Intelligibilité du numérique

N° 7 | 2025
Le 4ème paradigme en question

La fabrique de la liquidité des données

Plongements dans un milieu navigable et herméneutique(s) de l'espace latent

Simon DUMAS PRIMBAULT *Professeur junior*
InSHS
OpenEdition Lab
French National Centre for Scientific Research
Paul GUHENNEC *Docteur*
DHLAB
École Polytechnique Fédérale de Lausanne

Édition électronique :
URL :
https://intelligibilite-numerique.numerev.com/articles/revue-7/3211-la-fabrique-de-la-liquidite-des-donnees
Date de publication : 04/05/2025

CertiScience© *Certifié évalué par les pairs*

Cette publication est sous licence **CC BY-NC-ND** (Attribution - No commercial - No derivatives).

Pour **citer cette publication** : DUMAS PRIMBAULT, S., GUHENNEC, P. (2025) La fabrique de la liquidité des données. *Intelligibilité du numérique*, (7). https://doi.org/10.34745/

Selon les discours d’escorte du quatrième paradigme, l’une des conditions de sa réalisation serait la fluidité des données – « data liquidity » – sur lesquelles il se fonde. La métaphore du « lac de données », par exemple, – utilisée à partir de 2011 pour décrire un format de stockage de données de natures et formats très divers, structurées ou non, souvent sans savoir quel en sera l’usage – s’est par la suite émancipée de son contexte technique pour incarner l’idéal hydraulique de la recombinaison perpétuelle et sans friction. En somme, c’est la fluidité de masses informes de données supposées non structurées qui permettrait le « désilotage » des pratiques et la production de savoirs nouveaux. Avec elle s’ouvriraient les possibilités d’analyse nouvelles du quatrième paradigme : le changement de nature des résultats dans le traitement de grandes quantités de données, la mise en équivalence de matériaux empiriques de natures hétérogènes dans un espace commun, l’automatisation de ces procédés dans leur incarnation par des machines auto-apprenantes. Pourtant, rien de spontané ou de naturel dans la liquidité des données qui, paradoxalement, nécessitent d’être structurées afin de garantir une navigation « seamless » et sans friction. Cette contribution se propose de mettre au jour la fabrique de la liquidité des données en se penchant sur la chaîne opératoire de l’algorithme de plongement word2vec en humanités numériques.

The Making of Data Liquidity: Embeddings in a navigable milieu and hermeneutics of the latent space

According to the discourses accompanying the fourth paradigm, one of the conditions for its realisation would be the fluidity of the data—“data liquidity”—on which it is based. The metaphor of the “data lake”, for example—used from 2011 onwards to describe a storage format for data of very diverse natures and formats, structured or unstructured, often without knowing what use they will be put to—has subsequently emancipated itself from its technical context to embody the hydraulic ideal of perpetual, frictionless recombination. In short, it is the fluidity of supposedly unstructured masses of data that would enable practices to be “un-siloed” and new knowledge to be produced. This would open up the new analytical possibilities of the fourth paradigm: the

changing nature of results when processing large quantities of data, the equivalence of empirical materials of heterogeneous natures in a common space, the automation of these processes in their embodiment by self-learning machines. Yet there is nothing spontaneous or natural about the liquidity of data, which paradoxically needs to be structured to guarantee seamless, frictionless navigation. This contribution aims to shed light on the making of data liquidity by examining the operational chain of the word2vec embedding algorithm in digital humanities.

Read the English version

Mots-clés :
Humanités numériques, Données, Herméneutique, Digital Humanities, Liquide, Liquid, Hermeneutic, Data, Plongement, Embedding

---

1. Introduction[1]

Un déluge – c'est sous les atours d'un cataclysme apocalyptique que se présenterait l'avènement du "quatrième paradigme" dans les sciences. La fameuse "Révolution scientifique" de l'époque moderne, âprement débattue, fait pâle figure face à la prophétie d'un cataclysme hydraulique d'ampleur biblique qui engloutirait les paradigmes précédents, rendrait "obsolète la méthode scientifique" (Anderson, 2008), ferait table rase de l'histoire des sciences. L'irruption de "la donnée" en masses liquides informes représenterait donc un péril pour les sciences mais un péril gros d'une opportunité – une promesse qui justifierait *a priori* la destruction (créatrice) de l'édifice scientifique sédimenté au fil des siècles.

Fondé dans une épistémologie réductionniste qui définit *la* science comme "calcul à forte intensité de données [*data-intensive computing*]" (Hey *et al.*, 2009, p. xi), le quatrième paradigme prétend donc inaugurer la primauté des données dans la production de connaissances – et ancrer solidement dans cette primauté, grâce à des méthodes d'analyse réputées neutres, un inductivisme parfait. Renouvelant ainsi, paradoxalement, les promesses d'une prétendue Révolution scientifique arc-boutée sur quatre principes majeurs : un empirisme radical, un inductivisme radical, des données exhaustives, la neutralité du dispositif d'analyse. Ainsi produit, le savoir lui-même

“devient ‘liquide’ et son flux [*its flow*], de la recherche à la pratique, se fait sans friction [*frictionless*] et de manière immédiate” (Hey *et al.*, 2009, p. 61). Des données à la connaissance, la métaphore liquide infuse tout le “*workflow*” scientifique et ce “flux” se fait garant d’un savoir immédiat, c’est-à-dire à la fois “instantané” (Hey *et al.*, 2009, p. 59) et non médié, donc objectif, en adhérence avec le réel.

Cependant, afin de canaliser ce déluge de données et le rendre productif, les tenants du quatrième paradigme insistent sur la nécessité de développer et déployer à grande échelle une infrastructure pour la “capture”, la sélection [*curation*], l’analyse et la visualisation des données (Hey *et al.*, 2009, p. 17). Le chaos diluvien ne se ferait donc destruction créatrice qu’au prix d’un processus d’infrastructuralisation comme tentative paradoxale d’organiser l’instabilité[2]. Condition de possibilité du quatrième paradigme, ce que doit soutenir l’infrastructure c’est non seulement l’influx de données – *via*, par exemple, des *pipelines* – mais encore le flux des sujets de la connaissance au sein d’un véritable milieu “navigable et interopérable sans couture [*seamlessly navigable and interoperable*]” (Hey *et al.*, 2009, p. 185). Comme l’on surfe sur le Web, la “fin de la théorie” (Anderson, 2008) serait donc l’avènement d’un nouvel ordre du savoir, un ordre liquide qui ne connaîtrait ni cloison ni silo, un milieu parfaitement continu et continûment navigable où (se) croiseraient, im-médiatement et sans friction, flux de données et sujets de la connaissance[3].

La fabrique (asymptotique) d’un continu liquide ne se fait cependant pas par simple dissolution passive – *i.e.* en ne faisant que recueillir et laisser baigner des masses informes non structurées – mais résulte bien d’opérations diverses de “mise en solution” active grâce à des procédés techniques paradoxalement structurants, cristallisants[4]. C’est cela que nous proposons de démontrer par l’analyse sémio-technique d’un procédé phare du *machine learning*[5] : l’*embedding*, c’est-à-dire la génération d’espaces continus de représentation des données par plongement mathématique. En effet, le plongement, en générant un espace latent unique et partagé, place chaque *datum* dans un voisinage, un *continuum* dont la nature et la portée épistémique sont à interroger. Cette remédiation possible pour toute donnée d’entrée est devenue, avec l’avènement des modèles d’apprentissage profond autour de 2012, l’opération canonique de la fabrique du continu dans le cadre du quatrième paradigme.

Du côté de l’analyse des métaphores, le champ sémantique hydraulique a souvent été inventorié, parfois commenté, que ce soit pour désigner le numérique en général (Bernardot, 2018) ou les données en particulier (Puschmann et Burgess, 2014 ; Stark et Hoffmann, 2019 ; Nolin, 2020). Ici pétrole ([*oil*]), là eau, là encore liquide indistinct formant lacs[6], océans et déluges, les données dans leur *liquidité* métaphorique ont surtout fait l’objet de commentaires sur leur portée économique comme “ressources”. Cependant, le lien entre ces métaphores et l’ordre du savoir qui en découlerait sont peu discutés, en particulier dans une articulation avec les dispositifs sociotechniques qui seraient les garants de cette liquidité. En effet, du côté de l’archéologie des médias et des pratiques, si de nombreux procédés d’ordonnancement des données – en particulier

leur vectorisation – ont fait l'objet de commentaires tout à la fois techniques et philosophiques (Mackenzie, 2017 ; Offert, 2020 ; Dobson, 2022), ceux-ci ont surtout insisté sur l'opération analytique amont de transformation d'un corpus en sac de mots (ex. Gavin, 2020), ainsi que sur l'opération aval de leur visualisation et interprétation, plus rarement sur l'opération synthétique de génération d'un espace vectoriel continu permettant la navigation.

Au croisement de ces deux corpus, la présente contribution propose une étude du lien entre la rhétorique liquide des données et la création d'espaces continus de manipulation de celles-ci et de navigation des sujets de la connaissance. Cette problématique fondamentalement épistémologique de l'analyse de données textuelles au moyen d'algorithmes de traitement automatique des langues naturelles a pris la forme d'un débat sur la place réservée à l'herméneutique dans les sciences humaines et sociales (SHS) numériquement équipées – notamment les humanités numériques. Cependant, plutôt qu'une exception, les SHS sont à envisager comme un cas limite qui met en lumière les tensions essentielles à toute analyse de données. Ainsi, nous espérons ajouter une pierre au développement d'une "herméneutique (du) numérique" (Fickers *et al.*, 2022) qui permette la critique du quatrième paradigme jusque dans les sciences naturelles et l'informatique[7].

