Produire, analyser et partager des données ouvertes en Humanités Numériques : quelques bonnes pratiques

Gérald Kembellec

Produire, analyser et partager des données ouvertes en Humanités Numériques : quelques bonnes pratiques

Gérald Kembellec

MCF en Sciences de l'information et de la communication

Laboratoire Dicen-IdF, CNAM

gerald.kembellec@cnam.fr

Résumé

La réponse à des problématiques scientifiques liées aux humanités passe par le traitement numérique de corpus. Les humanités numériques deviennent un sujet d'importance qui regroupe des savoirs et des méthodes issus de diverses disciplines comme l'informatique, les statistiques, la sociologie, la cartographie ou encore la linguistique. Cet article, s'il est ancré dans les sciences de l'information et de la communication, convoque des méthodes périphériques et se propose comme un *vade-mecum* de la gestion des données des humanités : la qualification, la collecte, le traitement, l'enrichissement, la documentation et le partage des données des humanités. Nous mettons ici en avant le concept de « courtoisie du FAIR data » en contexte scientifique : la valorisation des corpus, en particulier par le partage de jeux de données de qualité, documentés et accessibles physiquement et légalement exploitables. Nous insistons également sur l'éthique lors des étapes de traitement et d'exploitation des données de la recherche.

Mots clés

Humanités numériques, courtoisie du FAIR data, bonnes pratiques, données numériques, données ouvertes, qualités des données.

Title

Producing, analyzing and sharing Open Data in Digital Humanities: some good practices

Abstract

Scientific problems associated with the Humanities lies in the digital processing of corpora. Digital Humanities are becoming an important subject that brings together pieces of knowledge and methods from various disciplines such as information and library sciences, computer science, sociology, cartography, statistics and linguistics. This article is a *vade-mecum* for humanities data management: the qualification, collection, processing, enrichment, documentation, and sharing of humanities datasets. We put forward here the concept of "data courtesy" in a scientific context: the enhancement of corpora, in particular through the sharing of quality, documented and physically and legally accessible data sets. We also emphasize ethics in the processing and exploitation of research data.

Keywords

Digital Humanities, FAIR data courtesy, good practices, Linked Open Data, Data Quality

INTRODUCTION

Les projets scientifiques liés aux activités de recherche en Sciences, Technique et Médecine (STM) produisent, analysent et partagent des jeux de données dont les mécanismes de traitement sont strictement encadrés par des comités scientifique et d'éthique, ainsi que par une cohorte de spécialistes de la qualité. Ces bonnes pratiques sur la collecte, la curation, la manipulation, la documentation et le partage des données de la recherche nécessitent des aptitudes qui s'acquièrent par l'expérience, mais aussi par l'analyse des projets passés et des revues de littérature (Schneider, 2013). Même si des scandales éthiques apparaissent dans des revues scientifiques dites « prédatrices » à la moralité douteuse (Gingras, 2018) et que l'on note des cas de fraudes expérimentales, il est évident que l'usage scientifique des données est une pratique qui évolue vers une certaine maturité en STM. Cela s'explique par le fait que les protocoles expérimentaux dans ces disciplines bénéficient d'une longue tradition qui s'est muée en une indéniable expertise. De plus, les avancées métrologiques constatées depuis le 18e siècle (cf. projet ANR, « Le Bureau des longitudes », Muller et al., 2019) couplées à la possibilité d'usage de logique computationnelle initiée au 20e siècle ont permis un traitement plus rapide et une réduction des erreurs de calcul.

