Données, sciences humaines et méthodes quantitatives

Ruffini-Ronzani, Nicolas; de Valeriola, Sébastien; Bertrand, Paul

Published in:
Revue belge de philologie et d'histoire

Publication date:
2024

100 • 2022

REVUE BELGE DE PHILOLOGIE ET D'HISTOIRE

FASC. 4: HISTOIRE

AFL. 4: GESCHIEDENIS

BELGISCH TIJDSCHRIFT VOOR FILOLOGIE EN GESCHIEDENIS

100 • 2022

SOCIÉTÉ POUR LE PROGRÈS DES ÉTUDES PHILOLOGIQUES ET HISTORIQUES
fondée en 1874

Président: Jean-Marie Duvosquel, 44 (boîte 1), avenue Adolphe Buyl, 1050 Bruxelles.
Secrétaire général: Denis Morsa, 29/3, avenue Émile Vandervelde, 1200 Bruxelles.
Trésorier: David Guilardian, 326 (boîte 5), Avenue Brugmann, 1180 Bruxelles.

L'organe de la Société est la *Revue belge de Philologie et d'Histoire*, recueil trimestriel dont le tome I est paru en 1922.

Het *Belgisch Tijdschrift voor Filologie en Geschiedenis* wordt uitgegeven door de Société. Het Tijdschrift werd gesticht in 1922.

REVUE BELGE DE PHILOLOGIE ET D'HISTOIRE
BELGISCH TIJDSCHRIFT VOOR FILOLOGIE EN GESCHIEDENIS

Website : http://www.rbph-btfg.be

Publié avec l'aide financière de la Politique scientifique fédérale (belspo), du Fonds de la Recherche scientifique - FNRS et de la Fondation universitaire. La Bibliographie de l'Histoire de Belgique est établie avec l'aide des Archives générales du Royaume, de la Commission royale d'Histoire et de la Loterie nationale.

Uitgegeven met de steun van het Federaal Wetenschapsbeleid (belspo) en van de Universitaire Stichting. De Bibliografie van de Geschiedenis van België komt tot stand dankzij de steun van het Algemeen Rijksarchief, van de Koninklijke Commissie voor Geschiedenis en van de Nationale Loterij.

Données, sciences humaines et méthodes quantitatives : typologie, programme et pistes de réflexion

Sébastien DE VALERIOLA,
Université libre de Bruxelles

Paul BERTRAND
Université catholique de Louvain

& Nicolas RUFFINI-RONZANI
Université de Namur & Archives de l'État à Namur

En 1956, dans la conclusion d'un article intitulé *Les mathématiques de l'homme*, l'anthropologue Claude Lévi-Strauss note :

> Faut-il en conclure qu'entre les sciences exactes et naturelles, d'une part, les sciences humaines et sociales, de l'autre, la différence est si profonde, si irréductible, qu'on doit perdre tout espoir d'étendre jamais aux secondes les méthodes rigoureuses qui ont assuré le triomphe des premières ? Une telle attitude [...] nous paraît entachée d'un véritable obscurantisme, en prenant ce terme dans son sens étymologique : obscurcir le problème au lieu de l'éclairer(1).

Ce constat volontairement provocateur formulé il y a plus d'un demi-siècle pourrait encore, dans une certaine mesure, figurer sous la plume de certains scientifiques actuels, tant l'application des méthodes quantitatives à de vastes corpus documentaires (textuels, mais aussi iconographiques et archéologiques) se fait avec difficulté dans la plupart des disciplines relevant des sciences humaines et sociales. C'est avec l'ambition de décloisonner les approches qualitatives et quantitatives – lesquelles sont souvent plus complémentaires que concurrentes – que les auteurs du présent article ont obtenu la création auprès du Fonds national de la Recherche scientifique (FRS-FNRS) du groupe de contact « Les humanités des données ». Celui-ci visera, d'une part, à stimuler le recours aux méthodes quantitatives, en illustrant leurs apports par des exemples concrets et en formant à des techniques novatrices, et, d'autre part, à réfléchir aux spécificités des données employées dans le champ des sciences humaines et sociales. Cet article constitue une version remaniée de la communication présentée à l'occasion de la rencontre inaugurale de ce groupe de contact, tenue à l'Université libre de Bruxelles le 7 novembre 2022. Il souhaite à la fois replacer la création de celui-ci dans son

* Les auteurs tiennent à remercier l'ensemble des participants à la rencontre inaugurale du groupe de contact F.R.S.-FNRS « Les humanités des données », et plus particulièrement Fanny Martin (Université de Namur), Johan Van der Eycken (Archives de l'État) et Clément Bert-Erboul (Université libre de Bruxelles), qui ont accepté de jouer le rôle de répondants lors de cette séance.

(1) Claude LÉVI-STRAUSS, « Les mathématiques de l'homme », dans *Esprit*, t. 243, 1956, 10, p. 525-538, ici p. 531.

arrière-plan historique, esquisser une typologie des méthodes quantitatives utilisées en sciences humaines et sociales, et définir des pistes de travail pour l'avenir.

Des premiers projets quantitatifs aux *data-driven humanities*

L'application de méthodes quantitatives à des problématiques issues des sciences humaines et sociales n'a rien d'une idée nouvelle. Les prémices s'en rencontrent dès les lendemains de la Seconde Guerre mondiale, lorsque le théologien italien Roberto Busa, équipé de machines IBM et de milliers de cartes perforées, se lance dans la conception de l'*Index thomisticus*, dans le but d'obtenir une concordance de l'ensemble des œuvres de Thomas d'Aquin(2). S'inspirant de ce modèle pionnier, d'autres équipes de recherche lui emboitent le pas au cours des années suivantes, avec une même volonté de rassembler des corpus textuels massifs et d'y appliquer des méthodes quantitatives à grande échelle. En Belgique, dès les années 1960, ces équipes sont celles du LASLA, fondée à Liège par Louis Delatte, et du CETEDOC, créé à l'Université catholique de Louvain sous l'impulsion de Paul Tombeur, appuyé, pour le volet diplomatique, par Léopold Genicot(3). Les opérations appliquées restent alors relativement simples sur le plan méthodologique (tris, filtres, sommes, moyennes, etc.) et correspondent aux limites des machines mobilisées. Elles suffisent néanmoins à répondre aux objectifs initialement définis au sein de ces projets, c'est-à-dire calculer des fréquences et obtenir des collocations pour de vastes ensembles de textes.

