créée en 1913. Comme l'affirme Jean-Louis Tissier (1991) : « Fonder une revue c'est, implicitement, affirmer que les périodiques existants ne peuvent assurer la diffusion régulière et le reflet fidèle d'une nouveauté scientifique ». Plus généralement, la revue incarne l'importance de la notion de réseau d'acteurs dont l'intérêt dans la science a bien été montré par Bruno Latour dans *Science in Action* (1987). Sur ces notions, il faut citer le travail récent d'Hervé Ferrière et Isabelle Thiébau (2017) qui, partant d'une perspective critique et de l'application à un cas concret, arrive à cette conclusion : « Après en avoir testé la validité, l'efficacité du concept réseau et acteur a été démontrée et s'avère être la notion la mieux adaptée à la singularité de notre objet ».

17 Par leurs aspects synchroniques (il existe plusieurs revues au même moment), il est possible de faire émerger des thèmes qui ont traversé l'ensemble des revues, mais aussi d'autres qui n'ont touché spécifiquement que certaines revues. Au-delà des thèmes, il est possible de faire des comparaisons inter-revues sur une multitude de facteurs. Par exemple, François Briatte compare l'entre-soi éditorial de toutes les revues du portail Cairn [18]. Cette approche n'est en soi pas incompatible avec la dimension diachronique car si les indicateurs choisis évoluent dans le temps, des groupes de périodiques avec des trajectoires différentes peuvent être mis en lumière. Les décalages temporels peuvent, par exemple, révéler des phénomènes de circulation des connaissances avec des processus variés d'appropriation ou de résistance. L'aspect comparatif aide à l'interprétation des résultats car il est plus aisé de discerner ce qui relève des tendances générales, d'évolutions spécifiques ou de cas particuliers.

18 Enfin, comme l'a montré Pierre Bourdieu (2001), les revues structurent avec d'autres dispositifs les champs disciplinaires. Il y a indéniablement des phénomènes de hiérarchies et de spécialisations. En effet, chaque comité de rédaction s'inscrit dans un champ, ce qui lui laisse des marges de choix variables pour tirer son épingle du jeu. Les analyses factorielles citées par P. Bourdieu sont des outils privilégiés pour mettre au jour des cartographies [19] des disciplines. Plus globalement, des graphes construits à partir de n'importe quel indicateur permettant de mesurer une proximité entre les revues (les auteurs, les citations, les lieux…) peuvent aboutir à des représentations d'une discipline par ses revues.

19 Maintenant que l'intérêt d'une épistémologie numérique disciplinaire à partir des revues est mieux précisé et conforté dans ses pertinences et ses possibilités, il est important de considérer ses difficultés.

Les difficultés d'une épistémologie numérique disciplinaire

La constitution d'un corpus représentatif

20 Il n'existe pas en géographie et à notre connaissance dans les SHS françaises, de recherches à l'échelle d'une discipline sur les revues réalisées dans une optique d'épistémologie numérique. Une difficulté majeure est de constituer un corpus représentatif. Une issue méthodologique est celle adoptée dans le cadre du projet *INTERCO-SHS*[20] par Johan Heilbron et Anaïs Bokobza. Ces auteurs font référence à des « revues-cœur » en s'appuyant sur une enquête du Centre national de la recherche scientifique (Heilbron et Bokobza 2015). Le caractère multidisciplinaire de leur étude dilue certainement un peu le problème de la représentativité du corpus qui serait à n'en point douter crucial dans le cas d'une étude disciplinaire. Quelles revues choisit-on pour représenter une discipline ? L'image finale dépendra de ce choix initial qu'il est difficile de justifier. Le problème, bien montré par F. Briatte (2008), est que plusieurs classements ont essayé de définir des « revues-cœur » et que les recoupements entre ces classements produisent au mieux un nombre très limité de résultats robustes. David Pontille et Didier Torny (2010) vont plus loin en montrant comment les classifications des revues ont évolué en fonction des objectifs qui sous-tendent ces classements. Par conséquent, prendre comme base ces classements pose question.