La présente contribution se propose donc de questionner la métaphore liquide aux racines du quatrième paradigme en se penchant sur la fabrique de la liquidité des données – en particulier de l'algorithme *word2vec* permettant de mobiliser, traiter et analyser du texte – *medium* et matériau originel et canonique du quatrième paradigme et du savoir depuis la révolution scientifique. En combinant analyse média-théorique (Geoghegan, 2023), philosophie des techniques (Bachimont, 2021) et archéologie des pratiques de données (McKenzie, 2017), nous proposons d'interroger ce passage du discret statique de la donnée au continu dynamique de la fluidité en détaillant la "chaîne opératoire" de l'opération de plongement numérique de *word2vec*. Dans un premier temps, cette contribution proposera une lecture proche du texte et de l'algorithme séminal de *word2vec* (Mikolov *et al.*, 2013), mettant en lumière le contexte historique et scientifique de production de ce dispositif de fabrique du continu, ainsi que les opérations matérielles d'ordre épistémologique que la méthode fait subir à ses données d'entrée afin de générer un espace continu de navigation. Nous nous attarderons ensuite sur les modalités de navigation que le sujet de la connaissance, médié par des algorithmes et outils de visualisation, peut déployer au sein de cet espace afin de produire du savoir – ainsi que hors de celui-ci, vers d'autres espaces qui forment un ensemble réticulaire. Enfin, à l'aune de ces analyses, nous reviendrons sur la place occupée par la continuité liquide au sein du quatrième paradigme et la conséquente nécessité de l'envisager, plutôt que comme déluge ou révolution, dans une relation de sédimentation avec les paradigmes et les pratiques disciplinairement marqués, qui le précèdent historiquement et lui survivent.

2. Synthétiser un espace continûment navigable

Si l'algorithme de vectorisation de données textuelles *word2vec* n'est pas le premier en son genre (sous-partie 1) et s'il a été depuis surclassé par l'introduction des grands modèles de langage (LLMs), il offre néanmoins un parfait observatoire de la méthode du plongement, *i.e.* de la synthèse d'un espace continu (espace latent) à partir de données discrètes (sac de mots) par leur vectorisation (sous-partie 2)[7]. Largement mobilisé pendant plus d'une décennie par les humanités numériques, il inaugure un espace de représentation de données textuelles qui permet un calcul algébrique vectoriel sur des entités minimales de sens. C'est ce calcul qui permet de re-produire du sens par déplacement, au sein d'un milieu continu, d'un sujet médié par des algorithmes de *clustering*, de réduction de dimensionnalité ou de visualisation (sous-partie 3). Du texte au sac de mot à l'espace latent, nous tentons de mettre au jour la chaîne opératoire de *word2vec* et des changements d'état encourus par les données pour produire un milieu continûment navigable, lui-même inscrit dans un espace réticulaire abritant d'autres représentations (sous-partie 4).

2.1. De la cybernétique aux computational humanities

Cette quête d'un espace mathématique où placer des représentations numériques d'unités de sens (mots, documents, images) n'est pas nouvelle et ponctue les premières décennies de la cybernétique dans tous ses domaines d'application. L'objet "espace vectoriel" a commencé à être mobilisé pour représenter du texte dès les années 1970 notamment par Gerard Salton (Dubin, 2004). Si cet espace répond à la problématique restreinte de la recherche d'information typique des bases de données (*information retrieval*), Salton en théorise plusieurs caractéristiques – comme la similarité cosinus ou l'encodage en vecteurs binaires 1 parmi *n* (*one-hot encoding*) – qui seront repris tels quels dans les premiers travaux d'humanités computationnelles à la fin des années 1980. Dans l'espace vectoriel de Salton, la vectorisation se fait *a priori* et nécessite des règles explicites d'encodage. Une décennie plus tôt, au milieu des années 1960, des chercheurs des Bell Labs, conscients de cette limite à venir, avaient d'ailleurs développé la méthode dite de "positionnement multidimensionnel non-métrique [*non-metric multidimensional scaling*]" (NMMDS). Le NMMDS sera rapidement adopté par les psychologues et anthropologues intéressés par la quantification de concepts élusifs, notamment pour sa capacité à transformer un jeu de données en espace où la proximité conceptuelle s'exprime visuellement (Seaver, 2022).

Dans cette même logique de mathématisation des sciences humaines, la linguistique de l'après-guerre se tourne elle aussi vers la statistique comme descripteur. C'est le contexte d'un mot qui le définit et permet de lever toute ambiguïté sémantique[8]. Ainsi, selon les acteurs de cette "sémantique distributionnelle", toute formalisation du langage passe par l'étude statistique de co-occurrences (Harris, 1954). Résumant ses

principes, le linguiste britannique John Firth déclare, dans une formule depuis régulièrement reprise par les acteurs de la vectorialisation : “On reconnaît un mot à ceux qui lui tiennent compagnie [*You shall know a word by the company it keeps*]” (1962, 11).

Dès les années 1990, la sémantique distributionnelle devient, grâce à sa facilité d’implémentation, au centre des premiers travaux de *computational humanities*[9]. Diverses approches co-existent : sacs de mots ([*bag of words*]), matrices terme-document ([*document-term matrix*]) ou allocation de Dirichlet latente ([*Latent Dirichlet Allocation*]) pour la modélisation des sujets ou des thèmes d’un corpus de documents. Ces vecteurs sont à ce stade rarement plus qu’un réservoir de nombres, et si chaque dimension quantifie l’appartenance à un groupement, il manque encore l’équivalence entre position et concept promise par l’espace vectoriel. C’est ce que proposent Bengio *et al.* (2003) : une représentation spatiale qui soit linguistiquement cohérente, par le biais d’un réseau de neurones adaptant ses paramètres selon les co-occurrences de mots au sein d’un corpus.

2.2. word2vec : vectoriser par plongement

Dix ans plus tard, *word2vec* améliore l’idée de ce modèle de langue par “réseau de neurones [*neural network machine learning*]” (NNML) et lui emprunte son ambition d’une équivalence spatio-linguistique (Mikolov *et al.*, 2013). Moins coûteux à entraîner, plus rapide, capable d’un passage à l’échelle promis par plus de données, et offrant une sensibilité contextuelle accrue, *word2vec* s’impose rapidement comme le nouvel étalon et impose à la computation linguistique le chemin du *deep learning*[10]. En particulier, c’est sa capacité à dépasser largement les modèles de langue classiques dans les tests statistiques de comparaison et sa versatilité – comme le montrera quelques mois plus tard une version améliorée intégrant différentes langues – qui marqueront le sacre du plongement vectoriel.

Au delà de son importance technico-historique, si nous prenons *word2vec* comme exemple, c’est aussi qu’il est symptomatique tant de l’enjeu crucial aujourd’hui de la compréhension toujours plus fine par les réseaux de neurones du langage naturel, que de l’inauguration d’espaces d’équivalence entre toutes les données – tous les contextes, tous les *media*, toutes les langues… – par le biais du vecteur mathématique – donc de l’inauguration de la donnée comme valeur (Gingold, 2026). Relevant de ce que Chartier-Edwards *et al.* (2024) ont appelé un “vectorialisme”, il n’est pas anodin que l’équipe de recherche derrière *word2vec* vienne exclusivement de Google. Hier comme aujourd’hui, les outils centraux de la vectorisation culturelle proviennent d’entreprises privées (Meta : LLaMa, Segment Anything ; Google : *word2vec*, Transformers, BERT ; OpenAI : GPT, Dall-E) avant d’être repris par des projets publics (Wark, 2015).

Si les enjeux politiques de cette concentration dépassent notre contribution, notons néanmoins que les intérêts de ces entreprises – toujours à l’affût d’une solution

optimale du point de vue des performances, aisément “scalable”et, donc, générique — sont difficilement compatibles avec une recherche fondamentale faite d’une pluralité de pratiques et de questionnements. Ceci est particulièrement patent lorsque l’on considère, par exemple, que les applications futures de ces nouvelles technologies relèvent dans l’article sur *word2vec* seulement d’un paragraphe pensé *a posteriori* plutôt que de la motivation intrinsèque d’un problème précis à résoudre[11]. Ainsi, la vectorisation du texte apparaît comme *la* solution conceptuelle, améliorable certes, mais prête à être déployée à n’importe quel problème textuel – qu’ils soient aussi différents que la “traduction automatique, la récupération d’information, et les systèmes génériques de réponse aux questions”, aussi bien que “des applications futures qu’il reste à inventer” (Mikolov *et al.*, 2013, p. 5)[12].

Espace vectoriel, équivalence spatio-conceptuelle, sémantique distributionnelle : ces trois concepts infuseront et restent aujourd’hui, après des décennies de mise en application, les trois briques sur lesquelles se fondent les modèles de langues par apprentissage profond. Plus encore, même si la langue en fut le premier terrain de jeu – tant pour des limitations techniques que par l’existence préliminaire d’un cadre théorique – force est de constater que la vectorisation des phénomènes culturels vise depuis l’avènement du *deep learning* de plus en plus de domaines. Comme le rappellent Cardon *et al.* (2018) : “selon la formule de Yann LeCun [un des pionniers de l’apprentissage profond], l’ambition des concepteurs des machines connexionnistes est de mettre le monde dans un vecteur (*world2vec*)” (p. 206).