Ces problématiques de collecte, d'usage, de calcul et d'interprétation des données de recherches interviennent également en SHS (Bastin & Tubaro, 2018). Les sociologues sont depuis toujours de fins statisticiens et les historiens ne sont pas en reste avec le développement de méthodes et d'outils numériques de collecte, curation et croisement de données documentaires comme Zotero, de partage, voire de sémantisation de contenus avec OmekaS. L'historienne Claire Lemercier déclarait dès 2008 à propos de l'utilité de la littératie numérique en SHS : « *compter, comparer, classer, modéliser restent des moyens utiles pour mesurer notre degré de doute ou de certitude, pour expliciter nos hypothèses ou évaluer le poids d'un phénomène* » (Lemercier & Zalc, 2008). Cette question de l'usage des données s'ancre tellement dans les SHS que même les disciplines issues des « studia humanitatis » au sens traditionnel, comme les langues anciennes, l'histoire de l'art, la théologie… usent maintenant de numérisation, de traitements automatisés pour leur développement scientifique. Ceci correspond à une acception définitoire parcellaire, mais qui est la nôtre, du phénomène des « humanités numériques » qui émerge actuellement en SIC comme un champ d'études à part entière. Dans cet article, nous proposons, au moyen d'exemples, une feuille de route pour penser « la donnée » au sein d'un projet d'humanités numériques. Nous y présentons les méthodes de collecte, d'enrichissement, de traitement, d'analyse, de visualisation et de partage des données en SHS avec une focale rétrospective sur des projets en humanités numériques dont nous tirerons des préconisations méthodologiques.

1 – DE LA PRODUCTION DE DONNÉES DES HUMANITÉS

1.1 Que sont les données de recherche en humanités ?

Nous situons la recherche en humanités dans un champ assez restreint, celui issu des « Arts libéraux antiques ». Même si le « Manifeste des *digital humanities* » de 2010 propose d'élargir les humanités numériques à « *l'ensemble des sciences humaines et sociales, des arts et des lettres* » et qu' « *elles s'appuient […] sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines, tout en mobilisant les outils et les perspectives singulières du champ du numérique* », nous pensons que si les méthodes peuvent donc être résolument modernes, les « terrains » d'études se doivent d'être plus classiques avec les lettres anciennes, l'art, l'histoire, l'histoire de l'art et la théologie… : ce qu'en histoire de l'éducation l'on nomme « humanités classiques et modernes » (Compère et Chervel, 1997).

On l'a compris, la définition et le périmètre des humanités numériques (ou encore *digitales* au sens de Le Deuff, 2015), ne sont pas encore des questions tranchées, les frontières sont

encore poreuses et c'est ce qui fait la richesse de ce débat : l'aspect pluridisciplinaire permet des avancées par la co-construction d'objets de recherche qu'il serait impossible de négocier sans un recours à plusieurs disciplines (Kembellec, Desfriches-Doria & Gispert, 2019).

Les données des humanités sont donc des corpus qui peuvent varier selon les disciplines et les champs d'études. L'exemple le plus parlant historiquement est bien sûr le travail de ce pionnier qu'était le père Roberto Busa avec son analyse philologique de l'œuvre St Thomas D'Aquin dont l'indexation avec l'aide informatique d'IBM a tout de même pris plusieurs décennies à partir de 1950 (Busa, 1974, 1980 ; Eberle-Sinatra & Vitali Rosati, 2014).

Les données des humanités peuvent donc être textuelles, mais il ne s'agit là que d'un aspect des contenus à traiter : certains corpus vont être iconographiques avec des images fixes ou animées, des partitions musicales, des cartes ou des implantations spatialisées. Il peut également s'agir de jeux de métadonnées avec des notices bibliographiques, etc. Tous ces ensembles de données ont cela en commun qu'ils étaient historiquement traités scientifiquement de manière manuelle depuis des sources primaires physiques dans les humanités et que pour transposer ce traitement chronophage à des processus automatisés ou semi-automatisés, ils doivent passer d'une forme physique à une forme numérique. C'est ce processus de collecte puis de prétraitement qui va conditionner la possibilité d'un travail ultérieur de qualité.

1.2 De la collecte des données

La question de la méthodologie de collecte avant même celle de l'analyse et de l'interprétation est cruciale, qu'il s'agisse de collecte de données de terrain ou d'analyse de traces sociales au sens du monde numérique (Merzeau, 2009). Dans le cadre strict des humanités, les enquêtes sociales, au sens moderne, ne sont pas convoquées. Il n'en reste pas moins que certaines des méthodes issues des sciences sociales peuvent être appelées ultérieurement pour l'analyse.

Les données des humanités ne sont par définition pas *nativement* numériques puisqu'elles préexistent généralement à l'arrivée de l'informatique. Pour effectuer une computation, ou un traitement numérique sur lesdites données, une première phase de numérisation va donc être indispensable.