21 Une première réponse naïve à cette question de sélection des revues d'une discipline est de se contenter de la classification des sites. Néanmoins, celle-ci n'est pas sans poser problème. Par exemple, faut-il prendre la *Revue européenne des migrations internationales* qui est classée en « Démographie » sous le portail Persée alors qu'elle est classée en « Géographie » sous Cairn. S'il est décidé d'ajouter quelques revues de catégories externes, comment le justifier par rapport aux autres revues des mêmes catégories ? Finalement, faut-il prendre le périmètre le plus large possible de revues ? N'y a-t-il pas un périmètre restreint de revues reconnues comme ayant joué un rôle plus important dans la discipline ?

22 Une réponse classique à cette question de sélection des revues les plus importantes à étudier est la bibliométrie avec l'étude des revues les plus citées. Or, cette utilisation pose de nombreuses questions particulièrement en SHS comme le souligne Ghislaine Filliatreau (2008) : « il existe dans ces disciplines des journaux qui ont une portée nationale et/ou un faible indice d'impact sans pour autant devoir être écartés, car ils jouent un rôle scientifiquement important pour leur communauté ». Les débats qui ont animé la géographie suite au classement de l'Agence d'évaluation de la recherche et de l'enseignement supérieur en 2008 (Piron 2019) montrent à quel point le sujet est sensible. Écarter des revues de la reconnaissance disciplinaire, c'est risquer d'entrer dans des fortes polémiques.

23 Pour finir, Ghislaine Filliatreau (2008) montre que les citations en SHS n'ont pas le même rôle que celles en sciences de la matière et de la vie où « on peut dire – en simplifiant beaucoup – que les connaissances

nouvelles sont concurrentes entre elles au moment où elles sont publiées, puis qu'elles deviennent cumulatives quand elles ont été vérifiées – et donc citées – par les travaux d'autres équipes ». *A contrario*, dans les SHS, « on procède plutôt par filiations sélectives, ou par juxtapositions de résultats » (Filliatreau 2008, 61-66), ce qui ne donne pas le même statut aux études bibliométriques. De plus, en pratique, les citations sont particulièrement difficiles à étudier dans les revues françaises de SHS. En effet, il faut d'abord choisir un corpus de référence, ce qui pose à nouveau la question du corpus. Puis, il faut que les citations soient harmonisées. Il n'est pas évident que, d'une revue à l'autre, elles soient écrites de la même manière. Tout un travail de détection et d'indexation a été mené par les plateformes [21] mais celui-ci n'étant pas complètement abouti, les résultats ne peuvent être que partiels. Une idée peut être de travailler avec des bibliographies plus globales comme la Bibliographie géographique internationale [22] pour la géographie. Cette optique ne permet pas néanmoins de résoudre le problème de l'importance des revues. Un périodique peut avoir publié beaucoup d'articles, tout en n'étant pas si important pour la discipline [23].

24 Une autre idée pour s'en sortir est d'imaginer une enquête pour demander à la communauté quelles revues ont joué un rôle important dans l'histoire de la géographie. Toutefois, cette solution est biaisée car les résultats dépendent beaucoup des personnes interrogées. C'est en quelque sorte reporter le problème de la sélection des revues sur les personnes à interroger. Cette question de la représentativité peut enfin faire penser à celle de l'échantillonnage. Cette perspective nécessite une première phase de deuil par rapport au fantasme de pouvoir choisir finement les revues représentatives et de pouvoir les traiter intégralement. En effet, un des intérêts initiaux mentionné en travaillant avec les outils des humanités numériques était de pouvoir analyser de larges quantités de données. Il est dommage dans ce cadre de devoir revenir à une technique classique de l'échantillonnage qui ne règle la question de la représentativité que de manière très théorique.