Depuis cette époque de pionniers, trois grandes évolutions se sont fait jour dans le champ du numérique en général. Elles correspondent, plus globalement, à des transformations qui ont affecté la société et la recherche scientifique dans leur ensemble. La première d'entre elles est intervenue au cours de la seconde moitié du XX^{e} siècle, à un moment où les ordinateurs se sont considérablement perfectionnés, avec une puissance de calcul en constante croissance, et ont progressivement fait leur entrée dans le quotidien des scientifiques, d'abord dans un cadre institutionnel, puis au niveau individuel. Suite à ce bouleversement, une série d'approches quantitatives nouvelles sont devenues accessibles aux chercheurs. À l'inverse de leurs collègues issus d'autres champs disciplinaires, les chercheurs en sciences humaines et sociales ont assez peu saisi ces opportunités, même si des travaux ont bien évidemment dérogé à la règle. Ceux-ci sont toutefois demeurés ponctuels et

(2) Voir, dans sa version informatisée, Roberto BUSA, dir., *Index thomisticus* [en ligne]. URL : https://www.corpusthomisticum.org/it/index.age. Au sujet de cette entreprise, on se reportera dernièrement à Pierre MOUNIER, *Les humanités numériques : une histoire critique*, Paris, Éditions de la Maison des Sciences de l'Homme, 2018, p. 21-43 (DOI : 10.4000/books.editionsmsh.12006).

(3) L'histoire de ces premières initiatives belges dans le domaine des humanités numériques reste à écrire. À l'heure actuelle, on doit se contenter de « points d'étape » livrés par l'un ou l'autre participant à ces projets. Voir, par exemple, Joseph DENOOZ, « La banque de données du Laboratoire d'analyse statistique des langues anciennes (LASLA) », ainsi que Élisabeth LALOU & Monique PAULMIER-FOUCART, « Visite au CETEDOC de l'Université de Louvain-la-Neuve. Interview de Paul Tombeur », dans *Le médiéviste et l'ordinateur*, t. 33, 1996, p. 14-20 et p. 25-29 (DOI : 10.3406/medio.1996.1440 et 10.3406/medio.1996.1442).

n'ont guère suscité d'émulation. Au titre de ces exceptions, on pourrait pointer les enquêtes de Jean-Philippe Genet et d'Alain Guerreau dans le secteur de l'histoire médiévale. Leurs contributions, qui visaient à mettre au point des méthodes factorielles ou de classification automatique sur des calculatrices programmables, étaient d'autant plus remarquables que les seules machines alors accessibles étaient d'une utilisation difficile(4). Deux disciplines se sont néanmoins montrées plus perméables au quantitatif. Il s'agit de la linguistique, un champ disciplinaire au sein duquel de nombreuses initiatives se sont rapidement développées en matière de traitement automatique du langage – au LASLA déjà mentionné, par exemple –, et, dans une moindre mesure, de l'archéologie, où des méthodes de classification automatique ont commencé à être mises en application dès la fin des années 1960(5). Les raisons de ces particularismes tiennent sans doute aux échanges que ces disciplines entretenaient depuis longtemps avec les sciences exactes.

La popularisation du web au cours des années 1990 et 2000 est à l'origine d'une deuxième évolution. Cet avènement a en effet bouleversé le quotidien de millions d'utilisateurs, tout en rendant les sociétés de plus en plus connectées. Dans le champ des sciences humaines et sociales, cette période voit la constitution de dizaines de corpus (textuels, iconographiques, sonores, etc.) sur support physique ou en ligne – avec parfois un passage de l'un à l'autre, comme dans le cas du *Thesaurus diplomaticus* conçu sur CD-Rom en 1997 avant de migrer sous format web et ouvert avec les *Diplomata Belgica* en 2015(6). Au fil des ans, cette tendance n'a fait que s'exacerber, avec une volonté de plus en plus notable d'encourager le partage des données, leur interopérabilité et leur pérennisation. Si les corpus disponibles en ligne sont aujourd'hui innombrables, rares sont les entreprises qui ont soumis ceux-ci à des traitements quantitatifs à large échelle(7). L'utilisation de telles techniques n'était d'ailleurs généralement pas envisagée par les concepteurs de ces corpus, qui concevaient ces derniers comme un décalque numérique des versions papier, à ceci près qu'ils étaient plus maniables et plus faciles à consulter. L'objectif était alors d'accélérer le dépouillement et le repérage d'informations, sans imaginer que la création de corpus pourrait engendrer une évolution structurelle dans les méthodes de travail.

Depuis les années 2000, une troisième évolution – que d'aucuns qualifient

(4) Par exemple, Alain GUERREAU, « Analyse factorielle au moyen d'une calculatrice programmable de poche » et Jean-Philippe GENET, « Comment faire des analyses factorielles et de la classification automatique », dans *Le médiéviste et l'ordinateur*, t. 2, 1979, p. 10-12 et 12-13 (DOI : 10.3406/medio.1979.905 et 10.3406/medio.1979.907).

(5) François DJINDJIAN, « La classification automatique en archéologie », dans *Le médiéviste et l'ordinateur*, t. 7, 1982, p. 7-9 (DOI : 10.3406/medio.1982.984).

(6) Philippe DEMONTY, Walter PREVENIER & Paul TOMBEUR, dir., *Thesaurus diplomaticus*, Bruxelles, Commission royale d'Histoire, 1997, CD-Rom ; Thérèse DE HEMPTINNE, Jeroen DEPLOIGE, Jean-Louis KUPPER & Walter PREVENIER, dir., *Diplomata Belgica. Les sources diplomatiques des Pays-Bas méridionaux au Moyen Âge*, Bruxelles, Commission royale d'Histoire, 2015 [en ligne]. URL : https://www.diplomata-belgica.be

(7) Dans le cadre d'un projet porté par le consortium COSME dont les résultats seront accessibles en 2024, Coraline Rey s'est essayée à réaliser un inventaire des outils et corpus numériques disponibles en ligne et portant uniquement sur la période médiévale. Les chiffres obtenus sont particulièrement impressionnants, puisqu'ils se montent à environ un millier de sites web et d'outils numériques différents.

de *data revolution* – prend progressivement place[8]. À l'instar des pouvoirs publics et des entreprises privées, les scientifiques ont pris conscience de la valeur ajoutée que peuvent représenter la récolte systématique, le traitement et la mise en œuvre de quantités massives de données. En conséquence, ces dernières ont désormais acquis une importance majeure, voire stratégique, dans la plupart des secteurs de la société. C'est dans ce contexte que sont apparues les méthodes dites *data-driven*, lesquelles sont regroupées dans une discipline parfois appelée *data science*, ou « science des données », qui relève à la fois de la statistique, de l'informatique et des sciences de l'information. Les contours effectifs de cette discipline tout comme ses relations avec certains domaines des mathématiques (notamment les statistiques) et de l'informatique (comme l'intelligence artificielle) suscitent aujourd'hui de nombreux débats au sujet de l'éventuelle autonomie de ce champ disciplinaire. Qu'il existe un certain flou autour des lignes de démarcation entre cette discipline et d'autres ne nous paraît pas constituer, en soi, un problème fondamental : nous envisageons ici les méthodes *data-driven* dans un sens très large, en y incluant l'ensemble des approches quantitatives développées dans un cadre statistique ou algorithmique.