L'un des vieux truismes du domaine de l'informatique : GIGO (*Garbage In, Garbage Out*) spécifie que les ordinateurs peuvent traiter, interpréter, stocker, récupérer et trier des données avec précision et fiabilité – à la condition toutefois de leur donner de bonnes instructions. La qualité du résultat est donc directement induite par celle des données initiales (Weinberg, 2003, p. 228-229). Il est donc évident que même avec l'aide précieuse d'une automatisation, un travail de recherche de qualité en humanités numériques ne peut s'effectuer qu'avec des données et des méthodes qualifiées.

Il existe plusieurs types de données, qui vont avoir des sources et des usages distincts. Pour un corpus d'images, les bonnes pratiques passent souvent par l'usage de la haute définition pour l'art et d'une définition raisonnable d'un minimum de 300 dpi pour l'archivage. Un outil de *Digital Asset Management* permet de classer et d'indexer les images tout en permettant leur accès, ce qui peut être idéal en contexte de gestion iconographique des humanités (Hamma, 2004 ; Wythe, 2007). Des outils spécifiques aux humanités comme Omeka sont tout indiqués pour ce genre de travail, particulièrement en Histoire et en Histoire de l'Art. Pour un projet d'archivage en Histoire des Sciences et Techniques, nous avons utilisé le format TIFF multipages en 300 dpi avec Omeka pour archiver un corpus d'images avec un plan de classement et un thésaurus *ad hoc* (Kembellec, Fournier & Cubaud, 2018).

Les données peuvent également être textuelles, dans ce cas il est d'usage que les documents originaux soient numérisés sous forme d'images, qui peuvent ensuite être versées dans une archive numérique comme Gallica pour la BnF ou sur un outil comme Omeka. Par la suite, pour un traitement computationnel sur le texte à proprement dit, les images de texte sont traitées par un outil de reconnaissance optique de caractères (OCR) et de transformation en texte brut. Cette phase peut être complexe, car les textes n'ont historiquement pas toujours été produits en série et certains textes médiévaux par exemple, ont des graphies « particulières ». Cependant la recherche numérique a aussi progressé en ce sens et il existe des logiciels de conversion de textes numérisés avec des options « paléographiques » pour répondre ce genre de problématiques. Bien sûr, même avec des techniques avancées d'apprentissage profond, la vérification manuelle par des spécialistes reste le plus souvent nécessaire. Cependant, la réciproque peut aussi être vraie comme nous allons ici le voir.

Dans le cadre d'un autre projet de recherche en humanités, nous avions établi un protocole de saisie de notices bibliographiques faisant appel à des chercheurs ou des étudiants qualifiés, spécialistes du domaine (Kembellec, Desfriches-Doria & Gispert, 2019). Pour chaque auteur de la cohorte traitée, la totalité des notices était validée avant de commencer la saisie du corpus. De plus, des masques de saisie bridaient au maximum les erreurs par des sélections en lieu et place des champs libres. Enfin, une phase de relecture systématique et de pointage était effectuée au sein du dispositif de recherche et d'affichage. Une fois la base manuellement validée de manière « experte », nous avons, par acquis de conscience, effectué une visualisation pour vérifier graphiquement s'il n'y avait pas de données dites « aberrantes ». La visualisation ci-après montre une forte concentration de documents publiés entre 1850 et 1970, ce qui correspond à la période étudiée. Le schéma en Figure 1 montre que les données non regroupées (6 valeurs) sont des erreurs : les dates entre les 1er, 4e, et 11e siècles sont peu susceptibles d'avoir été correctement saisies.