25 Pour finir, dans le meilleur des cas, il est possible de trouver un référencement des périodiques jugés historiquement importants [24]. Après s'être assuré que la liste a été constituée par une personne qui connaissait bien le champ disciplinaire, après avoir suffisamment cherché pour affirmer qu'il n'existe pas d'autres listes similaires dans lesquelles on peut avoir autant confiance, après avoir rappelé que cette liste ne sera pas prise pour argent comptant, après avoir démontré toute la difficulté qu'il y a à choisir un corpus représentatif par ailleurs, alors il nous semble scientifiquement raisonnable de prendre cette liste comme base. Un problème immédiat est que ce référencement contient des revues qui n'ont pas été numérisées. Il est possible d'imaginer une enquête à partir de cette liste pour sélectionner les 10 références qui semblent les plus pertinentes en prenant soin de choisir des géographes de divers courants. Mais pour finir de montrer la difficulté, la liste, qui est publiée dans la revue électronique *Cybergeo*, ne cite pas *Cybergeo* alors que cette revue est reconnue par nombre de géographes comme un support important [25]. La rajouter à la liste initiale pose le problème de justifier une telle exception. Sortir de la complexité d'établir un corpus représentatif d'une discipline est au final loin d'être simple comme le montre cet exemple développé sur la géographie.

Obtention et préparation des données

26 Suivant les portails, les articles les plus récents peuvent être payants, mais, après une certaine durée, appelée barrière mobile, ils sont librement accessibles et téléchargeables en fichiers PDF. La barrière mobile étant actuellement limitée à cinq ans sur le portail Cairn et étant souvent inexistante sur le portail OpenEdition Journals, une grande quantité d'articles est donc librement disponible. Le réseau Mir@bel [26] permet de visualiser facilement les différents accès pour chaque revue. Le portail Isidore [27] permet de faire des requêtes sur les principaux portails simultanément. Des entrepôts comme HAL, très utilisés actuellement, pourraient être mobilisés mais ils ne l'ont pas été dans cette étude. En effet, leur échelle de référence n'est plus la revue mais l'article. Le projet scientifique étant de relire l'épistémologie à partir des revues, de leurs trajectoires, il nous a semblé que la prise en compte des entrepôts relevait d'une autre logique.

27 Dans la récupération qui a été menée, il est nécessaire de distinguer les métadonnées (titre, auteur, date…), qui peuvent être facilement moissonnées, des données (texte) qui sont beaucoup plus difficilement récupérables. Les métadonnées, par exemple sur les auteurs, cachent des situations très variées. Certains portails ont fait un véritable travail de désambiguïsation des noms en adoptant un référentiel commun, en alignant parfois leur référentiel avec d'autres référentiels déjà existants et en proposant des premiers outils d'analyses [28]. D'autres portails n'ont pas fait ce travail, ce qui provoque une hétérogénéité des données. En imaginant que l'on profite du travail fait par le portail Persée sur la question de la désambiguïsation des auteurs en partenariat avec l'Agence bibliographique de l'enseignement supérieur, il est nécessaire de préciser que l'on se place alors dans une logique un peu opportuniste en constituant un corpus avec ce qui est disponible. Cette approche est fort différente de celle développée dans la partie précédente qui visait la construction d'un corpus *ad hoc*. Cette distinction (entre corpus *ad hoc* et corpus généraux), bien développée par Adrien Barbaresi (2015), permet d'appréhender une autre manière d'accéder à la discipline. Les résultats issus de corpus généraux sont toujours critiquables car dépendants de corpus moins finement construits, mais ils ne sont toutefois pas forcément inintéressants.

28 Une difficulté cachée du travail sur les revues réside dans l'accès aux données (texte). Sur quelques articles, il est bien entendu possible de faire des copier-coller directement à partir des portails concernés. Pour une demande plus importante et dans l'état actuel de la loi, il faut signer une convention avec la plateforme concernée. De prime abord, cela peut sembler une simple formalité, mais comme les services des plateformes ont par ailleurs beaucoup de travail et comme les conventions passent par les services administratifs des universités qui ne considèrent pas ce genre d'affaire comme leur priorité, il nous a fallu attendre par deux fois plus d'un an avant d'obtenir les conventions signées à force de relance. L'avantage de cette méthode est de pouvoir ensuite accéder à des données dans un format structuré, habituellement en XML-TEI.