La chaîne opératoire[13] de *word2vec* reprend un certain nombre d’éléments issus de la généalogie esquissée ci-dessus : d’abord, un *corpus* de textes – *i.e.* un tout organique et signifiant – est décomposé en tokens, *i.e.* des unités de sens d’échelle inférieure – en l’occurrence des mots, eux-mêmes réduits à leur forme lemmatique et dont ont été expurgés les “*stopwords*”, c’est-à-dire des mots très courants et peu signifiants hors contexte, notamment les conjonctions de coordination, les articles, les déictiques... Ainsi, au début de la chaîne spécifiée dans l’article, des étapes de tokenisation, lemmatisation et nettoyage convertissent un corpus (non spécifié) de textes provenant de Google News en un ensemble de 6 milliards de *tokens*. L’ensemble obtenu peut conserver l’ordre des mots du document ; dans le cas contraire, il s’agit d’un sac de mots (*bag of words*).

L’étape suivante de la chaîne opératoire consiste à assigner un vecteur à chaque token. La contribution majeure de *word2vec* réside ici dans ses deux architectures neuronales, *Continuous Bag of Words* (CBOW) et Skip-gram, qui toutes deux prédisent un ou plusieurs mots étant donnés un ou plusieurs mots de contexte : en faisant glisser une fenêtre contextuelle d’une largeur variable (entre 2 et 20 mots), il est possible de générer les milliards de contextes linguistiques restreints désirés par la sémantique distributionnelle. Au sein de ces micros-ensembles de n mots, l’enjeu pour les deux architectures proposées par *word2vec* est de compléter les trous : un mot manquant pour CBOW, (n - 1) mots à prédire pour Skip-gram. Le rapport entre ce qui est su et ce qui doit être prédit est, toutefois, inverse pour les deux tâches. D’un côté, CBOW

consiste à prédire un mot étant donnés ceux qui le suivent et le précèdent. De l'autre, Skip-gram cherche à identifier, pour un seul mot donné, les mots autour. L'entraînement neuronal consiste alors à ajuster les paramètres pour optimiser ces prédictions, et améliorer progressivement la conversion d'un mot en vecteur à plusieurs centaines de dimensions[14]. Le déséquilibre entre les deux tâches justifie les avantages et inconvénients des deux architectures : rapide à entraîner mais mauvaise pour les mots rares pour CBOW, lente mais plus exhaustive pour Skip-gram. Toutefois, et alors que c'est Skip-gram qui s'imposera pour ses meilleures performances, les deux architectures semblent atteindre les ambitions des auteurs : à l'issue de la chaîne opératoire, un espace vectoriel euclidien multidimensionnel est donc synthétisé où chaque mot est représenté par un vecteur au voisinage linguistiquement cohérent. Déjà dès cette étape une première réduction herméneutique est opérée : un mot est transformé en vecteur à nombreuses dimensions qui n'ont pas de sens explicite, dont le nombre peut changer sans perte drastique de performance et qui échappent
à toute représentation mentale. Dès cette étape, si l'on veut filer la métaphore liquide, le sens semblerait "couler entre les doigts".

2.3. Produire du sens par déplacement

Une opération mathématique rendue possible par la synthèse de cet espace retient particulièrement l'attention des auteurs : l'algèbre vectorielle. Plus encore que la structure statique des points dans l'espace, c'est l'opération de déplacement entre deux points (c'est-à-dire, le vecteur)[15] qui est le véritable objectif[16]. À tout déplacement dans l'espace, on peut associer une opération linguistique conceptuelle. Ainsi l'exemple devenu fameux : vecteur("Roi") - vecteur("Homme") + vecteur("Femme") ≈ vecteur("Reine") (voir illus. 1). C'est donc en se "déplaçant à travers les données [*moving through data*]" (Mackenzie, 2017, p. 69) que le sujet peut produire du sens : "La transformation vectorielle des données est le substrat mobile de l'apprentissage automatique à mesure qu'il se développe, mais elle est largement considérée comme allant de soi, comme un espace donné ou du simple bon sens. [*Vectorized transformation of data are the moving substrate of machine learning as it expands, but they are largely taken for granted as given space or commonsense ground*.]" (Mackenzie, 2017, p. 67) C'est à l'aune de cette opération que l'espace vectoriel est évalué. À mesure que l'espace construit améliore sa cohérence, tout rapprochement fait sur la base d'opérations mathématiques (proximité, déplacement, etc.) devient une potentielle découverte sémantique. S'ouvre alors la quête d'un espace vectoriel toujours plus précis et qui nécessite plus de ressources, plus d'entraînement, plus de données, plus de paramètres.

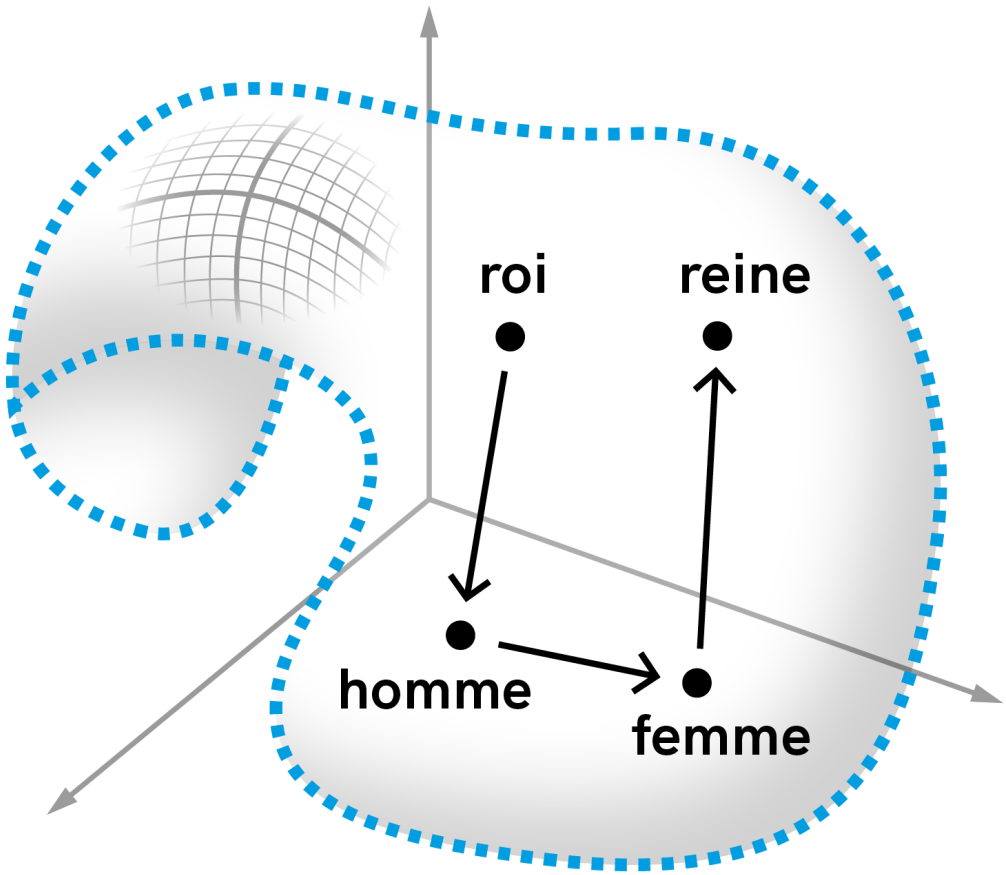


Illustration 1 : Représentation graphique de l'espace latent comme plongement mathématique permettant le calcul vectoriel.

Notons néanmoins que les passages qui rendent la vectorisation possible et opérante sont presque passés sous silence dans l'article séminal de Mikolov *et al.* (2013). Les premières étapes de tokenisation ne sont pas explicitées et aucune alternative ne semble possible. De même le choix des paramètres d'entraînement (*negative sampling*, *softmax* hiérarchique, fenêtre d'attention dynamique) sont mentionnés en passant, alors qu'ils forment la raison principale du succès de *word2vec* selon l'étude comparative de Levy *et al.* (2015). Ces étapes antérieures et postérieures de la chaîne opératoire – souvent oubliées par les discours d'escorte qui se concentrent sur la promotion d'architectures neuronales ré-employables et donc citables – doivent pourtant être au cœur de toute critique. En particulier, les opérations de préparation écrasent un grand nombre de dimensions du document original et sont révélatrices de pré-conceptions théoriques fortes.

Ces silences traduisent selon nous un *a priori* spatial chez les auteurs. Il y a, conceptuellement, une primauté de l'espace. Cette primauté s'exprime doublement. D'abord, temporellement, puisque c'est quelque chose que l'on impose à la donnée. Le plongement est une opération appliquée à des données, et non pas une propriété. Si les points sont *plongés* (ou même *embedded*) dans un espace continu, c'est qu'il faut bien d'abord construire mentalement cet espace et prendre comme *a priori* que la langue est continue – en dépit des instruments de la construction sociale de la langue dont les conventions (ex. grammaire), dispositifs (ex. dictionnaire) et institutions (ex. Académie française) en font un assemblage d'entités discrètes[17]. Avant toute opération algorithmique, c'est-à-dire avant même que cet espace ne soit tangible, on lit déjà le fantasme liquide. C'est sur la pré-existence de ce fantasme qu'il faudra s'interroger quant à la disparition de la théorie.