Affirmer, comme on le fait parfois, que ces techniques sont apparues récemment n'est en réalité pas tout à fait exact, car nombre d'entre elles existent depuis des décennies. Ainsi, les fameux « réseaux de neurones » – des modèles d'apprentissage supervisé souvent présentés comme radicalement neufs[9] – ont été introduits et développés dès les années 1950 et 1960. En fait, si les méthodes *data-driven* ont actuellement le vent en poupe, c'est sans doute à la suite des évolutions susmentionnées. Grâce aux ressources informatiques aujourd'hui disponibles, avec des ordinateurs à la puissance de calcul élevée, ces techniques permettent de traiter les quantités massives de données générées au quotidien dans différents secteurs de la société. L'impact de l'utilisation de ces méthodes sur la recherche scientifique est monumental. Pour certains auteurs, comme le très respecté James Gray, disparu en 2007, nous serions entrés dans un nouveau paradigme scientifique : le *data-intensive paradigm*[10]. De la médecine aux sciences de gestion en passant par la physique, l'avènement de ces méthodes a fait évoluer un grand nombre de disciplines. Si les sciences humaines ont, pour l'instant, été relativement peu affectées par ce *data turn*, il ne fait guère de doute que la plupart de ces techniques peuvent être appliquées en histoire, en archéologie, en sciences politiques, en communication, etc., avec des résultats parfois très enthousiasmants. Nous sommes, nous le pensons, au début d'une nouvelle ère pour les méthodes quantitatives en sciences humaines et sociales.

(8) Rob KITCHIN, *The Data Revolution. Big Data, Open Data, Data Infrastructures and Their Consequences*, Londres, Sage, 2014 (DOI : 10.4135/9781473909472).

(9) On y reviendra plus loin, mais les méthodes de reconnaissance automatique des écritures, disponibles à travers des outils tels que *Kraken* ou *Transkribus*, fonctionnent selon ce principe.

(10) Tony HEY, Kristin TOLLE & Stewart TANSLEY, *The Fourth Paradigm. Data-intensive Scientific Discovery*, New York, Microsoft Research, 2009.

Une typologie des méthodes quantitatives en sciences humaines

En sciences humaines et sociales, comme dans d'autres disciplines, les méthodes *data-driven* peuvent être mobilisées en vue d'atteindre des objectifs bien différents. Pour mieux comprendre l'apport de ces techniques, nous proposons au cours des pages suivantes d'en dresser une typologie. Celle-ci s'organise en fonction des quatre étapes principales qui composent le processus de recherche. Les frontières entre chacune de ces classes ne sont évidemment pas tout à fait hermétiques : certaines techniques peuvent intervenir à plusieurs moments dans un même projet de recherche ou être mobilisées pour traiter plusieurs étapes en même temps.

Acquisition des données

Un premier ensemble de méthodes vise à faciliter l'acquisition des données, en déléguant à l'ordinateur une partie de l'activité de dépouillement des sources. Dans le champ des études historiques et philologiques, la mise en œuvre de telles techniques se matérialise, en particulier, dans la reconnaissance automatique des écritures, ou *Handwritten Text Recognition* (HTR). Cette technique, qui implique la mise en place d'une forme d'apprentissage supervisé, permet d'extraire de vastes quantités de texte au départ de simples photographies de manuscrits ou de documents d'archives. Ces dernières années, de telles méthodes ont été mises en œuvre, avec succès, dans des dizaines de projets, portant sur les registres du Trésor des chartes, les livres aux délibérations capitulaires de Notre-Dame de Paris, les lettres de rémission modernes des anciens Pays-Bas ou encore les archives notariales françaises contemporaines (11).

Les sources textuelles ne sont toutefois pas les seules à pouvoir faire l'objet de telles approches. Des chercheurs les ont par exemple récemment mises en œuvre sur des documents iconographiques, en faisant en sorte que l'ordinateur reconnaisse automatiquement des détails précis au sein d'un corpus donné de peintures (par exemple, la présence de bâtiments qui seraient figurés à l'arrière-plan des œuvres) (12). La machine leur a ainsi permis de constituer, aisément et rapidement, un sous-corpus de représentations iconographiques. On comprend sans difficulté l'intérêt que pourrait revêtir une telle approche dans le champ de l'histoire de l'art une fois que l'on pourra l'appliquer à

(11) Pour une présentation de ces projets, voir, dans l'ordre, Dominique STUTZMANN, Jean-François MOUFFLET & Sébastien HAMEL, « La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet *HIMANIS* pour l'édition électronique », dans *Médiévales*, t. 73, 2017, p. 67-96 (DOI : 10.1215/00161071-7205281) ; Julie CLAUSTRE & Darwin SMITH, « *e-NDP* Notre-Dame de Paris et son cloître », dans *Revue Mabillon*, t. 94, 2022, à paraître ; Xavier ROUSSEAUX & Eddy PUT, éds, *Pardons. Topographies of Pardon Tales: Contextual Mapping of Pardon Letters in the Southern Low Countries, 15th-17th c.*, depuis 2022 [en ligne]. URL : https://pardons.eu/ ; Alix CHAGUÉ *et al.*, *LECTAUREP. L'intelligence artificielle appliquée aux archives notariales*, depuis 2018 [en ligne]. URL : https://lectaurep.hypotheses.org/

(12) Zhuohao CAI, Yi YANG, Zhiyao ZHOU & Lan LIN, « Automatic Detection of Landscape Painting Elements Based on Machine Learning », dans *2019 Photonics & Electromagnetics Research Symposium. Fall (PIERS – Fall)*, New York, IEEE, 2019, p. 432-439 (DOI : 10.1109/PIERS-Fall48861.2019.9021865).

des corpus plus vastes (pour reprendre l'exemple cité plus haut, on pourrait imaginer d'extraire automatiquement toutes les représentations de châteaux médiévaux peints à l'arrière-plan de tableaux modernes)[13].

