Humanités numériques

Serge Abiteboul
INRIA & ENS Cachan

Florence Hachez-Leroy
Université d'Artois & CRH (CNRS / EHESS)

> *The digital humanities, also known as humanities computing[1,2], is a field of study, research, teaching, and invention concerned with the intersection of computing and the disciplines of the humanities. It is methodological by nature and interdisciplinary in scope. It involves investigation, analysis, synthesis and presentation of information in electronic form. It studies how these media affect the disciplines in which they are used, and what these disciplines have to contribute to our knowledge of computing.* Wikipedia, http://en.wikipedia.org, 2014.

Introduction

La transformation des humanités par le numérique. L'expression « humanité numérique[3] » est aujourd'hui moquée au motif que l'association entre une technologie et les sciences humaines et sociales tiendrait de l'oxymore. Or, si l'on compare avec l'imprimerie, celle-ci n'a-t-elle pas été aussi une technologie dont l'apport transforma en profondeur les humanités d'alors ? Une autre critique plus incisive considère que l'adjonction de l'adjectif « numérique » relèverait de la tautologie. Dans un monde numérique, les humanités pourraient-elles éviter de suivre le mouvement et devenir elles-aussi numériques ? Peut-on par exemple imaginer aujourd'hui la géographie sans les systèmes d'information géographique ou la linguistique sans le traitement automatique de la langue ? Cela n'est pas plus censé que d'imaginer, par exemple, l'astronomie sans ses pipelines de calculs informatiques, ou la génomique sans ses algorithmes d'analyse de séquences ADN. Nous vivons donc bien une transition des humanités. En même temps, il existe un risque qu'à être partout, le numérique ne soit singulièrement nulle part, qu'il manque un lieu de *coordination* des efforts autour du numérique. Alors pendant cette transition au moins, il nous semble intéressant d'examiner spécifiquement les aspects numériques de cette transition.

Un chercheur en humanités aujourd'hui consulte des documents sur Internet, classe ses informations dans des bases de données, extrait des informations dans des bases de

[1] John Unsworth, « What is Humanities Computing and What is Not? », *in* Melissa Terras, Julianne Nyhan and Edward Vanhoutte (ed.), *Defining Digital Humanities: A Reader,* Londres, Ashgate, 2013.
[2] Willard McCarty, *Humanities Computing*, New York, Palgrave MacMillan, 2014.
[3] Marin Dacos et Pierre Mounier, *Humanités numériques : État des lieux et positionnement de la recherche française dans le contexte international*, Institut Français, 2014

données de plus en plus riches et nombreuses, tweete, blog, chate à distance avec ses collègues, ses étudiants, etc. Est-ce que les Humanités numériques se résument à cela ? Certainement pas, mais il n'est pas simple d'en donner une définition. Nous allons plutôt essayer de répondre à la question : en quoi est-ce que l'informatique et le numérique transforment les humanités ? Walter Benjamin a expliqué que la photographie a transformé l'art. De la même façon, nous allons chercher à mettre en évidence dans cet article comment le numérique et fondamentalement la science et la technique informatique ont transformé en profondeur les sciences humaines et sociales.

Une convergence entre sciences et humanités. Les humanités numériques sont un lieu privilégié de convergence entre sciences et humanités. La dissociation entre sciences et humanités s'est développée progressivement aux 19e et 20e siècles en particulier du fait de la spécialisation des recherches et du changement de regard sur les disciplines littéraires jugées peu compatible avec la « modernité ». L'enseignement a suivi au point que, par exemple, en France, on demande aux lycéens de première de choisir entre sciences et lettres. Un tel choix n'a pas véritablement de sens et il s'avère tout particulièrement absurde pour un élève qui serait intéressé par les humanités numériques. Un ordinateur est une machine à tout faire (« *general purpose* ») : le même système, par exemple un système de gestion de bases de données, peut être utilisé que la science soit « humaine » ou pas, et le même algorithme, par exemple un algorithme de recommandation, peut être utilisé dans les deux cas. Un historien ou un sociologue est aussi à même de saisir la poésie d'un code informatique qu'un physicien ou un biologiste. Les méthodes, les concepts, les techniques, les outils de l'informatique rapprochent les chercheurs de toutes ces disciplines. Les humanités sont donc en train de se réinventer par l'informatique et de se rapprocher des autres sciences, nous semble-t-il. Ces convergences sont si fortes que plutôt que de parler des humanités numériques, peut-être aurait-il fallu discourir de « sciences numériques » en général.

Organisation. Dans un premier temps, il est utile de délimiter le territoire et de préciser en particulier le sens d' « humanités » et de « numériques ». Nous examinerons ensuite divers aspects des humanités numériques. Celles-ci sont déjà, malgré leur jeunesse, d'une telle richesse que chaque choix d'une illustration en écartera de nombreuses autres tout aussi intéressantes. Evidemment, ces choix seront biaisés, discutables. Notre propos n'est pas d'établir un classement mais uniquement d'illustrer nos propos. Organiser les facettes du monde des humanités numériques n'est pas chose facile. D'une certaine façon, nous allons passer en revue des classes d'outils, mais ces outils existent par les techniques, les méthodes, les usages qu'ils impliquent. Mais considérons d'abord le territoire.

Le territoire

Les humanités. Le terme est imprécis. En anglais, par exemple, le terme « *humanities* » ne couvre pas les sciences sociales. Pour nous le domaine des humanités numériques est à prendre dans un sens très général, incluant l'histoire, la linguistique, la littérature, les arts et le design, mais aussi la géographie, l'économie, la sociologie, le droit, la théologie et les sciences des religions, etc.