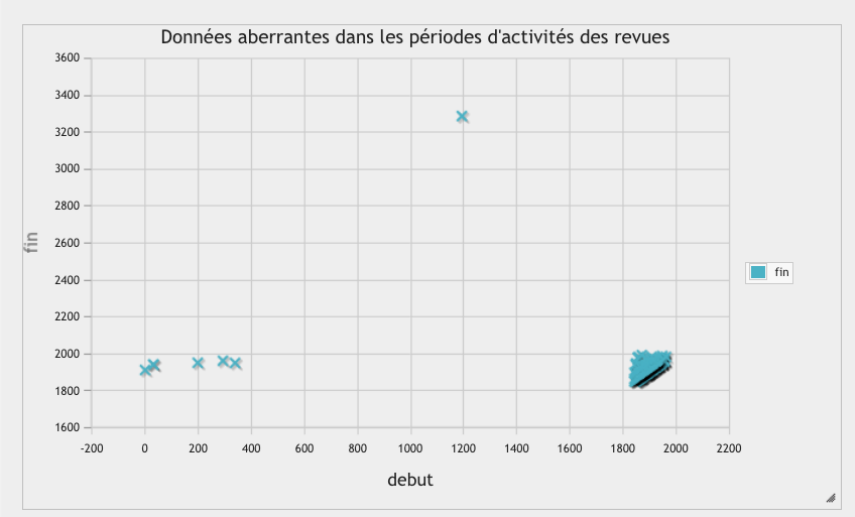

Figure 1. Détection de données aberrantes

1.3 De l'harmonisation des données

Il est d'usage de parler des données initiales, avant traitement, comme des données brutes, c'est-à-dire exemptes de tout traitement. Cette notion est fortement contestée, car tout traitement numérique peut être considéré comme un parti pris avec des choix impactant le format de stockage, les normes d'encodage, ou la structuration par exemple. Bowker y voit un oxymore autant qu'une aberration : les données doivent être structurées autant que possible

(Bowker, 2005 ; Plantin, 2013). Pour simplifier, les données initiales recueillies se doivent, pour être à même de subir un traitement numérique, d'être au maximum compatibles avec les logiciels les plus courants. Il est possible de tabler sur quelques formats dits « universels » pour stocker les jeux de données qu'ils soient textuels, numériques ou encore iconographiques. On peut catégoriser les données en partant des moins structurées vers les plus structurées de la manière suivante :

Faiblement structurées (un simple éditeur de texte suffit à les créer ou les lire) :

- Les fichiers texte TXT sans aucune mise en forme ;
- Les fichiers CSV ou TSV qui peuvent contenir des jeux de chiffres et de texte avec des séparateurs (virgule, point-virgule ou tabulation) en colonnes et lignes.

Peu structurées (peuvent nécessiter un éditeur spécialisé pour mieux appréhender la création la lecture ou le filtrage) :

- Les fichiers JSON qui permettent de formaliser des objets de manière plus ou moins complexe selon les besoins ;
- Certains fichiers XML avec des structures simples comme les notices bibliographiques MODS.

Fortement structurées (un logiciel dédié est absolument nécessaire pour la création et la consultation) :

- Des bases de données relationnelles, les données sont bien décrites individuellement et entre elles. Il est possible d'y stocker tout type de données.
- Les données RDF qui sont structurées et décrites, mais aussi inter-reliées selon les règles du Web des données.

1.4 De la réconciliation et de l'enrichissement des données

Une fois ces données « propres », sans erreur et correctement encodées, il reste possible de les enrichir par un processus dit de « réconciliation ». Le principe de la réconciliation des données est d'enrichir un jeu de données avec des informations connexes issues de dépôts d'information des autorités de la gouvernance du Web. Ainsi, lors d'un projet, si nous disposons d'un jeu de données sur une cohorte d'acteurs, il est envisageable d'obtenir des informations complémentaires et de désambiguïser certains homonymes. Avec une solution de type « OpenRefine [1] », il devient simple de désambiguïser et d'enrichir les jeux de données à partir de l'ISNI par exemple. Ces jeux, une fois augmentés d'informations issues de sources fiables sont d'autant plus pertinents pour être analysés.

2 – De l'analyse des données

2.1 Du traitement quantitatif (et éthique) des données

Dans l'introduction nous pointions que certaines analyses et interprétations des métriques aient pu conduire - à la marge - à des erreurs ou biais, parfois induites par une idéologique « particulière » comme dans le cas de celles commises en phrénologie [2]. Mais, au-delà du positivisme et de la qualification de la scientificité, ce sont souvent des questions de méthodologie qui pêchent dans le protocole de collecte ou d'analyse. Il arrive également que ce soient des raisonnements fallacieux (délibérés ou pas) qui posent le problème de l'interprétation et pas les données elles-mêmes.