Analyse exploratoire

D'autres méthodes, que nous proposons de regrouper dans une seconde catégorie, relèvent de l'analyse exploratoire, au sens où elles permettent au chercheur de mieux comprendre les données qu'il mobilise et les liens qui unissent celles-ci entre elles. En d'autres termes, grâce à ces techniques, l'ordinateur met au jour des phénomènes que le chercheur aurait pu percevoir de lui-même, en travaillant « à la main », mais qui risquaient fort de lui rester inaccessibles pour des questions d'échelle, en raison du caractère massif des données traitées. Les outils d'analyse de réseaux, fort en vogue depuis une quinzaine d'années, relèvent dans une certaine mesure de cette catégorie de méthodes. Grâce à l'ordinateur et à la mobilisation de différents algorithmes, il est à la fois possible de représenter graphiquement les liens qui unissent entre eux les acteurs d'un même réseau et de calculer un certain nombre de métriques sur celui-ci, comme des mesures de centralité ou de densité[14]. La machine révèle ainsi au chercheur des réalités dont il aurait difficilement pu prendre conscience en recourant aux méthodes traditionnelles. Ainsi, en 2017, les philosophes Mark Alfano, Andrew Higgins et Jacob Levernier ont apporté une contribution notable à l'étude du « système de valeurs » américain du début du XXIe siècle, en déterminant quelles qualités étaient mises en exergue de manière concomitante dans plusieurs milliers de notices nécrologiques publiées dans la presse régionale[15]. Cela leur a notamment permis d'établir que le critère du genre jouait un rôle primordial dans la façon dont les parents d'un défunt souhaitaient que l'on se souvienne de leur proche décédé, les femmes étant plus souvent dépeintes en mettant en avant des valeurs domestiques et les hommes en fonction du rôle qu'ils ont pu jouer dans la sphère publique, voire politique.

De la même façon, d'autres outils visent à comparer automatiquement des sources textuelles entre elles, en vue de repérer les similitudes, remplois et citations qu'un auteur fait d'un ou de plusieurs autres – un peu à la manière des logiciels de détection de plagiat utilisés dans le cadre de l'enseignement universitaire. Une telle approche a, par exemple, été adoptée dans le *Tesserae Project*, qui visait à développer un outil web permettant d'identifier les remplois de textes antiques latins ou grecs dans des sources postérieures, qu'elles soient antiques ou médiévales (seuls de « grands noms » tels qu'Avit

(13) Nous ne nous attardons pas sur la question, mais des entreprises de classification automatique en fonction des genres littéraires ont aussi été développées dans le domaine de la philologie. Voir notamment Pierre-Carl Langlais, « Reconstituer les genres romanesque sur *Gallica* : essai de classification automatisée de 1500 romans (1815-1850) », dans *Sciences communes*, 2019, [en ligne]. URL : https://scoms.hypotheses.org/986.

(14) Pour une introduction aux techniques et au vocabulaire de l'analyse de réseaux, voir Alain Degenne & Michel Forsé, *Les réseaux sociaux*, Paris, Armand Colin, 2004 (Collection U).

(15) Mark Alfano, Andrew Higgins & Jacob Levernier, « Identifying Virtues and Values Through Obituary Data-Mining », dans *The Journal of Value Inquiry*, t. 52, 2018, p. 59-79 (DOI : 10.1007/s10790-017-9602-0).

de Vienne, Bède le Vénérable ou Nithard sont cependant envisagés pour le Moyen Âge)[16]. Il n'est pas besoin de longs commentaires pour expliquer tout l'intérêt de telles techniques, qui permettent d'accélérer considérablement le processus de recherche…

Il ne s'agit ici que de quelques exemples de méthodes *data-driven* à vocation exploratoire. Nous aurions pu en citer d'autres, dans des champs disciplinaires différents, comme l'histoire de l'art, où des techniques permettent de détecter automatiquement des points de fuite dans des peintures anciennes[17], la musicologie, via l'extraction automatisée des *leitmotivs* utilisés dans une œuvre musicale et l'identification des contextes (harmoniques, scénaristiques, etc.) dans lesquels ceux-ci apparaissent[18], ou encore la science politique et la communication, à travers l'étude de corpus de tweets[19] ou de la manière dont des représentants se regroupent ou se désolidarisent à l'heure des votes[20].

Analyse interprétative

Une troisième classe de méthodes relève de l'analyse interprétative, dont les frontières avec ce que nous avons appelé « analyse exploratoire » sont relativement perméables. Dans le cas présent, l'ordinateur permet au chercheur de construire son argumentation et de bâtir son discours scientifique, en lui proposant des éléments d'interprétation ou d'explication auxquels il n'aurait pu aboutir sans recourir à la puissance de calcul de la machine. Dans le domaine de la philologie, ces techniques sont particulièrement utilisées dans l'exercice difficile de la critique d'attribution. Comme l'ont démontré, par exemple, les travaux des médiévistes de Gand autour d'Hildegarde de Bingen, de Mike Kestemont sur la poésie en moyen-néerlandais ou de Jean-Baptiste Camps et Florian Cafiero sur les pièces de Molière, il est en effet possible de recourir aux méthodes statistiques pour identifier les propriétés stylistiques caractéristiques d'une œuvre ou d'un auteur[21]. Partant, il devient

(16) Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde & Sarah L. Jacobson, « The *Tesserae Project*: Intertextual Analysis of Latin Poetry », dans *Literary and Linguistic Computing*, t. 28, 2013, 2, p. 221-228 (doi : 10.1093/llc/fqs033).

(17) Gilles Simon, « Jan van Eyck's Perspectival System Elucidated Through Computer Vision », dans *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, t. 4, 2021, 2, p. 1-8 (doi : 10.1145/3465623).

(18) Frank Zalkow, Christof Weiss & Meinard Müller, « Exploring Tonal-Dramatic Relationships in Richard Wagner's Ring Cycle », dans *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR, Suzhou, China*, s.l., 2017, p. 642-648 [en ligne]. url : https://archives.ismir.net/ismir2017/paper/000132.pdf.

(19) Yogev Matalon, Ofir Magdaci, Adam Almozlino & Dan Yamin, « Using Sentiment Analysis to Predict Opinion Inversion in Tweets of Political Communication », dans *Scientific Reports*, t. 11, 2021, 7250 (doi : 10.1038/s41598-021-86510-w).

(20) Bryce J. Dietrich, « Using Motion Detection to Measure Social Polarization in the U.S. House of Representatives », dans *Political Analysis*, t. 29, 2021, p. 250-259 (doi : 10.1017/pan.2020.25).

(21) Voir, dans cet ordre, Mike Kestemont, Sara Moens & Jeroen Deploige, « Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux », dans *Digital Scholarship in the Humanities*, t. 30, 2015, 2, p. 199-224 (doi : 10.1093/llc/fqt063) ; Mike Kestemont, *Het gewicht van de auteur. Stylometrische auteursherkenning in Middelnederlandse literatuur*, Gand, Koninklijke Academie voor Nederlandse Taal- en Letterkunde, 2013 (Studies op het gebied van de

donc possible de comparer sur des bases objectives les styles de deux auteurs ou, mieux encore, de déterminer si la façon dont un texte littéraire anonyme est composé correspond plutôt au style de tel ou tel auteur. Lorsqu'elle est utilisée en complément d'indices plus qualitatifs, la stylométrie offre donc des arguments quantitatifs solides pour attribuer une œuvre anonyme à un auteur déterminé(22).