D'un coté, les humanités, de l'autre, l'informatique et le numérique.

Le numérique et l'informatique

> *L'informatique est la science et la technique de la représentation de l'information d'origine artificielle ou naturelle, ainsi que des processus algorithmiques de collecte, stockage, analyse, transformation, communication et exploitation de cette information, exprimés dans des langages formels ou des langues naturelles et effectués par des machines ou des êtres humains, seuls ou collectivement*. L'informatique : la science au cœur du numérique[4], Société Informatique de France, 2014.

Il nous faut ici considérer l'articulation entre le monde numérique et la science qui en est au cœur, l'informatique[4,5]. Dans un premier temps, on a parlé de « *humanities computing* », en mettant l'accent sur les aspects tenant de l'informatique. Puis la focale a été élargie en préférant l'adjectif « *digital* » (en français, numérique) qui qualifie les activités s'appuyant sur la numérisation de l'information. Si l'informatique a rendu possible les humanités numériques, se limiter à cette science, à cette technique, serait réducteur. Par exemple, le Web qui a une place si essentielle dans les humanités numériques, au delà de ses aspects purement techniques, tient d'une philosophie de la mise à disposition pour tous, du partage. Cet aspect comme bien d'autres de la culture numérique font partie intégrante du cadre des humanités numériques. Au-delà de l'informatique, nous considérerons le numérique, ses usages, le monde qu'il construit, ses cultures.

Les humanités numériques se situent donc quelque part aux points de rencontre d'une part, des humanités et de l'autre, de l'informatique et du numérique. On y inclut typiquement de nombreux aspects comme, l'enseignement, la création artistique ou la recherche. Nous concentrerons notre propos sur cette dernière et considérerons comment

> *la recherche dans les sciences humaines et sociales évolue dans sa rencontre avec la science et la technique informatique, le monde et la culture numérique.*

Notre propos (et c'est un choix arbitraire) porte sur les humanités numériques uniquement lorsqu'un processus de recherche scientifique est impliqué. Par exemple, la conception d'un cours de grec ancien[6] par un Flot (Mooc en anglais) n'entre pas dans notre panel, pas plus que la conception d'une visite de musée à partir d'un logiciel spécialisé. En revanche, un chercheur intéressé par les flots ou la visite numérisée de musées et un ingénieur concevant un logiciel pour ce faire entrent dans notre champ.

Nous ne traiterons pas non plus de l'enseignement des humanités à l'heure du numérique. Evidemment, l'éducation change. Et il ne s'agit pas seulement d'utiliser des outils numériques pour enseigner, pas seulement d'enseigner autrement, mais également de faire évoluer les contenus aussi bien d'ailleurs dans les matières scientifiques que dans les humanités. Cet aspect est intimement lié au sujet de cet article, mais c'est une autre thématique. Et tout aussi arbitrairement encore, nous excluons d'autres aspects, comme la création d'art numérique ou les interactions entre les machines et les personnes, bien que certains les considèrent comme partie des humanités numériques.

En dépit de ces restrictions, le territoire reste vaste. Depuis les tout débuts de l'informatique, des chercheurs et des ingénieurs de laboratoires SHS et informatique se sont intéressés aux humanités numériques. Ils ont considéré des modèles quantitatifs, par exemple pour des analyses statistiques de texte s'appuyant sur la linguistique computationnelle. Ils sont passés au qualitatif avec des initiatives comme le « Text encoding

---

[4] Binaire. http://binaire.blog.lemonde.fr/informatique-quesaco/

[5] Milad Doueihi, *La grande conversion numérique*, Paris, Seuil, 2011.

[6] Les flots Sillages. http://flot.sillages.info/?portfolio=grec-ancien-debutants

initiative[7] » en développant des outils informatiques qui faisaient véritablement évoluer la représentation et l'échange de connaissances. Ils ont développé de nouveaux outils, de nouvelles techniques, conçu de nouvelles façons de faire la recherche. Ils ont bâti une discipline avec ses textes de référence (comme *Une introduction aux humanités numériques*[8]), ses revues (comme le *Journal of Digital Humanities*), ses sociétés savantes (comme l'*Alliance of Digital Humanities Organizations*), ses conférences (comme *Digital Humanities*), ses instituts (comme l'*Institut des Humanités Digitales de Bordeaux*), ses séminaires (comme « *Digital Humanities. Les transformations numériques du rapport aux savoirs* » à l'EHESS), ses financements (comme l'Axe 6 du Défi 8 du dernier appel ANR « *Révolution numérique et mutations sociales* ») et des chercheurs et enseignants-chercheurs qui se réclament spécialistes des humanités numériques. C'est de ce champ que nous allons essayer de dresser un rapide panorama.

La transformation des sciences par la pensée informatique

Au fil des recherches en humanités numériques, ce qui nous paraît le plus passionnant, c'est une remise en question radicale des SHS, de l'essence de leurs pratiques, une entreprise de transformation fondamentale de ces sciences par la pensée informatique. C'est ce que nous aurions aimé placer au cœur de cet article. Mais nous manquons sans doute encore de recul. Ces transformations sont encore pour une grande part en devenir. Si ces considérations sous-tendent tout notre article et plus spécifiquement les sections sur les connaissances, ou la simulation, l'accent est aussi mis sur les nouveaux outils et les nouvelles pratiques.