Comme l'utilité et la valeur des diverses données ne peuvent être séparées des opérations algorithmiques qui leur sont appliquées (Cardon, 2015), il paraît logique de comprendre *a*

minima comment agissent ces suites d'instructions simples résolvant un problème (Abiteboul & Dowek, 2017). Cependant, il a déjà été pointé que la qualification des chercheurs en SHS dans les méthodes de création, d'exploitation et de partage des jeux de données de la science était parfois biaisée.

En effet, comment organiser une étude sans connaître réellement les bonnes pratiques de méthodologie d'enquête, que ce soit de manière quantitative ou qualitative ? Comment dissocier les simples corrélations des facteurs de causalités si l'on n'est pas un minimum initié à la sociologie et aux statistiques (Pavel & Serris, 2016) ? Rappelons que Kenneth Rogoff et Carmen Reinhart, deux économistes de l'université américaine de Harvard ont publié une étude en 2010 contenant des conclusions erronées dues à une grossière erreur de calcul (Reinhart & Rogoff, 2010) qui a été pointée par un étudiant et ses encadrants (Herndon, Ash, & Pollin, 2014) obligeant ainsi les auteurs de l'erreur à publier un *erratum* (Reinhart & Rogoff, 2013).

À titre personnel, n'étant pas statisticien, nous n'avons pas hésité à demander de l'aide à une collègue statisticienne pour la bonne compréhension des phénomènes de corrélation ou de dé-corrélation sociale afin d'éviter de mal interpréter les liens pouvant exister de manière positive ou négative entre des individus par l'analyse de jeux de données déclaratives [3]. Un entretien informel avec un-e collègue spécialiste en statistiques, ou même juste d'Excel dans ce cas précis, aurait pu éviter à Reinhart et Rogoff de passer des moments socialement « complexes » dans leur communauté.

2.2 Gnose ou découverte ?

Pour travailler sur les jeux de données de manière quantitative, il faut se poser la question de l'objectif du travail de recherche. Cet objectif peut être (1) la confirmation d'une thèse formulée en hypothèse par un spécialiste du sujet et qu'il souhaite valider selon un modèle statistique, par exemple mesurer la relation entre des variables par un coefficient de corrélation ou encore (2) une approche agnostique aux connaissances qui va permettre grâce au regroupement automatique (*clustering*) des données de faire émerger de nouvelles questions de recherche [4] (Arruabarrena et al. 2019).

2.3 Visualisation, filtrage des données

Nous ne parlons ici que des difficultés évoquées précédemment, mais qu'en est-il également de la manière de représenter graphiquement des résultats sans connaître les bases de la sémiologie graphique (Bertin, 1970 et 1973) ou les biais introduits par les échelles ou les bases choisies ? Évidemment, la manière de représenter graphiquement l'information ne saurait être neutre, car le grain, l'échelle, la forme de représentation par exemple constituent une éditorialisation des données qui va amener le lecteur à les interpréter (ibid.). Il y a donc là, des jeux de sélection et d'échelle qui vont influer sur l'analyse et l'interprétation du lecteur, peut-être même influencer son jugement. Si la neutralité de présentation n'est pas possible, l'éthique du chercheur l'empêchera d'user de ficelles (parfois grossières) pour masquer des résultats peu convaincants : histogrammes tronqués, valeur en base 100 sur des séries chronologiques parcellaires ou des échelles disproportionnées (Kembellec, 2017).

Face à des corpus de données très étendus, les méthodes de représentation graphiques interactives, par exemple sur des graphes interpersonnels, des séries temporelles ou géographiques peuvent proposer une aide à l'analyse et au filtrage. Lors du *Hackathon* sur les données du projet « Bibliographies de critiques d'art francophones », un participant a

proposé une interface dynamique de représentation de corpus en réalisant un *clustering* des auteurs de la cohorte en fonction des revues dans lesquelles ils ont publié (cf. Figure 2). Chaque revue est représentée par un nœud, chaque auteur commun entre deux revues par un arc. À partir de ce travail, il est possible de proposer différents clusters, c'est-à-dire distinguer des groupes de revues ayant le plus d'auteurs en commun. Un petit cluster d'auteurs à la conjonction des critiques d'art en revues littéraires, plastiques et génériques a éveillé l'intérêt des chercheurs : il s'agissait de l'observation de l'émergence de la critique dans le 7e art naissant. Cela présentait à la fois une hypothèse sur l'« ADN » de ce nouvel art, mais aussi son impact social.

