Outillages numériques pour les humanités : cartographier des réseaux d'influences

S. Szoniecky, *MCF, Laboratoire Paragraphe - Paris 8*, M. Louâpre, *MCF, CERILAC - Paris VII*

***Résumé* — Cet article présente les outils numériques utilisés pour cartographier des réseaux d'influences entre littérature et biologie au XIXe siècle. Dans le cadre du projet ANR *Biolographes* réunissant des chercheurs français et allemands. Nous avons testé des outils existants et développé un prototype d'application Web pour la modélisation des influences. Les expérimentations ont montré l'utilité des outils existants pour rapidement constituer des données et les explorer à travers de visualisations cartographiques. Mais pour optimiser l'expression, la diffusion et l'interopérabilité des interprétations faite par les chercheurs, nous avons développé un prototype pour modéliser des points de vue sous la forme de monades numériques.**

***Mots clefs*— humanité numérique, modélisation, influence, monade, écosystème d'information.**

I. INTRODUCTION

L'USAGE des outils numériques dans le cadre des recherches en sciences humaines et plus spécifiquement des humanités (littérature, histoire, philosophie…) devient depuis quelques années un impératif pour gérer les quantités de plus en plus importantes d'information rendues disponibles par l'ouverture des dépôts d'archives numérisées. La constitution de ces archives numériques ouvre souvent la porte aux financements pour des études qui sans la dimension « numérique » auraient bien moins d'écho chez les financeurs. De fait, les outils informatiques prennent une place prépondérante à chaque étape du travail scientifique : définition des sources, méthodologie d'analyse, édition et diffusion des documents de recherche…

Parmi le vaste outillage numérique mis à disposition des chercheurs, nous nous sommes intéressé aux outils de visualisation, plus particulièrement ceux permettant de représenter des réseaux d'influences. Dans un premier temps, nous avons expérimenté des outils disponibles sur le Web en nous interrogeant sur leurs efficacités et leurs limites. A partir de ces analyses qui seront présentées dans la première partie de ce document, il nous est apparu qu'il était nécessaire de fournir aux chercheurs un outil plus efficace pour comparer les différents points de vue qui s'exprime sur un même objet de recherche. Pour ce faire nous conçu un prototype pour modéliser les points de vue des chercheurs sous la forme d'un réseau de relations spatialisées et temporalisées entre des documents, des acteurs et des concepts. Après avoir présenté ce prototype dans la deuxième partie de ce document, nous discuterons dans une troisième partie, des limites de notre approche et de ces éventuels dépassements par l'utilisation d'une méthode générique d'analyse des écosystèmes d'information par modélisation de monades numériques et par la mise en place de processus d'intelligence collective compris comme :

« la capacité des groupes humains à collaborer sur le plan intellectuel pour créer, innover et inventer. Cette capacité peut-être appliquée à n'importe quelle échelle, des petits groupes de travail jusqu'à l'espèce humaine en passant par des réseaux de toutes tailles. » [11, p. 105]

II. OUTILS DE VISUALISATION ET D'EDITION DE RESEAUX D'INFLUENCES

Dans le cadre du projet ANR Biolographes réunissant des chercheurs français et allemands, nous avons pour objectif de représenter les réseaux d'influences entre littérature et biologie en Europe au XIXe siècle.

Pour mener à bien ce projet, nous avons dans un premier testé des outils disponibles sur le Web avec les données d'un tableur conçu par Muriel Louâpre dans le cadre de ces recherches sur la biographie de Jules Michelet.

A. Google Fusion

L'outil Google fusion[1] nous a permis de consolider les données issues du tableur et de tester plusieurs modes de représentation. Grâce à cet outil, il est assez facile de « jouer » avec les données pour tester la pertinence des représentations.