Les bases : numérisation, bases de données, hypertexte et Internet

Le point de départ des humanités numériques est la représentation de connaissances sous forme numérique. La numérisation des textes a ainsi énormément bénéficié des outils d'OCR (« optical character recognition »). Mais la numérisation touche bien plus que le texte : les images fixes et animées, la musique, l'architecture, etc. Si quelques bastions résistent un peu comme l'odorat, le champ du numérique ne cesse de s'étendre.

L'objet numérisé apporte son lot de problèmes. Comment garantir des propriétés comme l'authenticité, l'intégrité, l'identité, la pérennité ? Ces problèmes ne sont pas nouveaux (les faussaires, les manipulateurs de documents, les tricheurs en tout genre n'ont pas attendu le numérique pour sévir) mais se posent avec plus d'acuité du fait que le numérique permet la reproduction du même objet quasi sans coût (on parle de biens non rivaux[9]). Pourtant, les techniques apportent aussi des solutions qu'il reste à développer, à déployer. Pour ne prendre qu'un exemple l'authenticité. En utilisant des algorithmes cryptographiques comme RSA[10], il est possible d'accompagner un texte d'une signature qui l'authentifie bien mieux que l'on ne pourrait authentifier quelque objet physique que ce soit. Reste que de telles signatures sont de nos jours encore trop peu

[7] Text Encoding Initiative. http://www.tei-c.org/

[8] Pierre Mounier (dir.), *Read/Write Book 2 : une introduction aux humanités numériques*, Marseille, OpenEdition Press, 2012 (généré le 26 octobre 2014). Disponible sur Internet : <http://books.openedition.org/oep/226>. ISBN : 9782821813250.

[9] Lawrence Lessig, *L'avenir des idées*, Lyon, Presses Universitaires de Lyon, 2005.

[10] Ron Rivest, Adi Shamir, Len Adleman, « Method for Obtaining Digital Signatures and Public-Key Cryptosystems », *Communications of the ACM 21* (2), 1978, p. 120–126.

utilisées. Reste aussi que le numérique accentue l'affaiblissement de la notion d'auteur, et la rigidité de la notion d'œuvre.

Le texte fut l'application liminaire de la numérisation. Une fois numérisé, il pouvait être indexé automatiquement. Une des premières utilisations massives de l'informatique dans les humanités numériques a été la gestion informatisée des catalogues de bibliothèques avec des index réalisés totalement manuellement dans les premiers temps. Les bibliothécaires voyaient leur monde évoluer sans peut-être imaginer qu'après les catalogues, les livres et les bibliothèques pouvaient un jour devenir numériques.

Au-delà du simple texte linéaire et des index, l'informatique offre le moyen de gérer en général des données et des informations. Les structures naturelles pour organiser l'information en mathématiques et en informatique sont les tableaux, les arbres et les graphes. Nous allons naturellement rencontrer ces trois types de structures : le tableau est la structure la plus utilisée dans les bases de données, l'arbre est la structure interne des documents HTML ou XML, et nous retrouvons les graphes dans les documents hypertextes.

> *Un système de gestion de bases de données sert de médiateur entre des individus et des machines. Dans le modèle relationnel proposé par Ted Codd dans les années 1970, les données sont organisées en tableaux à deux dimensions qu'on appelle des relations. Les données sont interrogées en utilisant des « requêtes » dans le langage SQL (inspiré de la Logique du premier ordre). Ces requêtes sont évaluées par le système en s'appuyant sur l'« algèbre relationnelle ».* Serge Abiteboul, Leçon inaugurale[11] du Collège de France, 2012.

Les chercheurs de SHS ont vite compris l'intérêt de réunir des données, de les organiser dans des bases de données notamment relationnelles. La production d'une base de données peut d'ailleurs être aujourd'hui l'objet d'un sujet de recherche au même titre que l'écriture d'un article dans la meilleure revue d'un domaine. La même situation existe dans d'autres domaines scientifiques notamment la biologie. Par exemple, le point de départ de la base de données de séquences protéiques SwissProt est le doctorat d'Amos Bairoch en 1986. A partir de données entrées manuellement, Swiss-Prot donne des informations extrêmement détaillées sur des séquences de protéines. Cette base de données est une contribution scientifique essentielle. De même en SHS, l'Internet Movie Database (IMDb) est une gigantesque base de données en ligne sur le cinéma mondial (films, acteurs, réalisateurs, etc.). IMDb a même été en 2011 parmi les 40 sites les plus visités du Web.

Après la base de données, un autre concept informatique essentiel pour les SHS est l' « hypertexte ».

> *Un système hypertexte est un système contenant des nœuds liés entre eux par des hyperliens permettant de passer automatiquement d'un nœud à un autre. Un document hypertexte est donc un document qui contient des hyperliens et des nœuds. Un nœud est une « unité minimale d'informations », notion assez floue qui signifie simplement que l'information d'un nœud sera toujours présentée entière. Les liens entre les parties du texte sont gérés par ordinateur et permettent d'accéder à l'information d'une*

[11] Serge Abiteboul, *Sciences des données : de la logique du premier ordre à la Toile*, Paris, Fayard, 2012.

manière associative ou, tout au moins, d'une façon de naviguer personnalisée, de manière non linéaire, au gré de l'utilisateur. Wikipédia, français, 2014.

Il n'est sans doute pas besoin d'expliquer plus loin la notion d'hypertexte, le lecteur ayant déjà l'expérience de navigation sur le Web, qui illustre la réussite planétaire des hypertextes. Les ingrédients du Web sont des plus simples : un réseau de machines, Internet, et un protocole d'hypertextes pour accéder à un réseau de textes disponibles sur ce réseau de machines, le World Wide Web.