L'utilisation de ces méthodes poussées se rencontre aussi en archéologie, avec le développement de « modèles à base d'agents », (*agent-based modeling)*. Créés en vue de mieux comprendre des systèmes complexes, ces techniques simulent des interactions d'entités (des personnes, des groupes, etc.) qui évoluent en fonction de critères objectifs ou de facteurs politiques, économiques ou sociaux. Pour revenir sur une réalité qui remémorera de douloureux souvenirs à tous, il s'agit, par exemple, de déterminer comment l'application de telle mesure de confinement ou l'avènement de tel variant plus contagieux peut avoir un impact sur l'évolution globale d'une épidémie et la saturation des services hospitaliers. Dans le champ de l'archéologie, des chercheurs néerlandais ont récemment tenté de produire des modèles qui visent à simuler les dynamiques antiques de peuplement et de production agricole dans le delta de la Meuse et du Rhin à la fin de l'Antiquité(23).

Exercice de la critique

Enfin, grâce à l'ordinateur, le chercheur peut également exercer une certaine forme de critique des résultats qu'il a obtenus, en faisant appel à des techniques qui permettent de mesurer la « robustesse » de ces derniers, c'est-à-dire leur capacité à ne pas être affectés par des perturbations importantes si les données sont légèrement modifiées. L'un des auteurs du présent article a ainsi entrepris de mesurer la robustesse des mesures de centralité (de degré, de proximité, d'intermédiarité et de vecteur propre) obtenues dans trois analyses de réseaux menées à partir de sources médiévales de natures différentes (actes épiscopaux des XI^e^ et XII^e^ siècles, chirographes du

oudere Nederlandse letterkunde, 5) ; Florian CAFIERO & Jean-Baptiste CAMPS, « Why Molière Most Likely Did Write His Plays », dans *Science Advances*, t. 5, 2019, 11, p. 1-14 (DOI : 10.1126/sciadv.aax5489).

(22) Nous ne nous y arrêtons pas, mais d'autres techniques permettent, un peu de la même manière, de déterminer automatiquement quelles sont les thématiques mises à l'avant dans de vastes corpus textuels. De telles méthodes ont été utilisées, par exemple, pour dépouiller automatiquement des centaines d'articles scientifiques. Voir Jan LUHMANN & Manuel BURGHARDT, « Digital Humanities: A Discipline in its Own Right? An Analysis of the Role and Position of Digital Humanities in the Academic Landscape », dans *Journal of the Association for Information Science and Technology*, t. 73, 2021, 2, p. 1-24 (DOI : 10.1002/asi.24533).

(23) Jamie JOYCE, « Modelling Agricultural Strategies in the Dutch Roman Limes via Agent-Based Modelling (ROMFARMS) », dans Philip VERHAGEN, Jamie JOYCE & Mark R. GROENHUIJZEN, éds, *Finding the Limits of the Limes. Modelling Demography, Economy and Transport on the Edge of the Roman Empire*, New York, Springer, 2019 (Computational Social Sciences), p. 109-127 (DOI : 10.1007/978-3-030-04576-0_7). Pour d'autres techniques d'analyse interprétative utilisées en archéologie du peuplement, voir, par exemple, Wesley BERNARDINI & Matthew A. PEEPLES, « Sight Communities. The Social Significance of Shared Visual Landmarks », dans *American Antiquity*, t. 80, 2015, 2, p. 215-235 (DOI : 10.7183/0002-7316.80.2.215).

XIII^e siècle, manuscrits hagiographiques)[24]. Pour ce faire, il a simulé des processus de dégradation ou de disparition de données similaires à ceux que l'historien rencontre quotidiennement. Il a ensuite mesuré la manière dont ces transformations des données avaient un impact sur les résultats obtenus, avec l'intention de déterminer, par exemple, si la perte de tel ou tel pan des données aurait pu avoir des conséquences importantes sur les mesures de centralité calculées par l'outil d'analyse de réseau. Au terme de cette enquête, il faut en conclure à la relative robustesse de ces mesures de centralité, et donc à leur fiabilité, dans le champ des études historiques. De la même manière, d'autres chercheurs travaillant sur des sources littéraires du XVIII^e siècle ont tenté de déterminer l'impact des erreurs d'OCRisation sur les résultats que l'on obtient en appliquant à ces écrits des techniques numériques classiques d'analyse textuelle (stylométrie, collocation, etc.)[25].

La question des données en sciences humaines

Ce rapide tour d'horizon permet de se convaincre de l'intérêt des méthodes *data-driven* dans le champ des sciences humaines et sociales. Leur application ne va toutefois pas de soi, loin s'en faut. En un sens, il s'agit de la confrontation de deux sphères a priori bien distinctes : les sciences humaines, d'une part, des techniques quantitatives venues des sciences exactes – et souvent développées pour elles –, de l'autre. Cette confrontation soulève par conséquent un certain nombre de questions épistémologiques, que d'autres spécialistes du numérique ont déjà pointées avant nous :

> Some of the key techniques that are deployed in the digital humanities […] importantly raise questions for epistemology in relation to knowledge production and analysis[26].
>
> Few things will cripple the humanities more than the uncritical "adoption of tools" or the continued encroachment of positivistic research methods borrowed from cognitive science, neuroscience, computer science, or elsewhere[27].
>
> In fact, there is very little research on the epistemological foundations of digital humanities and on how they differ from the "traditional" humanities; in other words, research that tries to answer the question: what is the impact of computational methods on the production of knowledge in the humanities[28] ?

(24) Sébastien DE VALERIOLA, « Can Historians Trust Centrality? Historical Network Analysis and Centrality Metrics Robustness », dans *Journal of Historical Network Research*, t. 6, 2021, 1, p. 85-125 (DOI : 10.25517/jhnr.v6i1.105).

(25) Mark J. HILL & Simon HENGCHEN, « Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study », dans *Digital Scholarship in the Humanities*, t. 34, 2019, 4 (DOI : 10.1093/llc/fqz024).

(26) David M. BERRY & Anders FAGERJORD, *Digital Humanities*, Cambridge, Polity, 2017, p. 105.

(27) Alexander R. GALLOWAY, « The Cybernetic Hypothesis », dans *Differences*, t. 25, 2014, 1, p. 107-131, ici p. 128 (DOI : 10.1215/10407391-2420021).

(28) Michael PIOTROWSKI, « Epistemological Issues in Digital Humanities », communication inédite donnée le 25 avril 2022 à Berlin, ici p. 2 [en ligne]. URL : https://

Si l'existence de tensions entre les questionnements des sciences humaines et les méthodes des sciences exactes a déjà été soulignée à maintes reprises au sein de la littérature scientifique, on rencontre encore assez peu de débats autour des conséquences pratiques et concrètes de cette confrontation, comme si les réflexions s'étaient bornées à constater la difficulté à les faire coexister. Or, bien des aspects de cette rencontre entre des disciplines très différentes doivent encore faire l'objet de discussions. Ce sera précisément l'un des objectifs du groupe de contact « Les humanités des données » que de réfléchir aux multiples dimensions de cette confrontation, en réunissant des spécialistes de différentes disciplines des sciences humaines.