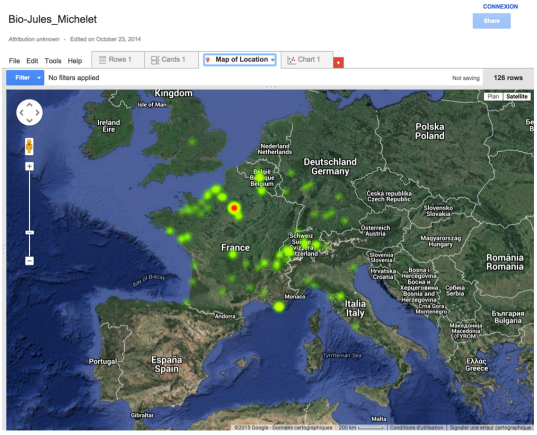

Figure 1 : carte des lieux fréquentés par Jules Michelet

[1] url vers les documents google fusion : https://www.google.com/fusiontables/DataSource?docid=1yXn2IaMaWFAThyZz0uAQPnFCGAhXcvfFfGrKH_3a#chartnew:id=4

Cette carte montre les lieux fréquentés par Jules Michelet ainsi que l'intensité de ces fréquentations : plus le lieu est brillant et rouge, plus le lieu est fréquenté. L'intérêt de cette carte et d'évaluer rapidement les zones d'activités de cette personnes ou au contraire les endroits qu'elle n'a pas fréquentés.

La représentation suivante montre les relations entre le type et le lieu du séjour.

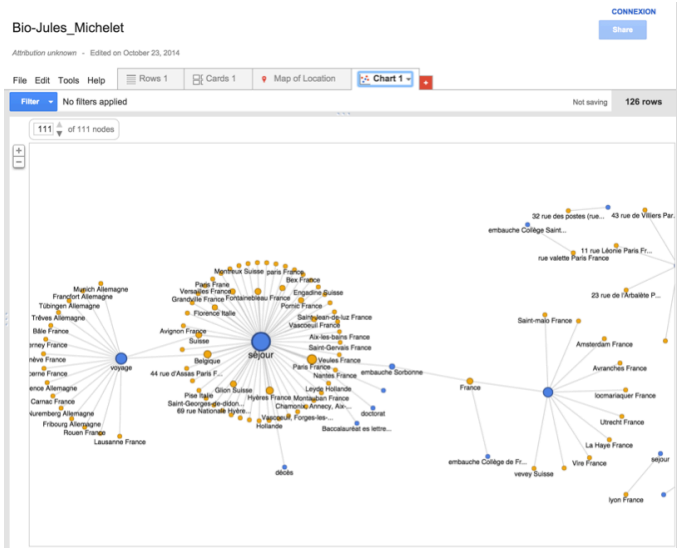

Figure 2 : Relation entre les types et les lieux de séjour

Ce premier travail de visualisation fait prendre conscience aux chercheurs des incohérences dans les données, notamment en terme de doublons, d'informations inexploitables ou d'ambiguïté dans le choix des catégories. Par exemple, il a fallut expliciter la différence entre « voyage » et « séjour » pour comprendre qu'il était question de durée du déplacement et de récurrence de ceux-ci.

B. Palladio

De même que Google Fusion, Palladio[2] permet de tester des représentations à partir d'un jeu de données. On retrouve les mêmes types de représentation par carte et en réseau avec la fonctionnalité supplémentaire d'associer une chronologie pour l'exploration des données.