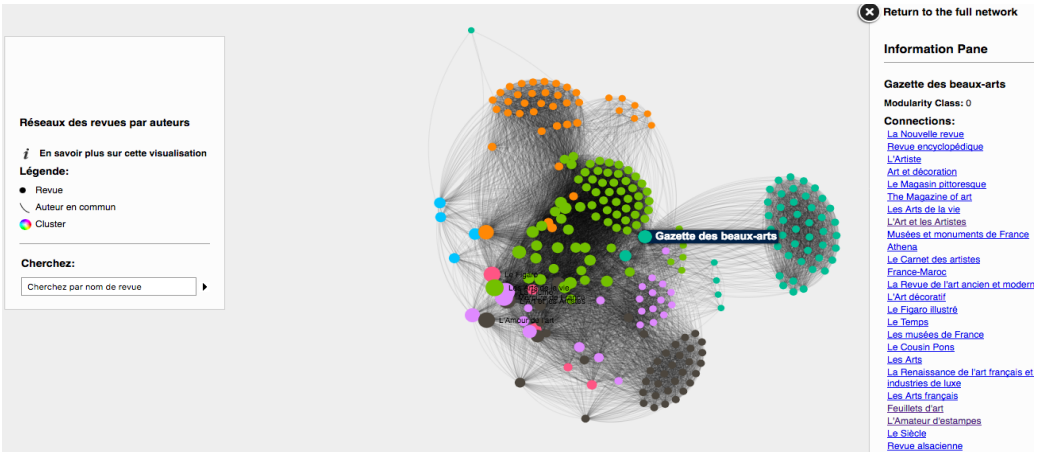

Figure 2. Langlais, P-C. (2017), Visualisation des clusters de critiques d'art par leurs revues de co-publication, In 1er Hackathon interdisciplinaire sur la critique d'Art, INHA, Paris

3 – DES INTERFACES DE RECHERCHE ET DE PARTAGE DE DONNÉES DES HUMANITÉS

L'exemple présenté dans la partie précédente démontre l'intérêt de la visualisation pour accélérer et faciliter la compréhension des résultats de *clustering*. Proposer une interface manipulatoire graphique de sélection et de filtrage en prise directe avec les données du corpus peut également être une option intéressante, comme dans le cas – par exemple - de la libraire R Shiny qui offre un filtrage cartographique intuitif. Comme vu précédemment, une fois les données sélectionnées, il reste encore un long chemin pour en faire des informations, voire des connaissances à l'aide de sa propre expertise du domaine et bien sûr l'usage éclairé de méthodes mathématiques et statistiques. Si ces traitements peuvent être automatisés, il faudra dans un premier temps s'approprier les jeux de donner, les « capter » au sens des « captas » de Plantin (2013) reprenant Drucker (2011) et soulignant l'effort déployé en ce sens. En effet, l'usager doit recopier ou « *scrapper* » les données par d'habiles manipulations techniques ou encore l'usage de logiciels dédiés afin d'obtenir des matériaux exploitables. Une autre solution, celle que nous nommerons la « courtoisie du FAIR [5] data » est d'exposer les données, de les rendre accessibles et documentées, prêtes à être exportées et travaillées (Wilkinson, 2016).

3.1 De l'exposition et du partage des notices

La notion de FAIR passe en premier lieu par l'exposition des données et métadonnées dans le cadre d'un projet scientifique. Il s'agit de la mise en œuvre de la médiation et de la

documentation des jeux de données et métadonnées pour maximiser les potentialités de leur réutilisation.