D'autre part, car l'espace est doté d'une intention propre. C'est ce que trahit l'usage du terme *espace latent*, aujourd'hui hégémonique en humanités numériques[18]. Cette équivalence spatio-conceptuelle comme nous l'avons appelée plus haut serait comme une heureuse conséquence : une cohérence conceptuelle qui, après être restée cachée dans la topologie se manifesterait à un certain moment, une fois l'apprentissage terminé[19]. Que cet espace ait un sens semble en effet surprendre ses concepteurs. Les auteurs de *word2vec* déclarent par exemple : "De manière quelque peu surprenante, il est possible de répondre à ces questions en effectuant de simples opérations algébriques sur la représentation vectorielle des mots. [*Somewhat surprisingly, these questions can be answered by performing simple algebraic operations with the vector representation of words*]" (Mikolov *et al.*, 2013, p. 5).

Une des raisons de cette surprise vient, au-delà d'un certain techno-optimisme, peut-être de la construction même de l'espace : entre l'inévitable *bricolage* des données et l'abstraction d'un corpus de texte en vecteurs numériques, l'émergence d'une forme de structure familière interpelle[20]. De la même manière qu'il faille passer par le discret des tokens pour créer le continu de l'espace latent, la théorie est (supposément) évacuée mais les résultats de sa recombinaison appellent une interprétation.

2.4. Inscrire le plongement dans un réseau

Ce potentiel interprétatif de l'espace vectoriel a rapidement séduit nombre de projets de recherche en humanités numériques où, à la suite des tentatives des *computational humanities*, le plongement a pris une place centrale dans l'arsenal des méthodes déployées. Mais comment, à partir d'un objet vectoriel aussi complexe, produit-on de la connaissance ? En effet, comme le rappelle James E. Dobson, le traitement de vecteurs nécessite une autre médiation algorithmique : "Les représentations de textes comme espaces vectoriels ne sont pas explicitement conçues pour être utilisées par des humains, mais pour être manipulées par des algorithmes. [*Vector space representations*

of texts are not explicitly designed for human consumption but for algorithmic manipulation]" (Dobson, 2020, p. 85). Nous pourrions formuler l'idée autrement : un espace vectoriel ne peut être interprété qu'en tissant un ensemble de relations internes et externes à l'espace et en naviguant ces relations. Ici, les nombreux projets d'humanités numériques nourrissent une taxonomie d'opérations.

Les opérations les plus fréquemment réalisées sont celles d'un calcul *au sein même* de l'espace de plongement – ce que Sybille Krämer regrouperait sous "sens intrinsèque [*intrisic meaning*]" (Krämer, 2021, pp. 7-9). Ces opérations de calcul intrinsèques à l'espace peuvent alors se décliner en deux types. D'une part, des opérations analytiques visant à découper l'espace en sous-ensembles signifiants (champs sémantiques ou *topics*) : par exemple, une analyse de la topologie de l'espace lui-même en visualisant des saillances, par *clustering* grâce à des calculs de similarités. D'autre part, des opérations de navigation au sein de l'espace permettant de générer ou classifier de nouveaux points.

Une seconde classe d'opérations regroupe celles relatives à la mise en relation de l'espace de plongement avec *d'autres* espaces – "sens extrinsèque [*extrinsic meaning*]", toujours selon Krämer (2021, pp. 7-9). L'espace de plongement peut alors être mis en relation avec des avatars de lui-même, comme c'est le cas par exemple avec le découpage d'un espace vectoriel en différentes tranches selon certains critères[21], ou bien par réduction de dimensionnalité, créant ainsi un espace équivalent non du point de vue mathématique, mais du point de vue de la description des données[22]. Mais il peut aussi être mis en relation avec des espaces autres, c'est le cas notamment pour des fonctions de traduction qui s'opèrent par la mise en équivalence d'espaces textuels de langues différentes ; ou encore pour la multimodalité qui "aligne" des espaces de natures parfois très différentes comme c'est le cas lorsqu'un *prompt* textuel est utilisé pour générer une image (voir illus. 2).

Notons que ces opérations sont souvent rendues possibles par des métadonnées discrètes extrinsèques à l'espace, permettant de filtrer selon certains critères, c'est-à-dire de distinguer et décrire. Mais par delà leurs différences, les usages interprétatifs du plongement sont structurés par un diptyque d'usage : un calcul interne à un milieu navigable, et l'inscription de ce milieu au sein d'un plus vaste réseau d'objets de natures diverses. L'on voit alors pourquoi et comment un espace de plongement ne peut se suffire à lui-même, si vaste et varié soit-il. En effet, la chaîne opératoire de *word2vec*, qui opère le passage du texte organique aux tokens discrets à l'espace continu, se doit à son tour d'être plongée dans un espace réticulaire. Apparaissent ainsi différentes strates comme autant de conditions de possibilité d'une herméneutique et parmi lesquelles l'espace de plongement continûment navigable.

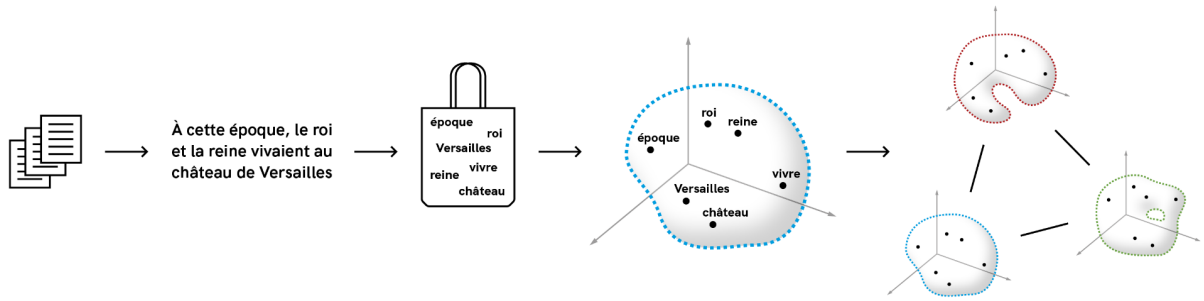


Illustration 2 : Schéma de la chaîne opératoire de word2vec (corpus – texte – sac de mots – espace latent – réseau d'espaces)

3. Des structures théoriques pour liquéfier

En-deçà de la promesse de fluidité athéorique faite par le quatrième paradigme, l'étude du plongement vectoriel, comme celui de *word2vec* ou ses concurrents modernes, montre que toute "mise en solution" nécessite à diverses étapes des *structures*. La fluidité des données et la navigation dans un espace sans couture, dépendantes de la synthèse d'un milieu continu, sont donc non seulement assurées par des opérations de structuration de celles-ci, mais encore s'inscrivent au sein d'une plus vaste écologie de *media* (sous-partie 1). Se dessine alors une généalogie d'ordres du savoir variés avec laquelle, loin de faire table rase, le quatrième paradigme entretient des relations de ruptures et de continuité (sous-partie 2). Dans cette perspective, il apparaît non seulement illusoire de se débarrasser de toute théorie, mais, dit autrement, la chaîne opératoire computationnelle est habitée ou travaillée par des structures théoriques, bien souvent spontanées (sous-partie 3), et qu'il convient de soumettre à une "herméneutique (du) numérique" afin d'en saisir les effets de sens (sous-partie 4).

3.1. Du plongement à l'espace latent comme medium

Du corpus de documents au sac de lemmes au plongement et à sa mise en relation avec d'autres objets, il paraît difficile de mobiliser le concept de représentation pour qualifier l'espace latent. En effet, on peut dire de celui-ci, avec James E. Dobson, qu'il est "anisomorphe au texte" (Dobson, 2022) quand bien même il *encode* une certaine information relative au corpus de documents et construite à partir de la fréquence d'occurrence relative des termes. Suivant Dobson, il est plus pertinent de mobiliser le concept de "médialité" pour comprendre la place de l'espace latent dans cette chaîne opératoire, et en particulier celui de re-médiation. Définissons largement un *medium* comme le lieu matériel et symbolique d'une médiation entre deux altérités – médiation

qui, par sa matérialité, façonne la relation qu'elle rend possible et l'émergence du sens en son sein (ex. Mersch, 2018).

On voit que l'espace latent joue le rôle de *medium* dans une chaîne de re-médiation, c'est-à-dire de traduction matérielle d'un *medium* à un autre (Dumas Primbault, 2025a), au cours de laquelle certaines qualités sont gagnées – calculabilité, visualisation (Krämer, 2022), "scalabilité" – tandis que d'autres sont perdues – épaisseur contextuelle, sens des textes, organicité du *corpus*. Dans cette perspective, l'espace latent apparaît comme une forme de médiation qui permet le calcul sur des entités mises en relation par plongement, donc *en son sein* et de manière continue. L'espace latent est donc un lieu au sens d'un espace fini et normé au sein duquel baigne le sujet de la connaissance médié par d'autres dispositifs algorithmiques comme des outils de clustering, de réduction de dimensionnalité ou de visualisation. Un espace dont la substance est la donnée même et dont la structure est le système d'interdépendance entre les unités discrètes qui ont présidé à sa synthèse et qui gisaient, non structurées, dans le sac de mot. Un "espace abstrait et inerte, en d'autres termes absolument maîtrisable" (Dumas Primbault *et al.*, 2021) comme condition de possibilité d'une pratique calculatoire mécanique et déterministe.