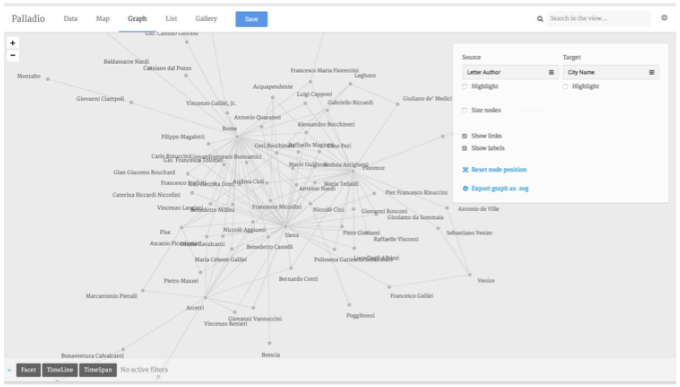

Figure 3 : Palladio représentation de réseau

[2] url vers l'outil : http://palladio.designhumanities.org

Conçu spécifiquement dans la perspective d'outiller les chercheurs en humanités numériques, Palladio est développé par le laboratoire de recherche en sciences humaines et Design de l'Université de Stanford[3]. Palladio est utile pour la visualisation et la découverte multidimensionnelle de relations dans le temps et l'espace. Pour créer un projet Palladio, les utilisateurs téléchargent leurs données dans un fichier de tableur ou CSV ou sous la forme d'une requête dans le format SPARQL (pour les utilisateurs avancés). Une fois qu'un utilisateur a créé un projet Palladio, il peut enregistrer le projet localement car aucune donnée n'est stockée sur des serveurs Palladio. L'utilisateur peut alors distribuer et partager son projet ou re-télécharger sur Palladio pour un développement continu.

Mais, comme le précise [7] : « Palladio does not have the most intuitive of interfaces, and the documentation could be more robust ». En effet, il est assez long est difficile de prendre en main cet outil notamment pour modéliser les relations entre différentes tables. De même, nous n'avons pas réussi à relier nos données avec SPARQL alors que cette fonctionnalité peut s'avérer très utile.

C. Keshif

Keshif[4] est une librairie informatique écrite en javascript qui permet de naviguer dans des données suivant des facettes thématiques, géographiques, chronologiques... Chaque facette est interconnectée avec les autres ce qui permet de filtrer rapidement les informations pour obtenir des données sur des questions complexes. Facilement paramétrable cette librairie permet de proposer rapidement aux chercheurs une interface pour manipuler leurs données.

Figure 4 : Keshif navigateur à facettes

Les possibilités de Keshif sont très étendues, toutefois elles ne sont accessibles qu'à partir du moment où les données sont parfaitement structurées. De plus, l'utilisation de cette librairie nécessite l'intervention d'un développeur Web pour modéliser les différentes facettes.

[3] url vers le site de l'université : http://hdlab.stanford.edu/tools/
[4] url vers la librairie : http://keshif.me/

III. PROTOTYPE POUR LA MODELISATION DES RESEAUX D'INFLUENCES

La présentation aux chercheurs des outils que nous avons testés, nous a permis d'attirer leur attention sur le vocabulaire à utiliser pour catégoriser des influences. Nous avons insisté sur la nécessité de rendre interopérable ces catégories notamment en terme de synchronisation spatio-temporelle et de référence à des personnages. Enfin, nous les avons sensibilisés à l'importance de préserver et partager les points de vue de chacun pour faire émerger un consensus à partir des controverses. C'est dans ces perspectives que nous avons développé un prototype pour la modélisation des réseaux d'influences.

A. Catégorisation des rapports

L'utilisation des outils précédents nécessite des données qui soient « propres » notamment pour éviter des doublons ou des mauvais formats de date, de lieux ou d'orthographe des personnes. Pour mener à bien ce travail, une première étape consiste à récolter les notions et concepts utiles pour les chercheurs afin d'avoir une première estimation d'un vocabulaire commun. Via un formulaire Google[5], les chercheurs ont pu choisir les catégories qui leur semblaient importantes pour décrire les relations d'influences entre biologie et littérature.

Figure 5 : Formulaire de catégorisation des rapports

[5] url vers le formulaire : https://docs.google.com/forms/d/1y-KWcgbQczgpDAFhdZSaHblqdbZqTLsCmjbRM_UHTXQ/viewform

Les données récoltées[6] ont permis de représenter les différents profils de chercheur suivant le nombre et la qualités des choix de catégorisation :