You affect the world by what you browse. Tim Berners-Lee

Bases de données et hypertextes se combinent à une troisième grande réalisation de l'informatique, Internet, pour former un des plus beaux succès des humanités numériques : les « bibliothèques numériques ». Par exemple, le Projet Perseus[12] de l'université Tufts s'est attaqué à la construction d'une bibliothèque numérique qui rassemble des textes du monde méditerranéen en particulier en grec, latin et arabe. Les textes numérisés, indexés, disponibles sur la Toile, sont facilement accessibles à tous. Son ambition affichée est : « *to help make the full record for humanity as intellectually accessible as possible to every human being, providing information adapted to as many linguistic and cultural backgrounds as possible*. »

Voilà un cas exemplaire de contribution aux humanités.

Cette diffusion s'accompagne du développement d'interfaces qui vont au-delà de l'accès classique aux connaissances par le texte. C'est le cas par exemple pour ce qui est de l'architecture[13] avec Usine 3D. Le « visiteur » peut découvrir la reconstitution du premier atelier automobile au début du 20e siècle, avec des chaines de montage. Il s'agit d'une véritable expérience qui va plus loin que la consultation de nombreux ouvrages et plans d'architectes. Photographies, images figurées, textes, plans, commentaires des historiens spécialistes, etc. sont analysés, les informations recoupées pour former un corpus hétérogène qui concoure à cette reconstitution scientifiquement encadrée.

La transformation radicale de l'accès à l'information fait qu'à l'heure du Web, des étudiants, mais aussi des amateurs, des journalistes, tous les citoyens ont accès à des sources considérables. Chacun peut s'improviser chercheur ou journaliste et contribuer à l'enrichissement des connaissances en rédigeant par exemple les articles d'une encyclopédie, domaine réservé jusque là aux chercheurs ou aux érudits. Observer les oiseaux et alimenter ensuite une base de données scientifique par ses informations est désormais à la portée de tous.

Le partage : interaction et communication

Depuis des siècles, le travail des chercheurs s'est appuyé sur la notion de réseau. On échangeait des lettres. On voyageait pour consulter une bibliothèque ; on en profitait pour rencontrer ses homologues locaux. Ces échanges, ces rencontres physiques participaient à élaborer et à enrichir les connaissances, à construire des réseaux.

Après Internet et le réseau des machines, après le Web des débuts et le réseau de contenus, le Web 2.0 s'est proposé pour faciliter les communications entre individus, enrichir

[12] Perseus Digital Library Project, http://www.perseus.tufts.edu/hopper/
[13] Usine 3D, http://www.usines3d.fr/

les interactions entre eux, réinventer les réseaux sociaux. Pour les scientifiques, cela permettait de redéfinir le travailler ensemble. On pouvait partager des textes, les annoter ensemble, les commenter, voire co-rédiger des contenus très riches en s'éloignant du texte linéaire bien défini aux auteurs bien précisés.

Les exemples fourmillent. Le réseau peut être par exemple massif et à vocation généraliste comme ResearchGate avec plus d'un million de membres ou dédié à un partage entre les chercheurs et le grand public comme Nature Network de la revue Nature. Pour citer un exemple riche en symbole, le projet[14] « Mapping the republic of letters » lancé par Stanford a permis de mettre en commun des recherches pour étudier comment, depuis la Renaissance, les lettrés européens partageaient leurs connaissances à travers des textes et des rencontres. Un réseau social numérique pour expliquer un réseau social « classique » !

Un scientifique comme Claude Bernard notait ses expériences dans des carnets[15] dont la consultation en ligne est une expérience chargée d'émotions. Que sont de tels carnets aujourd'hui ? Des notes partagées avec des collaborateurs sur Dropbox ? Des bases de données en ligne ? Des articles dans ResearchBlogging ? Des workflows d'expériences sur des sites MyExperiment ? Des carnets de recherche en ligne comme dans Hypotheses.org ?

En utilisant un jeu vidéo, Foldit, des internautes sont arrivés à décoder la structure d'une enzyme proche de celle du virus du sida[16]. Ils ont compris ce qui bloquait experts et ordinateurs, comment cette enzyme se repliait dans un espace en trois dimensions pour construire sa structure. Le jeu se marie ici au réseau, dans le plus pur esprit des réseaux sociaux. Si un individu a exposé le pliage, les scientifiques qui ont imaginé le jeu, les programmeurs qui l'ont implémenté, tous les joueurs qui ont essayé les pliages qui ne marchaient pas, tous ont contribué.

Les chercheurs en humanités n'ont pas attendu le Web pour se rencontrer, communiquer entre eux, collaborer. Il n'empêche que c'est seul devant la page blanche que s'élaborait le plus souvent une œuvre. Si cela reste en partie vrai, le travail est devenu typiquement plus collaboratif, les œuvres plus collectives. Le passage au travail en réseau s'accompagne de changements fondamentaux dans nos rapports aux connaissances. Un univers des fragments se substitue aux contributions monolithes. Les outils de recherche, les sites de co-rédaction encouragent cet effet, qui s'accompagne aussi de l'affaiblissement de la contribution de l'auteur individuel devant les contributions du groupe. Dans le cadre des revues académiques, s'étend l' « open access », c'est-à-dire la mise à disposition en ligne de contenus numériques. Et au delà, avec le « creative commons » se développe une approche globale pour libérer les « œuvres » des droits de propriété.