De plus, *medium* est un terme polysémique et s'il désigne l'espace dans lequel s'inscrivent des unités signifiantes et naviguent des sujets – ici un lieu plutôt qu'un milieu –, c'est aussi ce qui s'introduit entre deux altérités – un mi-lieu, un média. En effet, pris dans un réseau plus large d'autres objets et se prêtant ainsi à une série d'opérations de mise en équivalence, de comparaison, de mise en relation avec d'autres espaces et jeux de données, l'espace latent apparaît comme un *medium* inscrit dans une écologie de *media*. Une écologie cependant non continue, résultat de la mise en relation réticulaire d'outils et d'objets. La chaîne opératoire apparaît comme une succession de *media* : du corpus comme sac de mots (ensemble discret) à l'espace vectoriel comme espace continu à l'écologie d'outils comme espace réticulaire.

Medium résultant du plongement de données textuelles dans un espace vectoriel euclidien, l'espace latent est la double condition de possibilité de la navigation du sujet de la connaissance : d'une part au sein de l'espace lui-même par le calcul, d'autre part entre l'espace et d'autres objets. En effet, en inaugurant un espace continu de représentation des données textuelles et mettant celui-ci en relation avec d'autres espaces et d'autres données, il permet et conditionne la navigation d'un sujet médié au sein de cet espace même et entre d'autres espaces de médiation. Où l'on voit le rôle joué par l'infrastructure de données, en un sens métaphorique, qui est de canaliser le flux de données en imposant des conditions initiales et des conditions limites à l'écoulement liquide, en prescrivant une architecture informationnelle au milieu.

3.2. De la continuité de l'espace à la navigation liquide

Lieu à naviguer autant que *medium* inscrit dans un réseau de *media*, l'espace latent

généré par *word2vec* met en lumière deux principes majeurs de l'organisation des savoirs et de leur production à travers l'histoire (Foucault, 1969 ; voir aussi Mackenzie, 2017, ch. 3) : d'une part, la structure des espaces (physiques ou métaphoriques) de production, de recueil et d'ordonnancement des objets de la connaissance et, d'autre part, les modalités de navigation (physique ou médiée) des sujets de la connaissance au sein de ces espaces. Il en va ainsi de l'antique bibliothèque comme ambition de posséder tous les savoirs du monde selon un ordre encyclopédique et universel permettant le cheminement physique dans ses catalogues, ses rayonnages, ses magasins – autant que le contrôle de l'accès à ces savoirs. Et si l'on tire le fil des discours d'escorte relatifs à l'évolution récente de ce lieu de savoir paradigmatique qu'est la bibliothèque (Dumas Primbault, 2025b), il est possible de souligner dans quelle mesure l'espace continu synthétisé par *word2vec* s'inscrit à la fois en continuité et en rupture avec des ordres du savoirs antérieurs.

En effet, la bibliothèque numérique, permettant un accès plus granulaire aux documents qu'elle conserve et dont elle a "démembré les reliures" par numérisation, apparaît comme un instrument favorisant la recherche d'"information" entendue comme fragment réputé factuel. La bibliothèque en ligne, grâce à l'hypertexte, amplifierait les possibilités de navigation d'un écrit d'écran à l'autre, d'un fragment à l'autre au sein du navigateur, ce grâce à la "délinéarisation du document". Notons que cette structuration réticulaire de l'espace de représentation des objets de la connaissance n'est pas spécifique à la bibliothèque en ligne mais se retrouve dans une phase antérieure des humanités numériques qui faisait la part belle aux graphes de savoirs et aux ontologies. Enfin, la bibliothèque "en données" permettrait de "désiloter" les collections, les catégories, les disciplines afin d'ouvrir un espace de navigation sans couture, notamment au profit d'algorithmes permettant de traiter des jeux de données textuels non structurés[23].

L'ordre du savoir tel que prophétisé par le quatrième paradigme s'inscrit dans la continuité de deux tendances ici relevées : d'une part, la fragmentation selon une granularité toujours plus fine des unités de sens (rouleau – livre – chapitre – paragraphe – donnée) qui à la fois constituent l'espace à naviguer et sont représentées en son sein, et, d'autre part, les modalités de navigation de cet espace, toujours plus continues, fluides, liquides[24]. Cependant, tandis que la navigation était envisagée dans le paradigme computationnel (hypertextuel) comme le déplacement le long de liens entre des entités discrètes – comme c'est le cas dans un graphe, une ontologie, un vocabulaire contrôlé, un réseau –, le flux liquide du quatrième paradigme repose sur la synthèse d'un espace continu à partir d'entités discrètes[25]. Il y a donc, au cœur du quatrième paradigme, homologie entre liquidité de la donnée et navigation sans friction du sujet (médié) au sein d'un espace continu. Comme c'est le cas avec l'irruption des données dans les bibliothèques (Mesguich, 2023), la datafication du texte par *word2vec* inaugure un espace continu, *i.e.* "désiloté", faisant l'économie de la singularité des données, de leurs natures diverses, de leurs structures, afin de garantir une navigation sans friction en leur sein.

Loin d'être naturelle ou spontanée, l'homologie entre liquidité et navigation a pour condition de possibilité la continuité de l'espace au sein duquel données et sujets se meuvent – Lury *et al.* (2012) parlent "d'ordonnancement [de la culture] par le moyen de la continuité"[26]. C'est par une chaîne opératoire structurante (synthèse d'un espace vectoriel euclidien et mise en relation avec d'autres espaces) que, paradoxalement, le désilotage advient. En effet, l'espace latent généré par *word2vec*, s'il est navigable sans friction grâce à la médiation d'algorithmes d'analyse et de visualisation, est figé : le voisinage d'un vecteur ne change qu'à condition de synthétiser un nouvel espace à partir d'un nouveau jeu de données, d'une nouvelle paramétrisation ou d'un autre algorithme de plongement. Le plongement comme instrument de fabrique de la continuité est condition de possibilité de la navigation sans friction, au prix de la liquidité véritable des données qui se trouvent structurées, cristallisées.

3.3. Le retour de la théorie

Ainsi, l'objet vectoriel est structuré par un ensemble de postulats pourtant rarement explicités. Quels sont ces postulats ? Citons-en quelques-uns. Tout d'abord, le texte est pris comme suite de mots et non comme inscription matérielle dans un *medium* particulier (encre et papier, manuel ou machine à écrire, numérique). Cette suite de mots est par ailleurs "tokenisée", d'une façon stricte et homogène[27]. L'ordre des mots ne compte pas au sein des micros-ensembles d'entraînement, et seul le lemme est signifiant – faisant fi de toute nuance grammaticale. L'espace est statique, généré une fois à partir d'un corpus, et un *mot* y aura toujours la même représentation vectorielle.

Bien que certains de ces postulats n'aient plus cours à l'heure actuelle des grands modèles de langage qui permettent de prendre en compte toujours plus de contexte, ils illustrent néanmoins à travers *word2vec* notre propos sur la théorie implicite, et le même exercice s'appliquerait aux modèles contemporains[28]. Cette critique est particulièrement nécessaire alors que nous vivons depuis quelques années le passage, au sein des humanités numériques, d'un paradigme de l'hypertexte – celui des très nombreux projets de TEI – à la genèse d'un espace vectoriel unique où tout objet pourrait être plongé[29]. Par ceci, on évacue le processus d'explicitation manuelle de ces postulats, forcée par l'encodage en graphes et métadonnées. Si la vectorialisation n'invente pas la structure implicite de ses données, elle l'invisibilise.

Car derrière ces postulats, on trouve une épistémologie que l'on pourrait qualifier, avec Althusser (1974), de spontanée – c'est-à-dire qu'elle semble émaner directement de la *praxis* sans jamais être problématisée pour autant. Reprenons notre exemple : *word2vec* a dans sa généalogie le présupposé structuraliste selon lequel il est possible de capter une partie du sens du texte en le réduisant à des sacs de mots[30]. L'espace vectoriel lui-même charrie un ensemble d'implicites culturels : que la langue soit continue, comme discuté plus haut ; que son espace soit de plusieurs centaines de dimensions ; ou bien que la proximité sémantique soit une similarité conceptuelle – une

idée occidentale culturellement située comme le rappelle Seaver (2021, p. 48).

Gardant ceci en tête, on ne peut que reconnaître avec Andrew Goldstone et Ted Underwood la nécessité et la fertilité qu'offrent les variations de modèles de langages (2012). En jouant sur différentes étapes de la chaîne opératoire – du choix des données aux hyperparamètres de l'entraînement[31] – pour démultiplier les vectorialisations d'un même corpus, ils montrent la nécessité retrouvée d'une herméneutique textuelle plus traditionnelle, et complémentaire au kaléidoscope numérique.