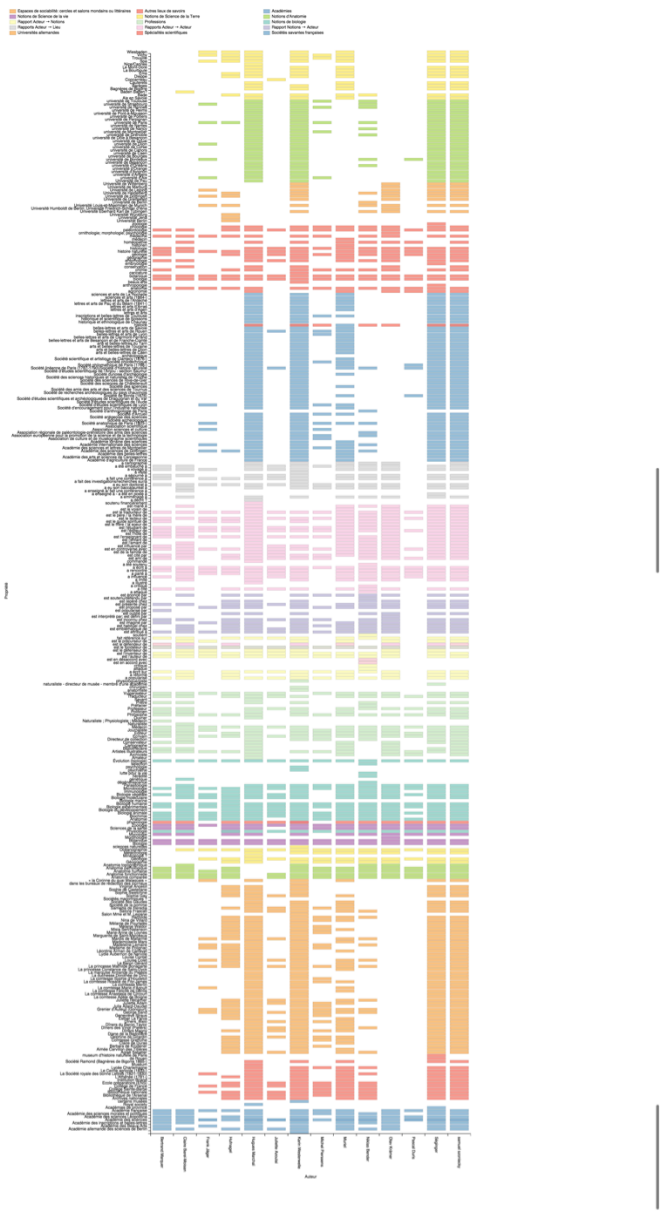

Figure 6 : comparaison des choix de catégorisation entre chercheur[7]

Cette représentation montre globalement qu'il y a plusieurs points de vue sur les catégories utiles pour modéliser les influences entre littérature et biologie. L'analyse précise de ces différents points de vue reste à mener mais ce premier résultat nous encourage à persévérer dans notre projet de concevoir un outil permettant aux chercheurs d'exprimer leurs points de vue. En effet, cette représentation donne aux chercheurs la possibilité de comparer les catégories qu'ils

[6] url vers les données récoltées : https://docs.google.com/spreadsheets/d/169QusRs3TyrqxzVb28hXX2E8FZsqWb01xIWpl5FDPns/pubhtml?gid=1344191823&single=true

[7] url de la représentation des choix de catégorisation : http://gapai.univ-paris8.fr/jdc/public/biolographes/comparecat

utilisent pour exprimer leurs points de vue, de les diffuser auprès des autres chercheurs et ainsi de faire émerger le processus de réflexivité nécessaire à l'élaboration d'un consensus scientifique.

B. Saisie assistée de réseaux d'influences

A partir de cette catégorisation collective, nous avons développé un prototype d'application Web pour la modélisation des influences. Cette application est conçue pour respecter et utiliser les standards du Web sémantique (dbpedia, databnf…) et ainsi facilité l'interopérabilité des données récoltées. Un accent particulier est mis sur l'enregistrement des différents points de vue des chercheurs. L'objectif n'est pas de simuler automatiquement par des algorithmes des points de vue sur les documents analysés comme peuvent le faire les outils de Traitement Automatique de la Langue, mais plutôt de fournir aux chercheurs des outils pour exprimer leurs point de vue sur les documents et le rendre interopérables avec d'autres points de vue.