La constitution de réseaux devient à son tour objet d'études. La théorie des réseaux s'applique en sociologie[17] aussi bien que pour les réseaux électriques.

---

[14] Mapping the Republic of letters. http://republicofletters.stanford.edu/
[15] La Salamandre. https://salamandre.college-de-france.fr
[16] Seth Cooper, Firas Khatib, Adrien Treuille et al., « Predicting protein structures with a multiplayer online game », *Nature*, 2010, vol. 466, n° 7307, p. 756-760.
[17] Duncan Watts J., *Six degrees: The science of a connected age*, New York, W. W. Norton & Company, 2004.

La connaissance et le Web sémantique

L'informatique permet de traiter des données, de l'information, *des connaissances*. Elle nous permet de transformer des données en information, en connaissances[11]. Dans sa forme la plus simple, cette connaissance permet d'expliquer le sens de documents textuels publiés[18], d'éléments qui les composent, de services Web proposés. C'est la base du *Web sémantique*[19]. Des balisages sont utilisés pour préciser le sens des mots d'un texte, pour faire des ponts entre des ressources distinctes avec le *linked data*. Avec les ontologies, un pas supplémentaire est franchi pour atteindre le monde des connaissances structurées, classifiées, organisées.

Quels sont les buts recherchés ? D'abord d'améliorer l'information publiée, de la préciser, en lever les ambigüités pour permettre par exemple de proposer de meilleures réponses aux questions des internautes. Ensuite, il s'agit de faciliter le fusionnement[20] de connaissances proposées par plusieurs bases de connaissances (de plusieurs chercheurs, plusieurs labos, plusieurs communautés) en une base de connaissances unique, pour aller pourquoi pas vers une base universelle.

Un des premiers exemples très populaire de balisage de texte est le « Text encoding initiative » initié en 1987. Le but du balisage était de permettre de trouver plus facilement de l'information dans de larges collections de textes de bibliothèques. (Les balises étaient en SGML ; elles sont passées depuis en XML.) TEI était défini comme « un système pour faciliter la création, l'échange, l'intégration de données textuelles informatisées »[21]. Nous sommes au cœur du sujet.

Plus récemment, le Conceptual Reference Model[22] est une ontologie développée pour décrire le patrimoine culturel. C'est une norme très utilisée notamment pour enrichir l'information et les connaissances autour des collections de musées.

[18] Davis Randal, Howard Shrobe, Peter Szolovits, « What is a Knowledge Representation? » *AI Magazine*, 14(1), 1993, p. 17-33.
[19] Tim Berners-Lee, James Hendler, Ora Lassila, « The semantic web », *Scientific american,* 2001, 284.5, p. 28-37.
[20] Marie-Christine Rousset et al., « Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL » *Revue I3 (Information-Interaction-Intelligence)*, 2002, Vol. 2, n° 1.
[21] Lou Burnard, *Le schéma XML TEI pour l'édition*, Cours donné à l' Université d'été de l'édition électronique ouverte, 2009.
[22] The CIDOC Conceptual Reference Model. http://www.cidoc-crm.org/

Connaissances présentées dans le moteur de recherche Google

Ces deux exemples sont dans des domaines larges mais ciblés. La base de connaissances Yago a une vocation plus encyclopédique. Elle a d'ailleurs été développée à partir de la version anglaise de l'encyclopédie textuelle Wikipédia, un bon point de départ pour développer une grande base de connaissances. L'ontologie Yago a été construite à l'aide d'un logiciel développé à l'Institut Max Planck[23]. En 2011, Yago avait déjà 2 millions d'entités et plus de 20 millions de relations entre ces entités. Un autre projet, Wikidata, se propose de réunir dans une base de données éditée de manière collaborative des données objectives, telles que les dates de naissance ou bien le PIB des pays, pour être utilisées en complément de Wikipédia. Pour les connaissances proposées typiquement à droite de l'écran quand une question comme « Paul Verlaine » est posée, le moteur de recherche Google va utiliser Wikidata. (Voir illustration)

Les ontologies servent aussi à formaliser les connaissances que nous pouvons acquérir (automatiquement ou pas) dans un domaine particulier. Par exemple, le projet ANR Hyperprince[24] s'est focalisé sur la mise en correspondance de plusieurs traductions du texte de Machiavel. Il a permis d'étudier notamment l'évolution du vocabulaire historique.

Pour pouvoir échanger des connaissances, pour pouvoir être aidés par des systèmes informatique pour stocker, sélectionner, rechercher, inférer des connaissances, etc., nous formalisons ces connaissances dans des ontologies qui ne capturent que des visions finalement assez limitées de ces connaissances. Pour représenter des connaissances dans un langage formel, il nous faut les structurer, les « normaliser », d'une certaine façon les affadir. Evidemment, nous pouvons repousser sans cesse les limites. Par exemple, les informations subjectives contenues dans les textes sont le plus souvent ignorées. Cette connaissance est donc le plus souvent absente des représentations numériques. Des techniques récentes d' « analyse de sentiment » basées sur la linguistique computationnelle (traitement du langage naturel) permettent d'enrichir la représentation de connaissances avec des informations subjectives. Un rôle des humanités numériques est de repousser toujours nos limites, dans la quête de toujours plus de connaissances.