Le premier point à présenter est l'exposition bibliographique. Cet aspect permet d'accélérer la collecte de notices lors d'une action de recherche et de filtrage sur un corpus. Dans un projet de collection, comme c'est le cas dans « Bibliographie de critiques d'Art francophones [6] », c'est une problématique centrale. Prenons l'exemple d'un étudiant ou d'un chercheur en Histoire de l'Art qui cherche à exporter les titres des productions écrites de la chronique « La Vie Artistique » de Guillaume Apollinaire dans le journal « *L'Intransigeant* » - 151 notices tout de même - il sera précieux de pouvoir faire une extraction en un clic à l'aide d'un logiciel de gestion de références bibliographiques comme Zotero. Ce type de méthode passe par l'application de normes techniques dédiées comme *OpenURL COinS* ou *unAPI* que nous ne détaillerons pas ici (Chudnov et al, 2006), mais qui permettent une interaction de collecte bibliographique quasi automatisée (Kembellec, 2012, p. 228-234). Cela représente premièrement un gain considérable de temps pour le chercheur-usager et deuxièmement un gain de qualité, car tout risque d'erreur de recopie est écarté. Des solutions techniques comme Omeka ou Refbase permettent de proposer ce genre de services sans posséder de compétence technique particulière. Il n'en va évidemment pas de même si l'on réalise un dispositif *ad hoc* pour des besoins spécifiques.

3.2 De la sémantisation des fragments documentaires

Pour continuer à décliner cette « courtoisie du FAIR data », focalisons-nous sur la possibilité de rendre les données et métadonnées « trouvables » et « interopérables ». Nous proposons de systématiser l'accès aux fragments documentaires de nos productions d'humanités en ligne. Cette opération est rendue possible par les ancres de localisation au sein des dispositifs Web et l'aide d'une des méthodes d'exposition de triplets sémantiques, à savoir le RDFa, les micro données ou équivalents (Gandon et al., 2012; Kemebellec & Bottini, 2017). La pratique de ces méthodes permet d'inte-relier des personnes, des concepts ou des objets grâce à la sémantique. Par exemple, en exégèse, si l'on veut mettre en relation la péricope biblique du « retour du fils prodigue » (Luc 15 :11-32) avec des œuvres éponymes l'illustrant (Gustave Doré, Rembrandt…), il faudra user de ces méthodes pour expliciter le lien entre les identifiants des œuvres illustrant le propos, leur auteur, le vecteur artistique et le propos exprimé. Les ontologies ainsi produites sont porteuses de connaissances intrinsèques qui vont être collectées et agrégées en « *Knowledge Graphs* » par les moteurs de recherche, qu'ils soient commerciaux ou spécialisés (Prime-Claverie & Kembellec, 2016). Ainsi, il sera possible d'exprimer des relations du type : telle œuvre d'art (que nous utilisons comme illustration), exposée dans tel musée, réalisée par tel artiste, à telle date, sur tel support se réfère à telle portion de telle œuvre littéraire (dont voici le texte). Ainsi, ce que l'œil de l'usager voit à l'écran, un système d'information est capable de le comprendre dans les moindres détails, tout du moins d'en avoir la vision proposée lors du travail d'éditorisalisation par une documentarisation auctoriale. La systématisation des liens conceptuels et de paternité est rendue possible dans le code source du dispositif grâce à des autorités, des schémas et les contenus sélectionnés par les acteurs du projet.

3.3 De l'Open Data en Humanités Numériques

Les enjeux de *l'Open Science* sont souvent cristallisés autour de la publication scientifique, mais il ne faut pas oublier que les données en sciences ouvertes peuvent aussi être des jeux de données réunies ou transformées lors d'un projet de recherche. Le nouveau genre littéraire

scientifique du *datapaper* est intimement lié à la production de données et à leur libération. De plus, de nouveaux services d'appui à la recherche émergent comme la mise à disposition entrepôts de données. En France, il existe des espaces de dépôt des jeux de données dédiées aux humanités comme celui d'Humanum avec une interface d'import-export nommée Nakala [7]. Il ne s'agit pas simplement d'y « déposer » des données brutes (*datasets*), il faut aussi les documenter, comme ce doit être le cas lors du dépôt du *Data Management Plan* (Reymonet et al., 2018). L'enjeu va bien au-delà de la proposition d'accès des exports de base de données : il faut proposer des jeux cohérents et compréhensibles avec des champs les plus « normalisés » possible. Cette normalisation va permettre de rendre les travaux de recherche *a minima* reproductibles, mais également possiblement réutilisables pour proposer d'autres analyses, voire contredire celles des auteurs avec d'autres hypothèses ou d'autres méthodes avec les outils de l'usager final.