Une première version de cet outil utilise les choix de catégorisation fait par les chercheurs pour proposer des « cribles » sous forme de menu déroulant proposant uniquement les catégories choisi par le chercheur. A partir de ces cribles, l'utilisateur peut construire un réseau d'influences suivant les critères génériques que nous avons défini pour qu'ils soient interopérables avec tous domaines de connaissances. Suivant ce principe nous définition la modélisation d'un réseau d'influence comme la création de relations spatialisées et temporalisées entre des documents, des acteurs et des concepts. En ce sens, nous nous rapprochons des travaux de Citton sur l'économie des affects :

« Cartographier les courants d'imitation et les flux d'énergie qui dirigent ces contagions et ces "cascades de magnétisations successives enchaînées", expliquer leurs modes de production, comprendre leurs régimes d'échanges et de transformations, voilà bien l'entreprise que nous proposent ensemble Spinoza et Tarde. » [4, p. 96].

La construction de cette cartographie se fait graphiquement grâce à des formulaires qui aident le chercheur à formaliser ces propositions pour qu'elles soient le plus interopérable possible. Par exemple, le formulaire de saisie des acteurs permet de qualifier l'acteur en retrouvant son n° ISNI[8] dans la base de données Data BNF grâce à une requête SPARQL.

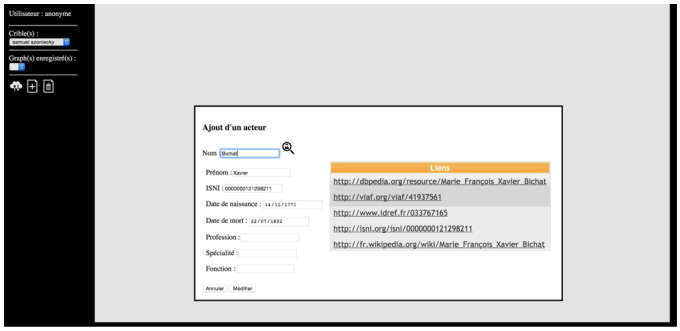

Figure 7 : éditeur de réseaux d'influences

L'outil est toujours en cours de développement et en phase de test mais une vidéo est disponible sur Youtube[9] pour montrer son usage. Nous prévoyons d'optimiser les fonctions d'édition et finaliser la gestion collaborative des réseaux (enregistrement, diffusion, annotation, …). Des tests utilisateurs doivent aussi être menés pour valider l'ergonomie des interfaces et optimiser la saisie des points de vue. L'outil sera ensuite disponible gratuitement et en Open Source pour modéliser tout type de réseaux d'influences.

IV. LIMITES ET PERSPECTIVES

La mise à disposition de nouveaux outils comme ceux que nous avons testés ou celui que nous sommes en train de développer, ne résout pas la question de la pertinence du numérique dans les recherches en sciences humaines et surtout les conséquences épistémologiques de leurs usages.

A. Exemples de limites

Les expériences que nous menons dans le cadre du projet ANR Biolographes montrent qu'il existe un fossé énorme entre les « informaticiens » et les « humanistes ». Les premiers ayant tendances à vouloir appliquer les solutions et les méthodes qui mettent en avant l'efficacité des machines, les seconds se laissant convaincre par les premiers que les machines leur feront « gagner » du temps dans l'exploration de leur corpus et « découvrir » les preuves qui confirmeront leurs hypothèses. Deux anecdotes qui se sont produites lors des réunions, nous semblent significatives de cette ambivalence entre le pouvoir d'agir de la machine piloté par l'informaticien et celui du chercheur en sciences humaines.