Le passage des données aux connaissances est au cœur du développement des humanités numériques parce que ce passage permet de mieux faire collaborer les chercheurs et les systèmes au développement de la recherche. Il exige des chercheurs d'aller plus loin dans la formalisation des connaissances dont ils disposent. C'est donc dans ce passage que peut le mieux se situer la nécessité de la « pensée informatique » : pour que cette collaboration puisse s'installer, les chercheurs sont amenés à penser autrement. Ils sont amenés à dépasser la question de comment utiliser des moyens numériques pour aller jusqu'à repenser leur façon de mener leurs recherches.

Toujours plus de connaissances

Dans une quête permanente de connaissances, des systèmes informatiques permettent d'obtenir toujours plus de connaissances.

[23] Yago2. http://www.mpi-inf.mpg.de/yago-naga/yago/
[24] Hyperprince. http://hyperprince.ens-lyon.fr/

Collecte et analyse de données. Il est intéressant de remarquer que c'est peut-être une des dimensions les plus anciennes des humanités numériques avec Roberto Busan, un jésuite italien, qui a imaginé dans les années 1940 et réalisé ensuite l'analyse linguistique basée sur l'informatique des œuvres complètes de Thomas d'Aquin. On retrouve les techniques d'analyse de texte qu'il a utilisées (indexation, cotexte, concordance, cooccurrence, etc.) dans de nombreuses disciplines notamment en histoire ou en littérature.

Les plus paradigmatiques exemples de cette analyse de données (notamment de par leurs masses) viennent peut-être de Google trends. Google trends permet d'avoir accès à la fréquence d'un mot dans les requêtes au moteur de recherche Google (près de 10 milliards de requêtes par jour en 2014). Il a donné lieu à de nombreuses études. Par exemple :

- La détection d'épidémies. Une corrélation entre la présence de certains mots dans les requêtes et le nombre de cas de grippes détectés a été mise en évidence.
- L'index d'orientation future. Il a été constaté que les internautes des pays dont le PIB brut est plus élevé sont plus susceptibles de rechercher des informations sur l'avenir plutôt que sur le passé.

Le système Hyperbase[25] d'Étienne Brunet est un exemple plus académique. Ce logiciel permet l'exploration et les analyses statistiques de corpus textuels. Il a été utilisé en linguistique, littérature, histoire, sociologie et sciences politiques. On peut aussi utiliser des outils numériques pour étudier l'intertextualité sous ses formes multiples de la citation au plagiat, en passant par l'allusion.

La production d'information atteint des niveaux que l'on pouvait difficilement prévoir. Par exemple, une personne « un peu connectée » construit une image digitale dans presque toutes ses activités : déplacements, communications, achats, travail, vie sociale, etc. Ces traces forment un matériel extraordinaire pour les chercheurs en sciences humaines. Encore récemment, pour étudier les déplacements estivaux d'une population, la collecte d'information était fastidieuse et coûteuse. Aujourd'hui on peut disposer de masses d'information considérables par les opérateurs téléphoniques (GPS de portables), ou des plateformes d'achat comme booking.com. Pour prendre un autre exemple, nous disposons de volumes de texte de plus en plus gigantesques. Cela a ouvert un champ d'étude considérable, celui des approches statistiques en linguistique computationnelle. Grâce à ces énormes corpus, l'extraction de connaissances à partir de textes a fait d'énormes progrès ; et en utilisant en particulier des textes multilingues alignés, la traduction automatique a progressé.

Evidemment, ces collectes massives d'information posent aux chercheurs de nouvelles questions comme les intrusions sur la vie privée d'un individu qui peut en résulter, ou le déséquilibre que les effets réseaux peuvent introduire : les plus grosses plateformes collectent plus d'information et introduisent des biais dans les concurrences commerciales.

L'inférence. Un avantage d'une représentation numérisée est que des programmes peuvent automatiquement aller chercher de l'information présente *implicitement* dans les données. Dans le plus simple des cas, c'est un système de gestion de bases de données qui stocke des données sur les joueurs de football et les matchs. Il peut facilement répondre à des requêtes comme de comparer les statistiques des joueurs sortis de différents centres de formation au cours du temps. Nous avons vu aussi comment des con-

[25] Hyperbase. http://fr.wikipedia.org/wiki/Hyperbase

naissances pouvaient être dérivées à partir d'analyses statistiques des données. Les systèmes autour des ontologies permettent également de réaliser des inférences. Pour être capable à partir de données brutes comme la consommation d'aluminium d'un pays, sa production, ses exportations, de dériver la place du recyclage dans ce pays, il faut disposer d'ontologies qui expliquent les concepts du domaine et relient les données disponibles à ces concepts. Dans un monde où nous sommes entourés de plus en plus de systèmes informatiques qui produisent, stockent et échangent de l'information, la place de l'inférence de connaissances ne peut qu'être appelée à grandir et ce notamment dans les humanités numériques.

La modélisation et la simulation

Dans les sciences physiques et les sciences de la vie, la modélisation numérique tient une place considérable. En simplifiant, le chercheur propose un modèle du phénomène complexe étudié, et le simule ensuite numériquement pour voir si les comportements résultants correspondent à ceux observés dans la réalité. Parmi les plus grands challenges actuels, on notera par exemple le « Blue brain project » à l'initiative de l'Ecole Polytechnique Fédérale de Lausanne qui vise ni plus ni moins que de simuler numériquement le cerveau humain.

La modélisation et la simulation tiennent une place de plus en plus importante en SHS.