29 Dans un premier temps, l'avantage n'est pas immédiatement perceptible car les formats récupérés ne sont pas directement lisibles par les outils d'analyse, ce qui nécessite une phase de transformation des données. Cette situation s'explique par le fait que les plateformes ont été conçues dans un objectif de diffusion de connaissances et non de recherche en fouille de texte. Dans un deuxième temps, la structuration des données est un vrai avantage car elle permet de faire une sélection de ce qui sera analysé et d'améliorer les résultats. En effet, certaines données, comme les notes de bas de page, peuvent apporter plus de bruit que d'informations importantes dans une analyse lexicométrique. Si ces données ont fait l'objet d'une balise spécifique, il est assez facile de les retirer. En revanche, les données qui n'ont pas été balisées sont beaucoup plus difficiles à retirer. Par exemple, sur des données issues du portail Persée, les répétitions en haut de page du titre de l'article ou du nom de l'auteur, souvent sous forme réduite, demandent quelques compétences en programmation pour être enlevées.

30 Un problème devant être soulevé est celui des formats utilisés par les différentes plateformes. En travaillant sur plusieurs revues (mais aussi souvent au sein d'une même revue car une partie ancienne des textes peut être sous le portail Persée et la partie récente sur les portails Cairn ou OpenEdition Journals), le chercheur est amené à travailler sur plusieurs schémas, ce qui pose la question de leurs interopérabilités. Le plus simple est de créer un schéma ou une base de données commune, ce qui nécessite un travail non négligeable [29]. Si, par exemple, quelques revues sont sur le portail Gallica et n'ont pas de balisage précis des notes de bas de page et des bibliographies, il faudra pour les intégrer soit abandonner les informations des autres revues sur ces éléments, soit essayer de créer un nouveau balisage permettant une homogénéité de traitement. Il existe certes des programmes comme Grobid [30] permettant de transformer des fichiers PDF en données structurées. Mais ces programmes ont surtout été entraînés sur des articles de sciences exactes qui ont un formatage standard. Face à des articles avec des formats moins homogènes, les résultats sont moins bons ou il faut passer beaucoup de temps à les entraîner de nouveau.

Choix des méthodes

31 L'épistémologie numérique telle que définie en introduction renvoie à une pluralité de champs et, par conséquent, de méthodes. Le travail réalisé sur la revue *Cybergeo* montre un exemple de méthodologie multiple abordant à la fois les lieux, les thèmes, les bibliographies et les auteurs (Pumain 2016). Toutefois, le fait de vouloir analyser plusieurs revues qui relèvent de portails différents engendre des problèmes importants comme ceux développés dans la partie précédente. De plus, comme le présente Denise Pumain (2016), il s'agit essentiellement d'une application permettant « de jouer librement des analyses cartographiques et sémantiques et de trouver vos chemins parmi toutes ces références ». Le travail d'interprétation épistémologique n'a pas eu lieu de manière approfondie.

32 Dans le cas d'un chercheur n'ayant pas une équipe à sa disposition et qui souhaiterait réaliser en aval un travail de comparaison des résultats obtenus avec ceux de l'épistémologie traditionnelle, il est nécessaire de

choisir les indicateurs étudiés et les méthodes utilisées. Cela est particulièrement frustrant pour un épistémologue – qui a pour objectif de produire un discours un peu général – de devoir se spécialiser de la sorte. Cela est d'autant plus décevant que beaucoup de champs à explorer par les revues semblent intéressants. Cependant, cela se comprend aisément à partir d'un exemple. Tout d'abord, en choisissant un indicateur simple et unique : les auteurs des articles. Ensuite, en prenant une méthode déjà implémentée : par exemple, celle de F. Briatte sur le degré d'entre-soi éditorial [31]. On peut penser qu'en prenant un indicateur aussi simple et une méthode déjà implémentée, il serait assez facile de réaliser une épistémologie numérique disciplinaire. Il n'en est rien car, comme nous l'avons mentionné, travailler à partir de plusieurs portails pose le problème de l'hétérogénéité des données et, par conséquent, de la désambiguïsation d'auteurs. Dans un cadre scientifique, il est nécessaire de s'intéresser aux méthodes déjà existantes sur le sujet avec une littérature assez riche. Après avoir choisi et justifié le choix de la méthode la plus adaptée, il faut l'appliquer et l'évaluer. Finalement, beaucoup de temps s'est écoulé, et il n'est envisageable que de produire des résultats encore à interpréter. Dans un tel cadre, impossible de multiplier les indicateurs et les méthodes.