3.4. Une herméneutique numérique

Car parmi les prophéties du quatrième paradigme, la "fin de la théorie" s'accompagne également de la fin de l'interprétation, la mort de l'herméneutique. En effet, l'inductivisme radical ancré dans une analyse athéorique de masses de données objectives est aussi la promesse d'une production "automatique" d'un savoir im-médiat, c'est-à-dire à la fois instantané et non-médié. En particulier non médié par une intelligence humaine à laquelle il reviendrait d'interpréter des données – ou le résultat d'un calcul sur celles-ci – car les données parleraient d'elles-mêmes :

> *Ce qui souvent permet de penser que l'écart entre calculi manipulés et signes interprétés peut être comblé est la sophistication des procédés techniques mobilisés. La complexité technique transformerait une différence de nature entre le sens et le calcul en une différence de degré : algorithmes plus performants, données plus massives seraient alors les moyens adéquats pour résorber cette déhiscence entre sens et calcul, interprétation et manipulation.* (Bachimont, 2021, p. 8)

L'ordre du savoir annoncé par le quatrième paradigme ne serait plus descriptif et explicatif, mais prédictif et normatif ; les savoirs produits sont évalués par leur action efficace sur le réel. En "adhérence" (Rouvroy et Berns, 2010) im-médiate avec le réel, les savoirs émergeraient de celui-ci sans thèse, modèle, hypothèse ou interprétation[32].

Le retour de la théorie en plusieurs endroits de la chaîne opératoire, censée pourtant l'escamoter pour assurer la liquidité des données et la navigation sans friction des sujets de la connaissance, annonce donc également le retour de l'herméneutique trop rapidement évacuée par les prophètes diluviens[33] – ou plutôt met en lumière son irréductible présence, car "*la manipulation technique ne peut être qu'une instrumentation permettant d'approximer le régime du sens et non de l'induire à partir des données textuelles ou sémiotiques*" (Bachimont, 2021, p. 3).

Identifiée comme une caractéristique majeure de l'épistémologie des humanités, la question herméneutique est beaucoup débattue dans le champ des humanités numériques depuis les années 2010 – parfois réduite à une opposition franche entre

calcul et interprétation, machine et humain, ou numérique et analogique (voir notamment Gibbs et Owens, 2012 ; van Zundert, 2016 ; Drucker, 2017 ; Dobson, 2022 ; Krämer, 2022). Ce que montre la chaîne opératoire de *word2vec*, c'est que pour transformer le texte en données calculables puis synthétiser un espace continu, il faut l'interpréter selon des schèmes, des théories, des modèles, c'est-à-dire faire travailler une herméneutique. Si l'on ne peut pas dire que le calcul soit une herméneutique – "il n'y a pas de texte pour une machine" (Bachimont, 2021, p. 10) –, on peut dire qu'il y a une ou des herméneutiques à l'oeuvre dans la chaîne opératoire, qu'elles soient pratiquées par des humains ou incarnées par des *media* – des théories, des modèles, des formats – qui sont nécessaires à la structuration des données, à la conception des algorithmes et à l'organisation visuelle des résultats[34]. C'est-à-dire que la chaîne opératoire de traitement des données mobilise nécessairement quelque chose qui n'est pas déjà dans les données mêmes, ainsi que le voudrait la fin de la théorie. Calcul et herméneutique sont bien distincts mais inséparables : si l'analyse de données ne produit pas de sens sans interprétation humaine, elle mobilise néanmoins des structures de sens déposées dans la machine par le sujet :

> *"les processus de sélection, d'analyse et de visualisation ont leur propre herméneutique. Le dialogue sur l'herméneutique des humanités numériques ne peut donc pas se contenter d'opposer le quantitatif et le qualitatif, et de reléguer l'herméneutique à un aspect qualitatif de l'interprétation de données, comme si ces données n'étaient pas déjà chargées de valeurs et interprétées. [processes of curation, analysis, and visualization have a hermeneutics of their own. The dialog on the hermeneutics of digital humanities cannot therefore simply posit a dichotomy between the quantitative and qualitative, and relegate hermeneutics to a qualitative aspect of interpretation of given data as if this data would not be value-laden and interpreted already.] "* (van Zundert, 2015, p. 335)

En définitive, le calcul est toujours déjà travaillé par une ou des herméneutiques et, en ce sens, procède de l'incalculable : la manipulation calculatoire n'est rendue possible qu'au prix de réductions interprétatives successives. Ainsi, l'objet de la présente contribution, et de toute pratique réflexive des humanités numériques, devient alors de développer une "herméneutique numérique" (Fickers *et al.*, 2022) afin de débusquer les effets de sens inscrits dans les chaînes opératoires en dépit des discours d'escorte du quatrième paradigme, c'est-à-dire une réflexion critique sur ce que le numérique fait aux matériaux de la recherche et aux épistémologies des humanités et des sciences sociales.

4. Conclusion : pour un archipel de communautés épistémiques

Point de déluge de données, en chassant la théorie par la porte, elle revient par la

fenêtre. Nécessaire à la structuration des données – de leur format, à leur encodage, leur analyse, leur visualisation et leur mise en relation avec d'autres –, elle est même une condition de possibilité de la continuité d'un espace dès lors navigable par un sujet médié de la connaissance. Ce dont atteste la survivance de la théorie au cœur d'un quatrième paradigme qui n'en voudrait pas, c'est l'inextricable coexistence de paradigmes sédimentés au fil de l'histoire des savoirs. Une diversité épistémique qu'il convient de cultiver plutôt que de prôner une table rase :

> *"En réalité, les méthodes numériques ne remplacent pas l'arsenal des méthodes traditionnelles utilisées dans les sciences humaines, mais sont plutôt un supplément qui s'y ajoute. [...] Ces procédures ne constituent pas en elles-mêmes une forme d'herméneutique, mais elles sont indissociablement intriquées aux décisions et interprétations humaines à toutes les étapes ; elles complètent les méthodes d'interprétation utilisées dans les sciences humaines. [In fact, digital methods are not a replacement, but rather a supplement and addition to the arsenal of methods in the Humanities. [...] These procedures themselves are not performing hermeneutics; but at all stages, they are indispensably interwoven with human decisions and interpretations; they are complementary to the interpretative methods of the Humanities.] "* (Krämer, 2023, p. 14)

Et s'il fallait trouver une métaphore liquide nouvelle pour remplacer celle du déluge et promouvoir la diversité épistémique, nous pourrions nous tourner vers le "paradigme de l'archipel" d'Edouard Glissant (Glissant, 1990; Noudelmann *et al.*, 2020). Le lac, plutôt qu'une masse informe, devient le point de rencontre de sources multiples et l'étendue liquide au milieu de laquelle sédimentent des "îlots de cohérence" (Boullier et Ghitalla, 2004, p. 174) structurés de manière critique et réflexive grâce à une "herméneutique numérique". Un archipel qui tirerait sa force de la mise en relation des diversités qui l'habitent.

5. Bibliographie

Althusser, L. (1974). Philosophie et philosophie spontanée des savants (1967). Paris : Maspero.

Anderson, C. (2008). The End of Theory: The Data Deluge That Makes the Scientific Method Obsolete. Wired. https://www.wired.com/2008/06/pb-theory/

Bachimont, B. (2021). La complexité herméneutique à l'épreuve du calcul. Interfaces numériques, 10(3). doi: 10.25965/interfaces-numeriques.4686

Bauman, Z. (2013). La vie liquide. Paris : Fayard.

Bengio, Y., Ducharme, R., Vincent, P. et Jauvin C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research 3, 1137–1155. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bernardot, M. (2018). Plongée dans les métaphores et représentations liquides de la société numérique. Netcom, 32(1/2). doi: 10.4000/netcom.2886

Boullier, D. et Ghitalla, F. (2004). Le Web ou l'utopie d'un espace documentaire. Information-Interaction-Intelligence, vol. 4, no. 1, 173-189.
Cardon, D., Cointet, J.-P. et Mazières, A. (2018). La revanche des neurones L'invention des machines inductives et la controverse de l'intelligence artificielle. Réseaux, 211(5), 173-220. doi: 10.3917/res.211.0173.
Castells, M. (2000). The Rise of the Network Society. Hoboken : Blackwell Publishers.
Chartier-Edwards, N., Grenier, E., et Roberge, J. (2024). If (world2vec) Then vec2politics: On Machine Learning and the Performativity of Recursive Power. In : Brzeziński, D., Filipek, K., Piwowar, K., et Winiarska-Brodowska, M. (dir.), Algorithms, Artificial Intelligence and Beyond. Londres, Routledge, 139-154.
Daston, L. et Galison, P. (2012). Objectivité. Dijon : Presses du réel.
Denis, J. (2018). Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales. Paris : Presses des Mines. doi: 10.4000/books.pressesmines.3934
Devlin, J., Chang. M., Lee, K., et Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Arxiv. https://arxiv.org/abs/1810.04805
Dobson, James E. (2022). Vector hermeneutics: On the interpretation of vector space models of text. Digital Scholarship in the Humanities, 37(1) 81–93,doi: 10.1093/llc/fqab079
Dubin, D. (2004). The Most Influential Paper Gerard Salton Never Wrote. Library Trends, vol. 52, no. 4, 748–764.
Dumas Primbault, S., Tortosa, P.-A., et Vailly, M. (2021). Introduction - Milieux, media, écologie des savoirs. Cahiers François Viète, III(10). doi: 10.4000/cahierscfv.381
Dumas Primbault, S. (2022). Leibniz en Créateur. In : Briegel, F. (dir.) Matières à raisonner. https://savoirs.app/en/articles/leibniz-en-createur
Dumas Primbault, S. (2025a). Introduction : Voir l'archive par le travers. Sens public. https://sens-public.org/dossiers/1785/
Dumas Primbault, S. (2025b). Du Poste de Lecture Assisté par Ordinateur (P.L.A.O.) au schéma numérique. Trois étapes de la re-médiation de la BnF. Sens public. https://sens-public.org/articles/1796/
Drucker, J. (2017). Why Distant Reading Isn't. PMLA, vol. 132, no. 3, 628-635.
Fickers, A., Tatarinov, J. et Heijden, T. (2022). Digital history and hermeneutics – between theory and practice: An introduction. In : Fickers, A. et Tatarinov, J. (dir.), Digital History and Hermeneutics: Between Theory and Practice, Berlin, Bosto, De Gruyter Oldenbourg, 1-20. doi: 10.1515/9783110723991-001
Firth, J. (1962). Studies in Linguistic Analysis. Oxford: Blackwell.
Foucault, M. (1969). L'Archéologie du savoir. Paris : Gallimard.
Gavin, M. (2020). Is There a Text in My Data? (Part 1): On Counting Words. Journal of Cultural Analytics, 5(1). doi: 10.22148/001c.11830.
Geoghegan, B.D. (2023). The Statistical Order of Discourse: How Information Theory Encoded Industrial and Political Discipline. In : Gabrillo, J. et Zetter N. (dir.), Articulating Media: Genealogy, Interface, Situation, Open Humanities Press, Ch. 3.
Glissant, E. (1990). Poétique de la Relation: Poétique III. Paris: Gallimard.
Goldstone, A. et Underwood T. (2012). What Can Topic Models of PMLA Teach Us About