Un exemple intéressant est celui du droit. Dans un ouvrage[26] maintenant célèbre Lawrence Lessig écrivait en 1999: « Code is law ». Bien sûr, nous ne pouvons plus ignorer l'importance sur notre société de codes informatiques comme celui qui classe les pages des résultats d'un moteur de recherche (PageRank) ou celui du DRM qui nous empêche de prêter à un ami un film numérique que nous venons d'acheter. Mais avant même ces codes informatiques, on peut voir le droit comme un ensemble d'instructions qui définissent le fonctionnement d'un système. Si pour des raisons historiques et involontaires, les lois sont imprécises, incomplètes, ambiguës, la pensée informatique peut permettre d'aller vers une plus grande formalisation des lois pour les rendre plus compréhensibles notamment de programmes informatiques. Dans cette direction, des outils informatiques sont mis au service de l'analyse des lois comme par exemple TheLawFactory.fr[27].

En SHS, les phénomènes sont souvent complexes et ouvrent des challenges pour la modélisation mathématique/informatique, voire la simulation numérique. La sociologie est en particulier un candidat évident. Il est possible de s'appuyer sur la modélisation (extrêmement simplifiée) des comportements d'un très grand nombre d'acteurs (agent dans une terminologie informatique populaire) et de leurs interactions avec leur environnement. La puissance de calcul de clusters d'ordinateurs permet ensuite de réaliser des simulations. Avec un millier de machines simulant chacune des milliers d'individus, il devient par exemple possible de simuler le comportement de population de millions de personnes. La comparaison des résultats avec la réalité permet de « paramétrer » le modèle, voire de le modifier, pour mieux coller à la réalité observée. Une fois le modèle mis au point (« tuné »), rien n'interdit les prédictions. Nous sommes dans des processus très semblables à ce qui se fait dans des domaines scientifiques comme la météo.

---

[26] Lawrence Lessig, *Code and other laws of cyberspace*, New York, Basic books, 1999.
[27] TheLawFactory.fr est une application Web en logiciel libre qui permet de visualiser des informations disponibles sur 290 projets de loi promulgués entre 2010 et 2014.

Nous retrouvons dans ce contexte par exemple l'étude[28] de Paola Tubaro et Antonio Casilli sur les émeutes de Londres. Ils ont cherché à savoir si la censure des médias sociaux, solution proposée par David Cameron pour calmer la situation, avait un effet sur le développement d'émeutes. A l'aide d'une simulation numérique, ils ont montré que la censure participait à augmenter le niveau général de violence.

La tâche paraît plus ardue dans d'autres disciplines des sciences humaines ne serait-ce que par la difficulté de développer des modèles assez formels pour se prêter à la simulation. On notera pourtant un ouvrage récent comme *Modeling complexity in the humanities and social sciences* [29] avec des sujets comme le développement et la transmission du langage ou la propagation des croyances, idées et idéologies.

L'archivage

L'archivage est devenu essentiellement numérique. On peut mentionner par exemple Europeana[30], une bibliothèque numérique européenne lancée en novembre 2008 par la Commission européenne qui compte déjà plus de 26 millions d'objets numériques, textes, images, vidéos, fin 2013. Les États européens (à travers leurs bibliothèques nationales, leurs services d'archivages, leurs musées, etc.) numérisent leurs contenus pour assurer leur conservation, et les mettent en commun. En partant de telles initiatives, nous pouvons imaginer dans moins de 50 ans des historiens qui trouveraient numériquement toutes les informations dont ils ont besoin, passant d'une archive à une autre simplement en changeant de fenêtre sur leur écran.

L'immatérialité de l'information et des données pose aux archivistes et aux bibliothécaires la question de leur conservation et de leur archivage à long terme.

Préservation dans un monde numérique. L'information sur un disque magnétique dure bien moins longtemps que, par exemple, dans un livre imprimé sur papier. Les solutions techniques existent : il faut par exemple répliquer l'information pour se prémunir contre les défaillances matérielles, et il faut la régénérer régulièrement pour lutter contre l'obsolescence rapide des supports d'information numériques. Et puis, il faut se protéger contre l'obsolescence également très rapide des formats informatiques, des logiciels informatiques ainsi que des matériels. L'obsolescence des contenus est souvent accentuée par la disparition des logiciels qui leurs donnaient vie, voire des matériels. C'est le cas notamment pour ce qui est des jeux électroniques. Le projet *Preserving Virtual Worlds* s'est par exemple attaché à archiver des jeux informatiques, des fictions interactives et des communautés virtuelles. Il ne s'agit pas de préserver seulement des formats qui décrivent les contenus, mais aussi des logiciels et les matériels qui permettent de les faire vivre.

---

[28] Antonio A. Casilli and Paola Tubaro, « Social Media Censorship in Times of Political Unrest-A Social Simulation Experiment with the UK Riots », *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique,* 2012, 115, p. 5-20.
[29] Paul A. Youngman and Mirsad Hadzikadic, *Complexity and the Human Experience: Modeling Complexity in the Humanities and Social Sciences*, New York, Taylor and Francis, 2014.
[30] Europeana. http://www.europeana.eu/

La préservation implique de préserver aussi la « qualité » de l'information. Les données numériques comme d'autres données, plus peut-être que d'autres données, sont susceptibles d'être manipulées. L'archivage doit pouvoir garantir leur authenticité.

Culture d'abondance et l'échelle. Avec le numérique, nous sommes passés pour l'information disponible d'une culture de rareté, à une culture d'abondance. Avec la baisse du coût des machines et surtout du stockage, on a cru à la possibilité de tout conserver. Il a fallu déchanter. La préservation de l'information numérique (pour les raisons citées précédemment) revient chère. Pourtant, dans un contexte où nous nous sommes habitués à l'abondance, il faut conserver massivement l'information. Devant le déluge informationnel, choisir ce qu'il faut archiver, voilà un vrai défi pour les humanités numériques. Et il s'agit vraiment de l'automatisation de ce choix parce que l'archiviste est tout petit devant la masse des données. Sa seule chance de réaliser la tâche qui lui est confiée est d'utiliser massivement des outils numériques.