33 Une donnée importante qui complexifie la recherche est la donnée temporelle. En effet, il est tentant de voir comment un indicateur évolue dans le temps. Néanmoins, il n'est pas évident de justifier le découpage temporel retenu. En effet, pour une revue, il est possible de prendre la succession des directeurs comme périodisation. Pour plusieurs revues, il est plus délicat de justifier d'une périodisation adéquate. De plus, dans de nombreux domaines, par exemple la détection de communautés d'auteurs ou de thématiques, les algorithmes qui prennent le mieux en compte cette dimension temporelle sont complexes et pas évidents à utiliser.

34 Enfin, le choix d'une méthodologie est important car il détermine des arrière-plans théoriques. Dans le cas du critère simple des auteurs, si le chercheur se place dans une perspective de travail surtout orientée par les réseaux, l'arrière-plan théorique sera dominé par les travaux de Bruno Latour, alors que dans une approche plus globale qui mettra en avant les hiérarchies entre les revues, l'arrière-plan théorique sera dominé par les travaux de Pierre Bourdieu. Certes, ces deux approches ne sont pas exclusives l'une de l'autre, mais leur articulation est loin d'être évidente. L'exemple des auteurs n'est pas un cas particulier. Une étude plus centrée sur le lexique demande de se pencher sur la qualité de la reconnaissance optique de caractères (ROC) pour les revues numérisées. Les outils d'analyses de données textuelles sont également nombreux avec des arrière-plans théoriques distincts (Lejeune 2017). Des analyses sur les bibliographies demandent aussi, comme on l'a vu, un important travail en amont et leur utilisation en SHS ne va pas de soi.

Conclusion

35 Ainsi, en l'état actuel, il est prometteur mais difficile d'aborder une discipline à partir de ses revues francophones dans le cadre d'une épistémologie numérique. Les chantiers esquissés sont nombreux : travail sur la représentation d'une discipline, amélioration de l'accès aux données, construction de référentiels et de schémas communs entre les portails, amélioration du texte et des bibliographies, articulation avec les outils d'analyse… Plus largement, on peut voir dans cette énumération les conditions de possibilité pour que les humanités numériques puissent se pencher de manière approfondie et réflexive sur les disciplines. Cette dynamique a déjà été anticipée par les principaux portails qui ont amélioré significativement certaines données. À n'en point douter, la question des référentiels et de la désambiguïsation des auteurs sera un jour réglée. Le portail Isidore permet déjà de faire des requêtes transversales, mais il est difficile de savoir quel pourcentage d'auteurs est bien aligné. Encore une fois, il est possible d'expérimenter sur le sujet, mais si le chercheur se place dans une perspective plus scientifique, il est amené, comme on l'a vu, à déployer des méthodologies plus lourdes qui l'obligent rapidement à se spécialiser. Cette double tension entre rigueur scientifique et expérimentation, d'une part, et volonté de lecture globale et spécialisation nécessaire, d'autre part, doit être prise en compte pour comprendre les balbutiements de ce champ de recherche.

36 Pour revenir à l'expérience du doctorat qui a été à l'origine de cet article, la mise en évidence de ces difficultés a été une étape réflexive déterminante qui a conduit à adopter une problématique épistémologique plus précise permettant ainsi de mieux justifier un corpus et une méthodologie. Cela a permis de mettre de côté le problème important de la représentation disciplinaire. Les chercheurs, comme Kristine Lund, qui font actuellement le choix de garder la focale d'une discipline se tournent en général vers les bases internationales, mais ont des difficultés à intégrer une grande part de la recherche francophone (Lund *et al.* 2017). Même si cette préoccupation n'a pas été initiatrice du projet, elle peut rejoindre une dimension politique revendiquée, par exemple, par Marin Dacos [32], pour rendre visible un monde scientifique autre que celui dominé par les sciences dures, l'anglais et les bases de données internationales. Sans être dans une vision simpliste, puisque de plus en plus de revues francophones sont présentes dans Scopus, des référencements, comme celui réalisé par JournalBase [33], montrent que de telles bases sont loin de couvrir la majorité des périodiques francophones et sont particulièrement limitées dès que l'on veut aborder la dimension diachronique. Analyser la bibliodiversité favorisée par l'*open access* et ce qu'elle apporte épistémologiquement est un défi non négligeable.