La première concerne la présentation d'un chercheur en histoire des sciences qui expliquait pourquoi il était très pertinent que la première traduction de Darwin en français soit l'œuvre d'une femme. L'intuition de ce chercheur l'avait fait tout de suite s'intéresser à cette caractéristique du contexte d'édition car sa connaissance intime de l'époque lui indiquait qu'il y avait là un détail particulièrement signifiant dans un monde où cela n'était pas l'habitude. Comment une machine pourrait-elle extraire d'une analyse automatique du corpus de Darwin cette information ? Qui penserait à programmer un algorithme pour comparer le sexe des traducteurs à différentes époques ? Nous nous confrontons ici à une des limites épistémologiques du pouvoir d'agir des machines que Bertin avait décelé dans sa sémiologie graphique et qui sont tout aussi valides dans le domaine des humanités numériques :

« Quel ordinateur nous dira qu'il lui manque tel algorithme ? […]
Quel ordinateur nous dira qui lui manque telles "données" ? Ces deux questions impliquent que nous fassions appel à des éléments extérieurs, qui sont nos connaissances et notre intuition... » [3, p. VII]

Face à cette limite, il est fondamental de placer au cœur de l'outillage des humanités numériques, l'expression et le partage de l'intuition des chercheurs car c'est elle qui

[8] définition du n° ISNI http://www.bnf.fr/fr/professionnels/isni_informer.html

[9] url vers la vidéo de présentation de l'outil : https://youtu.be/mrfkfzSULHU

différencie le pourvoir d'agir de l'humain et celui de la machine [15].

La deuxième anecdote s'est produite lorsque en réponse aux remarques récurrentes sur l'ampleur du travail de dépouillement qui était demandait aux chercheurs, il a été suggéré de faire appel à des étudiants pour récolter les informations de localisation, de datation et de personnes. La réponse de l'assemblée a été négative quasi-unanimement. Cette réaction est symptomatique d'une vision de la recherche qui ne prend pas en compte ce que l'organisation informatique de l'intelligence collective peut apporter dans un travail qui demande l'exploration fine d'un corpus de très grande taille. Pourquoi ne pas observer les corpus textuels comme le font les milliers de bénévoles qui recensent la biodiversité des jardins [jardin] ? Il y a sans doute des protocoles de recherche à inventer et à mettre en place pour organiser un travail collaboratif d'exploration et de cartographie des grands corpus de sciences humaines. Mais le monde universitaire est-il prêt à faire évoluer les systèmes de travail comme les y invitent Juanals et Noyer ?

« Ces systèmes de travail et de fonctionnement intègrent de nouvelles procédures normalisées de système d'écriture et de représentation performants. Ils supposent l'apprentissage et l'adoption de normes, de programmes, de routines, de dispositifs d'écriture et d'interfaces à la plasticité très grande. Ils supposent que soient développés et appropriés des modes de représentation et de navigation dans des espaces-temps coopératifs complexes et distribués. » [9, p. 38]

Limites mécaniques, limites humaines, les humanités numériques sont sans doute le terrain privilégié pour expérimenter de nouveaux modes d'existences [10] qui harmonisent au mieux le pouvoir d'agir des machines et des humains.

B. Exemple d'une perspective épistémologique

Faces aux limites dont nous venons de présenter deux exemples, nous envisageons de poursuivre nos recherches dans deux directions.

La première que nous explorons consiste à développer une méthode générique pour l'analyse des écosystèmes d'informations numériques qui composent un univers complexe [12] de relations à la fois complémentaires et antagonistes entre une multitudes d'existences humaines, mécaniques, institutionnelles… L'objectif est de décrire très précisément les existences pour que les chercheurs trouvent rapidement les analyses qui portent sur des existences similaires et puissent comparer les différences d'interprétations.

Pour parvenir aux analyses croisées de ces écosystèmes, nous modélisons les existences numériques en croisant les principes ontologiques de Spinoza [5] avec ceux de Descola [6]. En ce qui concerne Spinoza, nous reprenons les 3 dimensions d'existence (parties extensives, rapports, essences) corrélées à 3 genres de connaissance (chocs, logiques, intuitions). Chez Descola, nous utilisons les matrices ontologiques qui caractérises les rapports entre physicalités et intériorités. Plus particulièrement, nous nous focalisons sur l'ontologie analogiste qui correspond tout à fait au numérique dans sa capacité de transformation illimitée des physicalités, des intériorités et de leurs rapports :