the History of Literary Scholarship? Journal of Digital Humanities 2(1). https://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-under wood-and-andrew-goldstone/
Gibbs, F. et Owens, T. (2012). The Hermeneutics of Data and Historical Writing. In : Dougherty, J. et Nawrotzki, K. (dir.). Writing History in the Digital Age, Ann Arbor, University of Michigan Press, 159-170. doi: 10.2307/j.ctv65sx57.18
Gingold, A. (2026). Data manibus plenis: The future of an automatism. A paraître.
Hey, T., Tolle, K.M., et Tansley, S. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond : Microsoft Research.
Kinsey, C. (2020). Fluid Dynamics: On the Representation of Water and Discourses of the Digital. Art History, vol. 43, no. 3, 510–537. doi: 10.1111/1467-8365.12483
Krämer, S. (2023) Should We Really 'Hermeneutise' the Digital Humanities? A Plea for the Epistemic Productivity of a 'Cultural Technique of Flattening' in the Humanities. Journal of Cultural Analytics, 7(4). doi: 10.22148/001c.55592.
Krämer, S. et Ljungberg, C. (2016). Thinking in Diagrams: The Semiotic Basis of Human Cognition. Berlin : de Gruyter.
Langlais, P.-C. (2023). ChatGPT : comment ça marche ? Sciences communes. doi: 10.58079/twxr
Levy, O., Goldberg, Y. et Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the Association for Computational Linguistics, vol. 3, 211–225.
Lindgren, S. (2020). Data Theory. Medford, MA: Polity.
Lury, C., Parisi, L. et Terranova, T. (2012). Introduction: The Becoming Topological of Culture. Theory, Culture, & Society 29(4/5), 3-35. doi: 10.1177/0263276412454552
Mackenzie, A. (2017). Machine Learners: Archaeology of a Data Practice. Cambridge : MIT Press.
Mersch, D. (2018). Théorie des médias. Une introduction. Dijon : Les presses du réel.
Mesguich, V. (2023). Les bibliothèques face au monde des données. Villeurbanne : Presses de l'enssib. doi: 10.4000/books.pressesenssib.17655
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. et Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Arxiv. https://arxiv.org/abs/1301.3781
Nolin, J.M. (2019). Data as oil, infrastructure or asset? Three metaphors of data as economic value. Journal of Information, Communication and Ethics in Society, 18(1), 28-43. doi: 10.1108/jices-04-2019-0044
Noudelmann, F., Simasotchi-Bronès, F. et Toma, Y. (2020). Le paradigme de l'archipel. In : Noudelmann, F., Simasotchi-Bronès, F. et Y. Toma (dir.). Archipels Glissant, Vincennes, Presses universitaires de Vincennes, 7-11. doi: 10.3917/puv.noude.2020.01.0007.
Offert, F. (2020). "Intuition and Epistemology of High-Dimensional Vector Space", https://zentralwerkstatt.squarespace.com/blog/vsm
Puschmann, C. et Burgess, J. (2014). Big Data, Big Questions| Metaphors of Big Data. International Journal of Communication, vol. 8, 1690-1709.
Rouvroy, A. et Berns, T. (2010). Le nouveau pouvoir statistique Ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps « numériques »... Multitudes, vol. 1, no. 40, 88-103. doi: 10.3917/mult.040.0088.

Seaver, N. (2022). Computing Taste: Algorithms and the Makers of Music Recommendation. Chicago : The University of Chicago Press.
Stark, L. et Hoffmann, A.L. (2019). Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture. Journal of Cultural Analytics 4(1). doi: 10.22148/16.036.
Sutherland, T. (2013). Liquid Networks and the Metaphysics of Flux: Ontologies of Flow in an Age of Speed and Mobility. Theory, Culture & Society, 30(5), 3-23. doi: 10.1177/0263276412469670
s.a. (2019). FP 01. Chaîne opératoire. Techniques & Culture. doi: 10.4000/tc.11410
Wark, M. (2015). The Vectoralist Class. e-flux, vol. 65. https://www.e-flux.com/journal/65/336347/the-vectoralist-class/
Weaver, W. (1949). Translation. Memorandum to Rockefeller Foundation.
Wevers, M. et Koolen, M. (2020). Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings. Historical Methods: A Journal of Quantitative and Interdisciplinary History 53(4): 226–43. doi: 10.1080/01615440.2020.1760157.
van Zundert, J.J. (2015). Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics. In : Schreibman, S., Siemens, R. et Unsworth, J. (dir.), A New Companion to Digital Humanities, Wiley-Blackwell, pp. 331-347. doi: 10.1002/9781118680605.ch23

6. Notes

[1] Ces auteurs ont contribué également à ce travail. Nous remercions chaleureusement nos deux évaluateurs, qui sont également les deux éditeurs de ce numéro spécial : Lise Verlaet et Bruno Bachimont. Leurs commentaires nous ont permis en plusieurs endroits de clarifier et souligner certains éléments cruciaux de notre raisonnement.
[2] Sur l'infrastructure et le "travail invisible des données", voir Denis (2018).
[3] Plus généralement, ces métaphores s'inscrivent dans un imaginaire liquide contemporain, celui d'un réel sans structure (Bauman, 2006), dans un permanent mouvement de flux et reflux d'information et de sujets (Castells, 1998).
[4] Cette problématique de la synthèse du continu évoque des précédents dans l'histoire des sciences et des techniques. Celui du calcul infinitésimal de Leibniz, par exemple, qui permet de reconstruire, asymptotiquement, la continuité d'une courbe ou d'une aire par l'intégration – somme continue – d'éléments discrets infinitésimaux. Ou bien, dans un prolongement technique à cette problématique, inauguré par les travaux de Shannon en traitement du signal, la reconstruction d'un signal sonore à partir de son échantillonnage.
[5] D'autres procédés que celui choisi relèvent de cette même logique de construction du continu liquide par structuration des données : l'alignement de vocabulaires contrôlés ou d'ontologies, en particulier dans l'objectif de converger vers un vocabulaire ou une ontologie uniques, ainsi que les modèles de diffusion qui produisent des images à partir du bruit.
[6] L'usage du terme "lac de données" reste encore fortement inscrit dans son contexte

infrastructurel original : un réservoir de larges quantités de données, le plus souvent organisé autour d'un graphe de connaissances. Si ce terme est plus proche de l'esprit du web sémantique que de l'apprentissage profond, la métaphore liquide nous interpelle néanmoins et pointe vers le même fantasme de fluidité. Elle témoigne selon nous d'un point commun avec ce que nous développerons : la conversion de larges données en un format homogène, qui ouvre un régime de comparaison, de combinaison, d'inférence sans limite.

[7] "L'exploration du potentiel herméneutique du calcul est un défi que les humanités numériques pourraient se lancer, à elles-mêmes et à l'informatique au nom des sciences humaines. Cela n'implique pas nécessairement de transformer la nature du calcul, mais cela doit passer par une remédiation de la nature de l'herméneutique par le biais une logique et une conception computationnelles fortement influencées par un dialogue avec les sciences humaines. Il est évident que les méthodes scientifiques issues des sciences humaines seraient plus appropriées que le scientisme pour l'intelligence artificielle et l'informatique lorsqu'elles interagissent avec les sciences humaines. [Exploration of the hermeneutic potential of computation is a challenge that digital humanities could pose, to itself and to computer science on behalf of the humanities. This need not imply transforming the nature of computation, but it must involve remediating the nature of hermeneutics through computational logic and design informed strongly by a dialog with humanities. It is apparent that scientific methods deriving from the humanities would be more appropriate than scientism for artificial intelligence and computer science when interacting with the humanities]" (van Zundert, 2016, pp. 344-5)

[8] Notre analyse s'applique donc, mutatis mutandis, aux autres solutions algorithmiques au problème de la représentation vectorielle du monde, et qui à l'ère du deep learning et de l'hégémonie des LLMs nous apparaît centrale.