37 Pour finir, une remarque de Michel Pierssens (2005) permet de saisir un autre enjeu important de ce travail : « le lecteur se livre à la pure exploration de données, le *data mining*, libéré de toute temporalité. Dès lors, qu'importe qu'un article ait été choisi par une revue plutôt que par une autre ? ». S'il est nécessaire d'avoir une vision nuancée du *data mining* qui peut, bien entendu, intégrer une dimension temporelle, il est vrai que la navigation Web tend parfois à faire oublier les contextes historiques et spatiaux d'élaboration des connaissances. Un tel projet peut

servir à rétablir ces dimensions importantes pour la compréhension du travail scientifique. Les quelques difficultés bien mises en avant par cet article peuvent servir par conséquent d'état des lieux, non seulement à quelques épistémologues, mais – espérons-le – à une communauté plus large de praticiens des humanités numériques soucieux de telles problématiques.

Bibliographie

Abbott, Andrew et Emily Barman. 1997. « Sequence Comparison Via Alignment and Gibbs Sampling : A Formal Analysis of the Emergence of the Modern Sociological Article ». *Sociological Methodology* 27 (1) : 47-87. https://doi.org/10.1111/1467-9531.271019.

Barbaresi, Adrien. 2015. *Ad Hoc and General-Purpose Corpus Construction from Web Sources*. PhD thesis, ENS Lyon. https://tel.archives-ouvertes.fr/tel-01167309/document.

Boelaert, Julien, Nicolas Mariot, Étienne Ollion et Julie Pagis. 2015. « Les aléas de l'interdisciplinarité ». *Genèses* 100-101 (3) : 20-49.

Boukacem, Chérifa. 2014. « Les couleurs de la publication scientifique. Mutations dans la sous-filière de la revue scientifique STM, analysées par les industries culturelles ». *Les Enjeux de l'information et de la communication* 15 (1) : 49-65.

Bourdieu, Pierre. 2001. *Science de la science et réflexivité : cours du Collège de France, 2000-2001*. Paris : Raisons d'agir.

Briatte, François. 2008. « Comparaison inter-classements des revues en sociologie-démographie et en science politique ». *Bulletin de méthodologie sociologique. Bulletin of sociological methodology* 100 (octobre) : 51-60.

Broc, Numa. 2001. « École de Grenoble contre école de Paris : les Alpes enjeu scientifique ». *Revue de Géographie Alpine* 89 (4) : 95-105. https://doi.org/10.3406/rga.2001.3059.

Ferrière, Hervé et Isabelle Thiébau. 2017. « Apports et limites des concepts d'acteurs et de réseau dans l'étude d'une revue coloniale : les *Annales d'hygiène et de médecine coloniales* (1898-1940) ». *Clio@Thémis. Revue électronique d'histoire du droit* 33.

Filliatreau, Ghislaine. 2008. « Bibliométrie et évaluation en sciences humaines et sociales : une brève introduction ». *Revue d'histoire moderne contemporaine* 55-4bis (5) : 61-66.

Guérin-Pace, France, Thérèse Saint-Julien et Anita W. Lau-Bignon. 2012. « Une analyse lexicale des titres et mots-clés de 1972 à 2010 ». *L'Espace géographique* 41 (1) : 4-30.

Haaf, Susanne, Alexander Geyken et Frank Wiegand. 2014. « The DTA "Base Format" : A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources ». *Journal of the Text Encoding Initiative* 8.