Une question des plus basiques, mais des plus importantes, concerne la pertinence de l'utilisation du terme « données » en sciences humaines et sociales. Quel sens convient-il de lui donner, alors que les chercheurs actifs dans ces disciplines lui préfèrent souvent des mots comme « sources », « documents » ou « corpus » lorsqu'il s'agit de désigner le matériau sur lequel ils travaillent ? Les résultats d'une enquête menée en 2017 à l'Université libre de Bruxelles par le groupe de travail « Gestion des données de la recherche » illustrent bien cet état de fait. Le sondage réalisé par cette équipe visait à « recueillir des informations sur les pratiques des chercheurs de l'ULB dans la gestion de leurs données de recherche »[29]. Seuls 3.5 % des chercheurs de la Faculté de Philosophie et Sciences sociales et 1.5 % des ceux de la Faculté de Lettres, Traduction et Communication ont répondu à cet appel, contre 39 % en Faculté des Sciences, 21.5 % en Faculté des Sciences appliquées et 13 % en Faculté de Médecine. Les taux très bas des facultés rassemblant des chercheurs en sciences humaines sont en partie imputables à un problème de vocabulaire : une part non négligeable de ces chercheurs ont considéré que la question ne les concernait pas, parce qu'ils estimaient ne pas manipuler de données dans leurs activités de recherche… Dès lors, il importe de se demander pourquoi l'utilisation de ce terme dans le contexte des sciences humaines pose tant problème, alors que son emploi paraît peu remis en question dans d'autres champs disciplinaires.

Pour mieux le comprendre, un détour par les travaux de chercheurs qui se sont posé la question s'avère assez éclairant. La plupart des définitions proposées en dehors des sciences humaines prêtent aux « données » une dimension factuelle et objective, comme si le concept avait un caractère « naturel[30] ». C'est aussi ce que suggère l'étymologie commune du mot « donnée » et de son équivalent anglais *data*, qui renvoie à l'idée d'une réalité offerte ou octroyée à titre gratuit. Cette posture convient plutôt mal aux sciences humaines, dont les méthodes consistent notamment en une prise de distance critique vis-à-vis de l'objet étudié, quel qu'il soit, ce qui implique de considérer le matériau sur lequel travaille le chercheur comme une construction ou comme un artefact au moins en partie élaboré par celui-

zenodo.org/record/6498979#.Y61hSR3jJTY

(29) GROUPE DE TRAVAIL « GESTION DES DONNÉES DE RECHERCHE » (UNIVERSITÉ LIBRE DE BRUXELLES), *Résultats de l'enquête sur les usages des chercheurs de l'ULB en matière de gestion des données de recherche*, Bruxelles, 2017 [en ligne]. URL : https://gdr.ulb.ac.be/stat_survey.html

(30) Jennifer ROWLEY, « The Wisdom Hierarchy: Representations of the DIKW Hierarchy », dans *Journal of Information Science*, t. 33, 2007, 2, p. 163-180 (DOI : 10.1177/0165551506070706).

ci. En conséquence, il importerait de toujours garder à l'esprit les processus de transformation, d'altération ou encore de perte par lesquels ce matériau serait passé, de sa création jusqu'à son exploitation

Dans le même sens, d'aucuns ont parfois souligné que les contextes dans lesquels apparaissent les données en sciences humaines sont sensiblement différents de ceux dans lesquels on rencontre celles-ci dans d'autres champs disciplinaires. Que le chercheur travaille sur des manuscrits, des dessins, des œuvres musicales ou sur des artefacts issus de fouilles, les éléments exploités dans le cadre de l'enquête ne viennent généralement pas seuls, mais apparaissent dans un contexte qui peut être extrêmement complexe, en raison de son origine humaine. Le processus par lequel ces éléments sont extraits de leur contexte d'origine est donc un aspect central du travail du chercheur, ce à quoi l'utilisation du terme « donnée » ne rendrait pas justice. Dès lors, le matériau du chercheur en sciences humaines compterait plus de dimensions que celui des scientifiques actifs dans d'autres disciplines. Cette apparente inadéquation a poussé plusieurs auteurs, comme Howard Jensen en 1950[31] ou Johanna Drucker plus récemment[32], à proposer d'abandonner le terme « données » au profit du mot « captées » (en anglais *capta*), afin de mieux mettre en évidence que le matériau du chercheur en sciences humaines a été « arraché » au contexte dans lequel il apparaît. Par ailleurs, la question de la « qualité » des matériaux employés dans le cadre de la recherche pourrait être invoquée pour justifier l'emploi de termes différents, les chercheurs en sciences humaines et sociales étant bien souvent contraints de travailler avec des matériaux incomplets, imprécis, incertains ou dont l'état de complétude est difficile, voire impossible, à évaluer.

Si ces objections ne sont pas sans pertinence, il semble nécessaire de les confronter avec des définitions de la « donnée » formulées dans des contextes plus proches des sciences humaines et sociales. Donnons-en deux exemples, l'un en histoire, l'autre dans le champ des humanités numériques. Le médiéviste Jean-Philippe Genet a proposé, en 1986, de parler de « données » pour désigner le résultat d'un processus d'extraction des informations disponibles dans un corpus de sources (le « réel historique »). Il insiste sur la formalisation des définitions et des hypothèses de travail faites au cours de cette démarche (la « théorie initiale locale »). Toutes les données extraites selon ce processus (la « méta-source ») constituent pour lui un système fermé pouvant être soumis à l'ordinateur[33].

Pour sa part, le *digital humanist* Trevor Owens voit dans la donnée « un objet multiforme qui peut être mobilisé comme preuve à l'appui d'un argument ». Il définit trois manières d'aborder le concept dans le cadre des sciences humaines. Premièrement, traiter la donnée comme un artefact, c'est-à-dire un objet qui a été créé de façon active et délibérée par un être humain. Deuxièmement, la considérer de la même manière qu'un texte, en conservant à l'esprit que

(31) Cité dans Howard BECKER, « Science, Culture, and Society », dans *Philosophy of Science*, t. 19, 1952, 4, p. 273-287 (DOI : 10.1086/287212).

(32) Johanna DRUCKER, « Humanities Approaches to Graphical Display », dans *Digital Humanities Quarterly*, t. 5, 2011, 1 [en ligne]. URL : http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

(33) Jean-Philippe GENET, « Histoire, Informatique, Mesure », dans *Histoire & Mesure*, t. 1, 1986, 1, p. 7-18 (DOI : 10.3406/hism.1986.904).

la donnée est sujette à interprétation, en particulier par le chercheur qui l'a créée. Troisièmement, l'envisager comme une information susceptible d'être traitée par l'ordinateur, et qui peut dont être analysée à travers le recours aux méthodes quantitatives[34].