« Cette lutte continuée entre un océan vertigineux et des réseaux de relations toujours en train de multiplier leurs connexions définit en rigueur l'analogisme, mot qui résume et peint à merveille notre monde objectif, nos travaux cognitifs, nos rêves subjectifs ainsi que les collectifs qui naissent aujourd'hui et feront la politique du futur. » [13, p. 85]

A l'aide de ces principes nous composons des représentations uniques que nous qualifions de monades. Elles se composent de quatre ensembles : les documents, les acteurs, les concepts et les rapports. A l'intérieur de chaque ensemble, les éléments entretiennent des relations de *sémantique différentielle* [2, p. 142] suivant la position relative d'un élément par rapport à deux axes, celui du père et du frère dans un arbre.

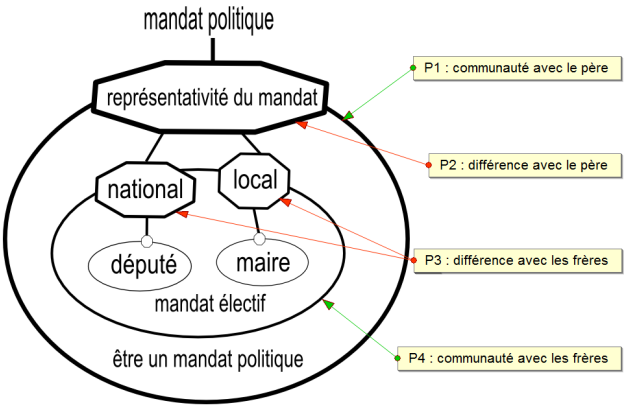

Figure 8 : Principes de sémantique différentielle

Les niveaux d'échelle définis par la position des éléments dans l'arbre des hiérarchies père-fils mis en rapport avec le nombre d'élément dans chacun des ensembles donne une métrique précise de la monade. Cette métrique permet de connaître le niveau de complexité d'une existence afin de pouvoir comparer automatiquement des interprétations qui portent sur les mêmes documents, les mêmes acteurs ou les mêmes concepts. Nous nommons cette métrique Indice de Complexité Existentiel (ICE) et nous développons un outil pour calculer automatiquement cet indice à partir de la modélisation d'une existence.

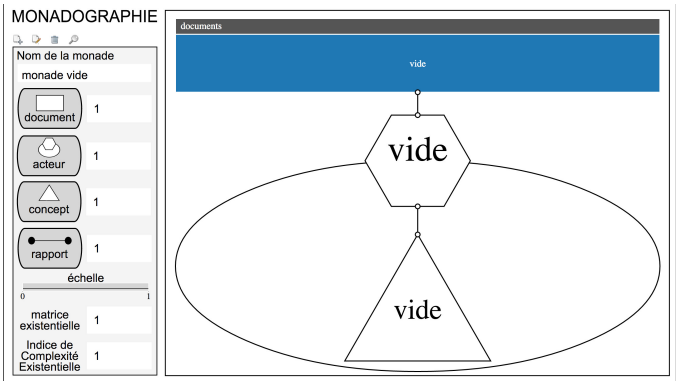

Figure 9 : modélisation d'une existence vide

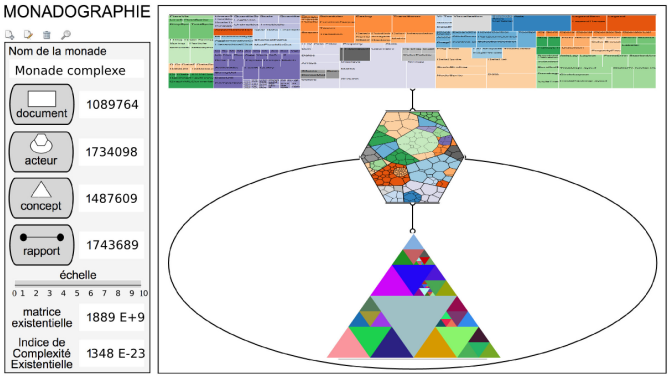

Figure 10 : modélisation d'une existence complexe

Il est trop tôt pour savoir si ce type de modélisation peut intéresser les chercheurs en humanités numériques et surtout si cette communauté acceptera de formaliser précisément ses recherches. Quoi qu'il en soit, vu la multiplication des productions scientifiques, il sera de toute façon nécessaire de trouver les moyens d'organiser ces flux de connaissances pour qu'ils profitent à d'autres que ceux qui en sont à la source.