Haraway, Donna. 1988. « Situated Knowledges : The Science Question in Feminism and the Privilege of Partial Perspective ». *Feminist Studies* 14 (3) : 575. https://doi.org/10.2307/3178066.

Heilbron, Johan et Anaïs Bokobza. 2015. « Transgresser les frontières en sciences humaines et sociales en France ». *Actes de la recherche en sciences sociales* 210 (5) : 108-121.

Joseph, Bernadette. 1997. « Liste chronologique des principaux périodiques français de géographie ». *Cybergeo : European Journal of Geography* 19. https://doi.org/10.4000/cybergeo.5399.

Latour, Bruno. 1987. *Science in Action : How to Follow Scientists and Engineers through Society*. Cambridge : Harvard University Press.

Lebart, Ludovic et André Salem. 1994. *Statistique textuelle*. Paris : Dunod.

Lefort, Isabelle. 2011. « Une revue de géographie sur la place lyonnaise : géographie d'un périodique ». *Géocarrefour* 86 (3-4) : 201-211. https://doi.org/10.4000/geocarrefour.8411.

Lejeune, Christophe. 2017. « Analyser les contenus, les discours ou les vécus ? À chaque méthode ses logiciels ! » Dans *Les Méthodes qualitatives en psychologie et sciences humaines de la santé*, édité par Marie Santiago-Delefosse et Maria del Rio Carral, 203-224. Paris : Dunod.

Lund, Kristine, Heisawn Jeong, Sebastian Grauwin et Pablo Jensen. 2017. « Une carte scientométrique de la recherche en éducation vue par la base de données internationales Scopus ». *Les Sciences de l'éducation – Pour l'Ère nouvelle* 50 (1) : 67-84.

Mayaffre, Damon. 2006. « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? ». *Texto*, 15-25.

Moravec, Michelle. 2015. « The Historian's Altmetrics : How Can We Measure the Impact of People in the Past ? ». *LSE Impact Blog*. http://blogs.lse.ac.uk/impactofsocialsciences/2015/08/13/what-would-a-historians-altmetrics-look-like/.

Pierssens, Michel. 2005. « Revues savantes : quel avenir ? ». Dans *Le Savoir des livres*, édité par Benoît Melançon, 71-94. Montréal : Presses de l'université de Montréal. https://doi.org/10.4000/books.pum.1342.

Pontille, David et Didier Torny. 2010. « Revues qui comptent, revues qu'on compte : produire des classements en économie et gestion ». *Revue de la régulation. Capitalisme, institutions, pouvoirs* 8 (2nd semestre). https://doi.org/10.4000/regulation.8881.

Pumain, Denise. 2016. « Les réseaux de Cybergeo ». *Cybergeo : European Journal of Geography*. http://journals.openedition.org/cybergeo/27665.

Roth, Camille. 2008. « Coévolution des auteurs et des concepts dans les réseaux épistémiques : le cas de la communauté "zebrafish" ». *Revue française de sociologie* 49 (3) : 523-558.

Tissier, Jean-Louis. 1991. « 1891 : Rappels ». *Annales de géographie* 100 (561) : 513-520.

Tran, Quoc-Tan. 2017. « The Emergence of the Digital Humanities : An Epistemological Cartography of Thematic Issues in French Academic Journals ». Communication à *AIUCD 2017 Conference & 3rd EADH Day*, Rome.

Valette, Mathieu. 2006. « La genèse textuelle des concepts scientifiques : étude sémantique sur l'œuvre du linguiste Gustave Guillaume ». *Cahiers de lexicologie* 89 : 125-142.

Voisenat, Claudie. 2013. « Bérose ou l'émergence des savoirs ethnographiques ». *Port Acadie : Revue interdisciplinaire en études acadiennes / Port Acadie : An Interdisciplinary Review in Acadian Studies* 24-25-26 : 418-423. https://doi.org/10.7202/1019148ar.

Zuckerman, Harriet et Robert K. Merton. 1971. « Patterns of Evaluation in Science : Institutionalisation, Structure and Functions of the Referee System ». *Minerva* 9 (1) : 66-100.