De telles définitions « alternatives » pourraient être multipliées. Les propositions de Genet et d'Owens ont néanmoins retenu notre intérêt car elles tiennent pleinement compte de la situation particulière des chercheurs en sciences humaines et sociales, des spécificités des matériaux qu'ils exploitent et des méthodes qu'ils mettent en œuvre. Dans leurs définitions, les données conservent un statut de matériau destiné à être soumis à des traitements quantitatifs, sans pour autant que l'on ne fasse fi des processus de création et d'extraction aboutissant à leur constitution. D'une part, leurs réflexions démontrent que parler de « données » en sciences humaines n'a rien d'incongru, puisque l'approche développée par le chercheur ne présentera pas nécessairement un caractère positiviste. D'autre part, leurs suggestions permettent de mettre en relief certaines étapes clés de tout processus de recherche en sciences humaines impliquant le recours aux méthodes quantitatives : ce sont des étapes de modélisation, au cours desquelles le chercheur transforme ses sources en les réduisant, en les simplifiant, en les interprétant, en les catégorisant, etc. En suivant Genet et Owens, on répond donc à la plupart des griefs formulés contre l'emploi du terme « donnée » en sciences humaines. C'est, d'après nous, l'élément essentiel, car il est vain de chercher à s'accorder autour d'une définition consensuelle, valable pour toutes les disciplines en même temps.

Les étapes de modélisation, sur lesquelles insistent Genet et Owens, font partie intégrante du processus de recherche. Elles ne peuvent d'ailleurs être accomplies que par un chercheur connaissant les matériaux exploités et maîtrisant la problématique envisagée. Sans cela, la construction du modèle risque de s'avérer erratique tandis que les conclusions ont de grandes chances de prendre un caractère fallacieux. Il n'est pas rare, pourtant, que des techniciens sans aucun bagage en sciences humaines s'attaquent, seuls, à des problématiques pour lesquelles une réelle expertise est pourtant requise… Trop souvent, leurs résultats à première vue chatoyants et porteurs auprès des médias ne résistent pas à une critique raisonnée de spécialistes de la problématique étudiée. Tel est le cas, par exemple, d'un article paru en 2020 dans la prestigieuse revue *Nature Communications*, et ayant depuis fait l'objet de plusieurs amendements de la part de ses auteurs en raison de faiblesses méthodologiques avérées[35]. Dans cette enquête, les auteurs appliquent une méthode d'apprentissage automatique qui permet de détecter automatiquement « des caractéristiques faciales spécifiques, telles qu'une bouche souriante ou des yeux plus grands » dans des portraits peints entre 1500 et 2000. Ils mettent ensuite en lien ces caractéristiques faciales avec des indices économiques comme le PIB (éventuellement reconstruit) des pays dont proviennent les sujets

(34) Trevor OWENS, « Defining Data for Humanists. Text, Artifact, Information or Evidence? », dans *Journal of Digital Humanities*, t. 1, 2011, 1 [en ligne]. URL : http://journalofdigitalhumanities.org/1-1/defining-data-for-humanists-by-trevor-owens/.

(35) Lou SAFRA, Coralie CHEVALLIER, Julie GRÈZES & Nicolas BAUMARD, « Tracking Historical Changes in Trustworthiness Using Machine Learning Analyses of Facial Cues in Paintings », dans *Nature Communications*, t. 11, 2020, 1 (DOI : 10.1038/s41467-022-31843-x).

des peintures. Ils en concluent que leurs résultats témoignent de l'existence de relations entre la richesse économique et la « confiance sociale » affichée par les individus figurés dans les portraits – en termes plus cavaliers, plus on est riche, plus on aurait tendance à sourire. L'équipe n'inclut évidemment aucun historien de l'art… Selon nous, la mise en œuvre de tels projets requiert pourtant que l'alchimie se fasse entre trois éléments : a) La manipulation des données et la mise en œuvre des méthodes demande des connaissances en informatique, sans pour autant exiger du chercheur des talents d'expert en programmation ; b) L'application des méthodes elles-mêmes, depuis les hypothèses sur lesquelles elles reposent jusqu'à l'interprétation de leurs résultats, nécessitent une certaine compréhension de leur fonctionnement mathématique et algorithmique ; c) Le troisième ensemble de compétences nécessaires est celui que nous venons d'évoquer : l'expertise « qualitative » des données et du contexte dans lequel elles s'insèrent avec la problématique étudiée.

Épilogue : d'autres enjeux pour demain

Bien entendu, la question de l'emploi du terme « donnée » n'est pas la seule question qui naît de la confrontation entre les méthodes quantitatives et les sciences humaines. Leur rencontre soulève une série d'autres enjeux, sur lesquels nous projetons de revenir lors de futures réunions. En guise d'épilogue, mentionnons six d'entre elles, sans entrer dans le détail.

1) L'une pourrait d'abord porter sur la difficulté d'atteindre une « vérité absolue » (*ground truth*) en sciences humaines – c'est-à-dire la possibilité de baliser clairement et objectivement des données, sans aucune place laissée à l'ambiguïté ou à l'interprétation –, alors que celle-ci est indispensable pour entrainer efficacement certains modèles (en particulier des modèles d'apprentissage supervisé).

2) Cette problématique rejoindrait, d'une certaine manière, celle de l'exigence du formalisme, dans la mesure où l'ordinateur peut uniquement traiter et produire des informations complètement formalisées, c'est-à-dire définies sans aucune ambiguïté. Or, les chercheurs en sciences humaines manipulent des concepts qui restent souvent volontairement flous et imprécis, et donc pas du tout formalisés.

3) De même, le recours aux méthodes *data-driven* permet de réinterroger la notion de « preuve ». Ce concept épistémologique particulièrement complexe prend des formes très différentes d'une discipline à l'autre. Prouver un fait en mathématiques, en chimie, en économie, en archéologie ou en études littéraires n'implique pas nécessairement les mêmes démarches. Il est dès lors nécessaire de s'interroger sur la manière dont la rencontre entre le discours scientifique des sciences humaines et les méthodes des sciences exactes pèse sur ce processus, dans la mesure où le recours aux techniques quantitatives suppose de suivre un certain nombre de règles précises pour établir des faits. La sous-question de la réplicabilité pourrait à cet égard constituer un point de friction. Est-il en effet pertinent de soumettre les sciences humaines aux mêmes exigences que les sciences exactes en termes de réplicabilité des méthodes et des résultats ?

4) Par ailleurs, va-t-il de soi que les méthodes quantitatives importées des sciences exactes soient appliquées telles quelles et sans adaptation – dans une logique *one size fits all* – aux données de sciences humaines, dont la nature est profondément différente ?