L'archivage du Web. Il a fallu faire évoluer des centres classiques d'archivage comme les Archives nationales, la BNF et l'INA, et des outils anciens comme le dépôt légal. Un exemple donnera toute la dimension de l'archivage dans le monde numérique : l'archivage du Web. Nous trouvons de plus en plus d'informations sur le Web. Que seront devenues ces informations dans 50 ans quand des chercheurs en auront besoin ? La plus importante initiative pour archiver le Web est venue de la fondation Internet Archive[31] qui archive des pages du Web depuis 1996. A titre d'exemple, la Library of Congress a récupéré 170 milliards de tweets juste entre 2006 et 2010 représentant 133.2 téraoctets ! Cela donne une idée des dimensions du problème.

Il faut aussi avoir conscience que, même si l'archivage du Web a fait énormément de progrès, une part importante du Web n'est pas aujourd'hui archivée pour de nombreuses raisons comme la taille du Web, les pages qui changent rapidement de contenu, les pages protégées par mot de passe (comme Facebook) ou du fait des « exclusions[32] de robot », et surtout des pages générées par des requêtes (comme eBay).

Conclusion : limites, difficultés et ambitions

Les humanités numériques modifient les modes de travail et de pensée dans les sciences humaines et sociales, non sans difficultés.

Limites de la technique. Si les opportunités sont nombreuses, tout n'est pas possible. Certains problèmes demandent des puissances de calcul dont nous ne disposons pas ou que nous ne sommes pas prêts à mobiliser pour un problème particulier. Les simulations peuvent être trop coûteuses (trop d'acteurs en sociologie), les problèmes trop complexes dans le cadre des connaissances actuelles (analyse sémantique de texte). Les machines gagnent sans cesse en puissance, vitesse des processeurs et des réseaux, taille des mémoires et des disques, parallélisme massif. La qualité et les performances des algorithmes ne cessent de progresser. Reste que les plus grandes avancées en humanités (numériques ou pas) reposent sur l'imagination d'humains qui trouvent la bonne question, énoncent la bonne hypothèse, proposent l'approche révolutionnaire. Les machines et les algorithmes ne sont pas prêts à remplacer ces personnes ; ils sont à leur service, évoluant sans cesse pour repousser le domaine du possible.

---

[31] Internet Archive. https://archive.org/

[32] Le propriétaire de la page interdit son archivage.

Limites de l'objectivité. Dans le cadre des SHS, il faut aussi savoir accepter les limites de l'objectivité. Les humanités numériques ne peuvent se réduire à des équations ou des algorithmes (les plus beaux soient-ils) et des nombres. Le sujet principal est l'être humain, souvent trop complexe pour être mis en équation ou même en algorithme.

Mutabilité du monde actuel. Une limite aussi est la rapide mutabilité du monde actuel. La culture est devenue numérique et évolue sans cesse ; l'homme évolue également avec le numérique dans sa manière de lire, d'apprendre, etc. Le numérique tient une place essentielle *dans le sujet même* des humanités numériques, comme l'explique Milad Doueihi[33], en conduisant à la définition de nouvelles humanités. Une difficulté est donc cette fluidité du contexte où à la fois le sujet même et les sciences qui nous permettent de l'étudier changent à l'unisson.

Mutualiser les savoirs. Une conclusion à tirer de cette étude serait peut-être la similarité des idées communes à un spectre très large de disciplines scientifiques (humaine ou pas) et trouvées avec le numérique, par de nouveaux outils, de nouvelles approches, de nouveaux savoirs. C'est au niveau des outils que cette évidence s'impose. On rappellera qu'un ordinateur est « general purpose », et que le même système de gestion de bases de données peut servir pour des données en histoire, en astronomie, etc. Une ambition du chercheur en humanités numériques et plus généralement en sciences numériques doit donc être de mutualiser ses réalisations (logicielles, conceptuelles, ou autres) au delà de son propre domaine. Les principes même de la « pensée informatique » (*computational thinking*[34]) sont généraux et se doivent d'être également « mutualisés ».

Inventer un nouvel humanisme. La culture numérique est fortement imprégnée de valeurs humanistes. Si Internet et le Web sont au départ des techniques et des outils, ils sont inséparables de modèles sociaux et économiques basés sur le partage et l'échange. Nous avons mentionné par exemple Wikipédia et les Creative commons dont le caractère humaniste doit être souligné. Il nous semble donc que, par essence, les humanités numériques ne peuvent se résumer au seul développement des connaissances scientifiques. Elles se doivent d'adopter les ambitions humanistes déjà présentes dans la culture numérique. Et par exemple, quelle plus grande ambition humaniste que de mieux diffuser les connaissances et la culture à tous. L'ambition des humanités numériques doit donc bien être de participer à l'invention d'un nouvel humanisme.

Remerciements. Merci à Gilles Dowek, Mathieu Latapy, Pierre-Michel Menger et Marc Tommasi pour leurs critiques de versions préliminaires de ce texte.

[33] Milad Doueihi, *Pour un humanisme numérique,* Paris, Seuil, 2011.
[34] Jeannette M. Wing, « Computational thinking », *Communications of the ACM,* 2006, 49.3.