Notes

1 Traduction de *science mapping* assez problématique pour un géographe attaché à une certaine définition de la cartographie.

2 Car il peut exister des exceptions dans certaines disciplines et cela dépend aussi du seuil à partir duquel on considère une revue comme grande.

3 http://www.vosviewer.com.

4 https://webofknowledge.com.

5 https://www.elsevier.com/solutions/scopus.

6 http://www.persee.fr/.

7 http://journals.openedition.org/.

8 https://www.cairn.info/.

9 http://gallica.bnf.fr/.

10 https://www.erudit.org.

11 Par exemple, Guérin-Pace, Saint-Julien, et Lau-Bignon (2012) ; Boelaert *et al.* (2015)...

12 Ce doctorat est préparé à l'université de Lyon sous la direction d'Isabelle Lefort, professeure en géographie, et de Sabine Loudcher, professeure en informatique. Il est soutenu financièrement par la région Auvergne-Rhône-Alpes.

13 Voir par exemple le livre de référence : Lebart et Salem (1994).

14 D'après Zuckerman et Merton (1971).

15 Voir par exemple cette intervention de 2015 : https://www.canal-u.tv/video/fmsh/nouvelles_formes_d_edition_vincent_lariviere.21144.

16 Dans une perspective kuhnienne, imposer un paradigme dans la durée, même si ces deux concepts, les programmes de recherches développés par Irme Lakatos et les paradigmes développés par Thomas Kuhn, doivent être différenciés.

17 http://www.berose.fr.

18 http://politbistro.hypotheses.org/2482.

19 Dans un sens métaphorique et toujours problématique pour un géographe.

20 International Cooperation in the Social Sciences and Humanities : http://interco-ssh.eu.

21 Notamment au niveau du portail OpenEdition Journals avec la création de Bilbo (*Bibliographical Robot*) : http://lab.hypotheses.org/955.

22 http://bgi-prodig.inist.fr/.

23 On retrouve ici une distinction classique entre quantité et qualité. Il est possible de penser également aux *Altmetrics* pour pouvoir aborder ce problème mais les recherches sur le passé sont dans ce domaine émergentes et pas directement applicables à notre objet d'étude : voir, par exemple, Moravec (2015).

24 Par exemple, pour la géographie, voir Joseph (1997).

25 Cette situation peut s'expliquer par le fait de la forme entièrement électronique de la revue qui en faisait à l'époque un support à part. De plus, le concept de périodique utilisé dans l'article en question peut être discuté pour ce type de publication à périodicité irrégulière. Enfin, Madame Joseph en tant que conservatrice de la bibliothèque de l'Institut de géographie qui, à cette date, n'avait aucun poste de consultation mis à disposition du public, a dressé un état des lieux qui s'est limité au papier du fait des circonstances de son élaboration.

26 Mutualisation d'informations sur les revues et leurs accès dans les bases en ligne : http://www.reseau-mirabel.info/.

27 https://www.rechercheisidore.fr/.

28 Par exemple, voir http://data.persee.fr/.

29 En France, il n'existe pas à notre connaissance de travail sur la question. À l'étranger, voir Haaf, Geyken, et Wiegand (2014).

30 https://github.com/kermitt2/grobid.

31 Le code est directement disponible sur http://politbistro.hypotheses.org/2482.

32 https://bn.hypotheses.org/11585.

33 https://journalbase.cnrs.fr/.

Auteurs

Max Beligné

UMR 5600 EVS et EA 3083 Eric, université Lyon 2, Lyon, France
Max Beligné est doctorant à l'université de Lyon.
beligne.max@gmail.com

Isabelle Lefort

UMR 5600 EVS, université Lyon 2, Lyon, France
Isabelle Lefort est professeure à l'université de Lyon.
ialefort@gmail.com

Sabine Loudcher

EA 3083 Eric, université Lyon 2, Lyon, France
Sabine Loudcher est professeure à l'université de Lyon.
sabine.loudcher@univ-lyon2.fr

Droits d’auteur

Les contenus de la revue *Humanités numériques* sont mis à disposition selon les termes de la Licence Creative Commons Attribution 4.0 International.