5) Dans la même veine, on peut également se demander si l'ensemble des méthodes quantitatives développées en sciences exactes sont transposables aux sciences humaines, au sens où, dans ce dernier secteur, il est souvent impossible d'atteindre le seuil critique nécessaire à une bonne calibration de ces méthodes qui ont été conçues à l'origine pour manipuler de très vastes ensembles de données – le fameux *big data*, tant à la mode dans le vocabulaire managérial du dépôt de projet.

6) Enfin, d'un point de vue plus pratique, se pose aussi la question de la gestion à long terme des données de la recherche. Les méthodes évoquées plus haut reposant sur la manipulation de quantités importantes de données, il est sans doute nécessaire que les chercheurs en sciences humaines et sociales s'inspirent des bonnes habitudes de leurs confrères de sciences exactes en matière de gestion, de partage et de conservation des données. Mais nos communautés sont-elles prêtes à franchir le pas ? Cela semble encore loin d'être acquis, hélas, puisque la pratique de déposer ses données de recherche dans des répertoires numériques prévus à cet effet reste marginale, tandis que le lourd travail de préparation des données n'est généralement ni reconnu ni valorisé scientifiquement, en dépit de son caractère fondamental.

RÉSUMÉ

Sébastien DE VALERIOLA, Paul BERTRAND & Nicolas RUFFINI-RONZANI, *Données, sciences humaines et méthodes quantitatives : typologie, programme et pistes de réflexion*

Cet article constitue une version remaniée de la communication présentée à l'occasion de la rencontre inaugurale du groupe de contact FNRS « Les humanités des données », tenue à l'Université libre de Bruxelles, le 7 novembre 2022. Nous y retraçons d'abord l'évolution des relations entre les sciences humaines et les méthodes quantitatives, en distinguant trois grandes phases d'évolution successives. Nous dressons ensuite une typologie de ces méthodes d'après l'étape du processus de recherche dans laquelle elles interviennent. Après avoir observé, avec la littérature, que leur application à des problématiques issues des sciences humaines pose question sur le plan épistémologique, nous nous intéressons à un problème plus précis, celui de la pertinence de l'utilisation du terme « données » en sciences humaines. L'article se clôt sur l'évocation d'autres questions de ce type qui nous semblent d'intérêt et qui pourraient être discutées lors des prochaines réunions du groupe de contact.

Méthodes quantitatives – humanités numériques – sciences humaines

SUMMARY

Sébastien DE VALERIOLA, Paul BERTRAND & Nicolas RUFFINI-RONZANI, *Data, Humanities, and Quantitative Methods: Typology, Agenda, and Food for Thought*

This article is a revised version of the paper presented at the inaugural meeting of the FNRS contact group "Data-driven Humanities", held at the Université libre de Bruxelles on November 7, 2022. We first trace the evolution of the relationship between the humanities and quantitative methods, distinguishing three major successive phases of evolution. We then draw up a typology of these methods according to the stage of the research process in which they are used. After having observed, with the literature, that their application to problems in the human sciences raises epistemological questions, we turn to a more specific problem, that of the relevance of the use of the term "data" in the humanities. The article closes by mentioning other questions of this type that we think are of interest and that could be discussed at future meetings of the contact group.

Quantitative methods – Digital humanities – Humanities

SAMENVATTING

Sébastien DE VALERIOLA, Paul BERTRAND & Nicolas RUFFINI-RONZANI, *Gegevens, menswetenschappen en kwantitatieve methoden: typologie, agenda en denkpistes*

Dit artikel is een herziene versie van de paper die werd voorgesteld op de openingsvergadering van de FNRS-contactgroep "Data-driven humanities", gehouden aan de Université libre de Bruxelles op 7 november 2022. Wij schetsen eerst de

evolutie van de relatie tussen de menswetenschappen en de kwantitatieve methoden, waarbij wij drie belangrijke opeenvolgende fasen van evolutie onderscheiden. Vervolgens stellen wij een typologie van deze methoden op volgens de fase van het onderzoeksproces waarin zij worden gebruikt. Nadat wij samen met de literatuur hebben vastgesteld dat de toepassing ervan op problemen in de menswetenschappen vragen oproept op epistemologisch niveau, gaan wij over tot een meer specifiek probleem, namelijk dat van de relevantie van het gebruik van de term "gegevens" in de menswetenschappen. Het artikel sluit af met een bespreking van andere dergelijke kwesties die volgens ons van belang zijn en die tijdens toekomstige vergaderingen van de contactgroep kunnen worden besproken.

Kwantitatieve methoden – digitale menswetenschappen – geesteswetenschappen

RÉDACTION: 4, boulevard de l'Empereur, 1000 Bruxelles.
Prière d'adresser à la Rédaction les *manuscrits* et les *ouvrages pour compte rendu.*

REDACTIE: 4, Keizerslaan, 1000 Brussel.
Gelieve *teksten* en *boeken ter recensie* aan de Redactie te zenden.

Prière d'adresser les demandes d'abonnements, les commandes diverses, etc.,
Revue Belge de Philologie et d'Histoire
KBR - Bibliothèque Royale
4, boulevard de l'Empereur, B-1000 Bruxelles.
rbph@belgacom.net
Tous les paiements doivent être faits au compte bancaire 000-0131507-72
(IBAN BE38 0000 1315 0772 - BIC BPOTBEB1)
de la Revue Belge de Philologie et d'Histoire, B-1050 Bruxelles.

Informations pratiques: http://www.rbph-btfg.be

Chaque article est signé. L'auteur est responsable des idées qu'il émet. La *Revue* n'accepte qu'une seule réplique à un article ou à un compte rendu. L'auteur de celui-ci aura la faculté de la faire suivre de ses observations. Après quoi, le débat sera tenu pour clos.

Voor abonnementen en andere bestellingen, zich wenden tot
Belgisch Tijdschrift voor Filologie en Geschiedenis
KBR - Koninklijke Bibliotheek
4, Keizerslaan, B-1000 Brussel.
rbph@belgacom.net
Alle betalingen dienen te gebeuren op bankrekeningnummer 000-0131507-72
(IBAN BE38 0000 1315 0772 - BIC BPOTBEB1)
van het Belgisch Tijdschrift voor Filologie en Geschiedenis, B-1050 Brussel.

Praktische informatie: http://www.rbph-btfg.be

Elke bijdrage vermeldt de naam van de auteur. Deze alleen is voor de in zijn studie uiteengezette opvattingen en verdedigde zienswijzen verantwoordelijk. Het *Tijdschrift* aanvaardt slechts één repliek op een artikel of een recensie. De schrijver ervan mag op de ingezonden repliek antwoorden. Van verdere polemiek wordt beslist afgezien.

Imprimerie GROENINGHE Drukkerij, Kortrijk

ISSN 0035-0818

Éditeur responsable et directeur de la publication

Michèle Galand, 106, rue de Rosières, 1332 Genval

HISTOIRE – GESCHIEDENIS

Articles - Artikelen

Varia

Bibliographie - Bibliografie

ISSN : 0035-0818