V. Conclusion

Le domaine des humanités numériques est en pleine expansion tant par le nombre de chercheur et d'institution qui s'y intéresse que par l'ampleur des travaux qui sont à mener. Cet engouement ne doit pas masquer les questions que pose l'usage des technologies numériques dans les sciences humaines. A travers nos expériences sur les outils de cartographie des réseaux d'influences, nous en avons décelé un manque de culture dans le domaine de l'épistémologie de l'information notamment en ce qui concerne les limites de la calculabilité et son dépassement par l'intelligence collective. Pour beaucoup de chercheur en sciences humaines, les outils informatiques sont vus comme des boites noires dont on ne connait pas ou pire on ne veut pas connaître le fonctionnement. Il en ressort une incompréhension sur la nécessité de la formalisation et sur le manque de pertinence des résultats qui en découle. Du coté des informaticiens, l'accent est mis sur les performances algorithmiques des machines au point d'en oublier parfois l'expertise des chercheurs.

Il faudra sans doute du temps pour harmoniser les pratiques et parvenir à une culture commune des bonnes pratiques dans le domaine des humanités numériques. Espérons que la mis en place en France d'un baccalauréat humanités numériques sera l'occasion de faire progresser la science dans ce domaine.

References

[1] A. Angjeli, "ISNI : un identifiant passerelle," *Documentation et Bibliotheques*, vol. 58, no. 3, pp. 101–108, Sep. 2012.

[2] B. Bachimont, *Ingénierie des connaissances et des contenus : Le numérique entre ontologies et documents*. Paris: Hermes science publications, 2007.

[3] J. Bertin, *Sémiologie graphique les diagrammes, les réseaux, les cartes*. Paris: Ed. de l'EHESS, 1999.

[4] Y. Citton, "Les lois de l'imitation des affects," in *Spinoza et les sciences sociales : de la puissance de la multitude à l'économie des affects*, Paris: Éd. Amsterdam, 2008.

[5] G. Deleuze, *Spinoza et le problème de l'expression*. Paris: Éditions de Minuit, 1968.

[6] P. Descola, *Par-delà nature et culture*. Paris: NRF : Gallimard, 2005.

[7] K. Gallant, E. Lorang, and A. Ramirez, "Tools for the digital humanities: a librarian's guide," 2014.

[8] H. Hampartzoumian, R.-L. Preud'Homme, G. Loïs, R. Raymond, È. A. Bühler, and Y. Hanachi, "L'Observatoire agricole de la biodiversité (OAB) : une pédagogie active autour d'un projet de sciences participatives," *Pour*, vol. N° 219, no. 3, pp. 169–180, Nov. 2013.

[9] B. Juanals and J.-M. Noyer, "De l'émergence de nouvelles technologies intellectuelles," in *Technologies de l'information et intelligences collectives*, Hermes Science Publications, 2010.

[10] B. Latour, *Enquêtes sur les modes d'existence : Une anthropologie des Modernes*. Editions La Découverte, 2012.

[11] P. Lévy, "De l'émergence de nouvelles technologies intellectuelles," in *Technologies de l'information et intelligences collectives*, Hermes Science Publications, 2010.

[12] E. Morin, *La Méthode, tome 4 : Les Idées*. Seuil, 1995.

[13] M. Serres, *Ecrivains, savants et philosophes font le tour du monde*. Paris: Pommier, 2009.

[14] S. Szoniecky and H. Hachour, "Monades pour une éthique des écosystèmes d'information numériques," presented at the Digital Intelligence, Nantes, 2014.

[15] S. Szoniecky, "Le langage du Web du symbolique à l'allégorique, vers une représentation de la connaissance en train de se faire," in *ISKO - Magreb 2011*, Hammamet, Tunisie, 2011.