[9] Financé par la fondation Rockefeller, important mécène de la cybernétique d'après-guerre (Geoghegan, 2023), Warren Weaver préconise, d'abord dans une lettre à Nobert Wiener de 1947 puis dans un mémorandum de 1949 sur la traduction automatique, l'usage d'un "masque opaque" pour déterminer le sens d'homonymes (Weaver 1949). Ce qui est depuis appelé une "fenêtre d'attention" est aujourd'hui un composant de tout modèle de langue, y compris word2vec.

[10] Pierre-Carl Langlais rapproche la sémantique distributionnelle de Weaver et Firth des travaux de Roman Jakobson dont l'aura intellectuelle aurait influencé Tomas Mikolov, premier auteur du papier de word2vec (Langlais, 2023).

[11] Il faudrait toutefois nuancer ici que world2vec ne propose pas, à proprement parler, une architecture neuronale profonde. C'est au contraire sa shallowness qui rend le modèle si facile d'implémentation. Il n'en reste que sa méthode d'entraînement aussi bien que son ambition épistémique seront celles du deep learning.

[12] "Nous prévoyons également que les vecteurs de mots de haute qualité deviendront un élément constitutif important des futures applications de TALN (traitement automatique des langues naturelles). [We also expect that high quality word vectors will become an important building block for future NLP applications.]" (Mikolov et al., 2013, p. 8)

[13] De fait, en 2019, Google passe d'un mode de recherche par mots clés, basés sur un

réseau de renvois et de mentions, à une interprétation du langage naturel par BERT, lui même pensé comme une extension de word2vec, capable de prendre le contexte d'un mot en compte, gérant ainsi les homonymies, et des phrases complètes (Devlin et al., 2018).

[14] Nous employons ici la définition de Robert Cresswell, dérivée des travaux d'André Leroi-Gourhan, à savoir une « série d'opérations qui transforment une matière première en un produit, que celui-ci soit un objet de consommation ou outil fini » (s.a., 2019).

[15] Le papier original fait osciller les dimensions de ses vecteurs entre 50 et 1000, une fourchette qui restait encore typique jusqu'à l'avènement des grands modèles de langage, aujourd'hui à plusieurs dizaines de milliers de dimensions.

[16] Notons ici la difficulté sémantique : chaque mot est associé, dans un espace, à un point – que l'on nomme vecteur de par ses larges dimensions, mais le déplacement entre ces mots-points est aussi nommé vecteur. De cette confusion naît aussi l'idée que le déplacement-vecteur peut être associé à un mot-vecteur.

[17] Ce que les auteurs appellent "régularités linéraires" : "Nous essayons d'optimiser la précision de ces opérations vectorielles en développant de nouvelles architectures de modèles qui préservent les régularités linéaires entre les mots. [We try to maximize accuracy of these vector operations by developing new model architectures that preserve the linear regularities among words.]" (2013, 2)

[18] Par ailleurs, la continuité de l'espace dans lequel on projette la langue ouvre l'épineux problème de l'interpolation de cet espace. Dans le cas des images, elle est source de nouvelles représentations (hallucinations ou autres) à la base de l'IA générative. Qu'en est-il de l'interpolation textuelle ?

[19] Il convient de noter que, pour une fois dans le lexique mondial de l'apprentissage machinique, les domaines linguistiques voisins convergent tous vers cette définition d'une émergence de sens – puisque l'on parle de latent space, spazio latente, latenten Raum ou bien espacio latente.

[20] Nous adaptons ici la définition de latent donnée par le TLFi : "Qui n'est pas manifeste, qui reste caché mais demeure susceptible d'apparaître, de se manifester à un certain moment".

[21] Un entre-deux que remarquait déjà Nick Seaver : "Bien qu'ils présentent la rigueur de la théorisation mathématique et du traitement algorithmique, ils répondent rarement aux normes de la preuve mathématique ou de la vérification expérimentale. [...] Ils font quelque peu maladroitement le pont entre les tendances méthodologiques interprétatives et positivistes. [Although they present the rigour of mathematical theorizing and algorithmic processing, they rarely meet standards of mathematical proof or experimental verification. [...] They somewhat awkwardly bridge interpretative and positivist methodological tendencies.]" (2021, 58)

[22] C'est le cas par exemple des études diachroniques où le jeu de données est "feuilleté" en fenêtres temporelles et génère plusieurs espaces de plongement qu'il est ensuite possible de mettre en relation et comparer par des métriques diverses (ex. Wevers et Koolen, 2020).

[23] appelons qu'après avoir placé les points dans un espace à grandes dimensions, il est commun d'avoir recours à une réduction de ces dimensions, grâce à certaines techniques comme t-SNE ou UMAP. Ceci présente le double avantage d'un côté, d'éviter

le “fléau de la dimension [curse of dimensionality]” (où l’espace est incalculable) et de l’autre, de pouvoir rendre imagé cet espace inimaginable. Ainsi, cet espace vectoriel est visualisable et peut rejoindre la longue généalogie des méthodes scientifiques basées sur le diagramme (Krämer et Ljungberg, 2016).

[24] Ainsi, lieux historiques de production, de structuration, et de mise à disposition de données (Bermès, préface à Mesguich, 2023), les bibliothèques deviennent, à l’aune du quatrième paradigme, “bases de données” (Hey et al., 2009, p. 27).

[25] Notons que suivant cette tendance, plus l’espace est continu et les données liquides, moins le sujet incorporé se déplace, plus il est médié par des outils (numériques) qui se déplacent pour lui.

[26] Si nous nous en tenons aux discours d’escorte textuel ici, il existe néanmoins de nombreuses vidéos d’escorte qui font usage du morphing, ou fondu enchaîné, pour montrer la continuité liquide entre les choses (du passé et du présent, entre les images, etc…) et que l’on retrouve par ailleurs dans les modèles multimodaux permettant des voyages visuels dans l’espace latent.

[27] Citées par Mackenzie (2017) qui ajoute : “L'expression ‘ordonnancement de continuité’ est intéressante, car nous ne considérons généralement pas les continuités comme soumises à un ordre. À bien des égards, ce qui est continu porte en soi son propre ordre, sa propre sériation ou stratification immanente. [The phrase ‘ordering of continuity’ is interesting because we do not normally think of continuities as subject to ordering. In many ways, that which is continuous bears within it its own ordering, its own immanent seriation or lamination.]” (p. 65)

[28] Symptôme de ceci, les auteurs de word2vec sont incapables d’utiliser New York comme exemple de ville car ses deux mots doivent être séparés.

[29] Ce que leurs formats propriétaires, pour les plus hégémoniques, rendent impossible à investiguer.

[30] Le projet Impresso (https://impresso-project.ch/) est un excellent exemple de cette tendance. Il entend rassembler en un seul grand plongement des données des archives de presse et de radio conservées sous forme textuelle, sonore, parfois visuelle par plus de vingt institutions européennes. Un tel espace présumé “sans couture” [seamless] permettra par exemple aux usagers de naviguer d’une coupure de presse française des années 1920, numérisée et océrisée, à une émission de radio luxembourgeoise de 1950.

[31] En ce sens, on pourrait discuter de la lecture faite par Simon Lindgren (2020) qui rapproche word2vec – et ses vecteurs résolument stables et statiques – de la linguistique poststructurale de Mouffe et Laclau.

[32] Leurs travaux prédatent word2vec, et leurs vectorisations sont obtenues par LDA. Mais l’idée serait également applicable aux plongements modernes.

[33] Où l’on renoue ici à nouveau avec une antienne de l’histoire des sciences : l’instrument “autographe” qui permettrait à la nature de s’inscrire im-médiatement sur un support neutre (cf. par exemple Daston et Galison, 2012).

[34] Aucun terme du champ sémantique de l’interprétation, de la signification, du sens, de l’herméneutique n’apparaît dans l’article de Chris Anderson (2008), tandis que l’ouvrage de Hey et al. (2009) en fait un usage très sporadique, ex. “insensible aux interprétations individuelles [insensitive to individual interpretations]” (p. 16) ou “extraire du sens à

partir d'ensembles de données volumineux et complexes [extract meaning from large and complex datasets]" (p. 161).

[35] "Les choix et les méthodes utilisés dans la conception de logiciels façonnent l'herméneutique des humanités numériques. La modélisation incarne la vision du monde du concepteur du modèle, son contexte et ses décisions subjectives. Les modèles de données sont tout sauf neutres ; au contraire, ils constituent une sélection délibérément spécifique de catégories et de propriétés sémantiques. Les langages de programmation relèvent de paradigmes qui influencent l'herméneutique. De plus, l'influence réciproque des méthodes informatiques sur l'herméneutique des humanités numériques va au-delà de la conception de logiciels. Les choix effectués dans la conception analytique d'un projet donné en humanités numériques influencent son herméneutique. Les choix concernant les propriétés à quantifier, les fonctions de distribution de probabilité à sélectionner et les tests statistiques à utiliser sont, par essence, influencés par l'herméneutique. [The choices and methods involved in software design do shape the hermeneutics of digital humanities. Modeling encompasses the worldview of the model designer, her context, and her subjective decisions. Data models are anything but neutral – on the contrary, they are a purposefully specific selection of semantic categories and properties. Programming languages have paradigms that affect hermeneutics. Moreover, the reciprocal shaping of the hermeneutics of digital humanities by the methods of computer science extends beyond software design. The choices made in the analytical conception of any given digital humanities project affect its hermeneutic makeup. The choices of what properties to quantify, what probability distribution functions are chosen, which statistical tests are used, are in essence hermeneutically informed.]" (van Zundert, 2016, pp. 343-344)

[36]