Certains estiment *a contrario* que les données doivent être vierges de tout traitement pour laisser un maximum de champs des possibles aux usagers. Ils estiment préférable d'exiger de tendre à cet état de virginité de la data par « brutification », c'est-à-dire avec un minimum de transformation, dans des formats les plus neutres et ouverts que possible, aussi peu structurés que possible (Denis et Goëta, 2017). Ces débats sociotechniques, avec des implications politiques sur l'usage des normes, formats et standards sont passionnants et d'une évidence limpide en sciences dures. Cependant, en humanités, il est évident que plus la donnée est documentée et structurée, plus elle a de chance de devenir de l'information, voire de la connaissance scientifique. De plus, comme présentés précédemment, certains auteurs n'hésitent pas à qualifier la donnée brute d'oxymore : le paramétrage, le calibrage, les algorithmes, la programmation sont des activités orientées humainement ce qui relativise le concept de « brutification » (Bowker, 2006 ; Plantin, 2013). Pour les mêmes raisons que celles évoquées dans la partie de traitement, il serait aberrant de qualifier les données et de partager des contenus au formalisme et aux métadonnées dégradés.

Pour désenclaver les jeux de données, il convient de les signaler dans un espace dédié – sur une plateforme comme *Nakala* par exemple ou sur le site du projet, sous la forme d'hyperliens, avec pour chaque *dataset* un texte d'escorte décrivant les données. La page *OpenData* du site dédié au projet « Critiques d'Art francophones » montre des jeux dynamiques décrits dont les variables sont proposées dans des formats normalisés [8], par exemple la norme iso-3166-1 pour les noms de pays ou iso-8601 pour les dates : ce qui les rend directement exploitables dans n'importe quel logiciel de traitement.

Enfin, il est utile de préciser les conditions souhaitées de réutilisation des données : la licence d'exploitation. Les principales licences sont celles d'*Etalab* et *Creative Commons* [9], il faudra en en choisir une parmi celles proposées, en fonction des règles de crédit, de partage, et de réexploitation souhaitées (Fabry et al., 2017; Castets-Renard et Gandon, 2016 ; Maurel, 2018) [10].

CONCLUSION

Cet article s'est présenté comme un *vade-mecum* de l'usage des données dans les humanités numériques avec des questions sur leur production et les méthodes d'analyses qui y sont liées. Nous avons également proposé des points de vue sur le partage des données, leur ouverture et leur propagation avec une focale sur le web des données liées et l'interopérabilité.

Lors d'un projet en humanités numériques, nous avons montré que les *digital skills* nécessaires à la création et à l'exploitation d'un corpus sont nombreuses et variées. Il faut connaître les normes d'encodage, les formats de stockage, parfois les bases de données, il faut également avoir des connaissances en statistiques pour l'analyse. Toujours pour l'analyse, une capacité à représenter les données est fondamentale pour en permettre une appréhension fine, ce qui doit se faire dans le respect de l'éthique pour limiter l'impact visuel de la présentation – tout du moins ne pas essayer de tirer avantage des effets de la sémiologie graphique.

Enfin, nous avons introduit le concept de « courtoisie du *Fair DATA* » en humanités qui comprend une somme de bonnes pratiques facilitant l'accès à nos données : les rendre accessibles (légalement et physiquement) et repérables par les outils des chercheurs, qu'il s'agisse de notices bibliographies, iconographiques ou encore des données textuelles sémantiquement liées à des autorités et schémas consensuels.

NOTES

[1] Gratuit, accessible à l'url http://openrefine.org/

[2] Pseudo-science anthropométrique du XIX^{e} siècle qui se proposait d'évaluer les caractéristiques morales et intellectuelles d'un individu à partir de la forme de son crâne.

[3] Dans le cadre de l'analyse des systèmes de recommandation sociaux.

[4] Un exemple sera proposé plus loin.

[5] FAIR est l'acronyme de « *Findable, Accessible, Interoperable, Reusable* », notre expression « courtoisie du FAIR data » vient d'une analogie avec le *data steward*, l'intendant de la donnée dont le métier est d'effectuer un travail invisible mais indispensable : (avoir la courtoisie de) mettre à disposition de les données exploitables par un *data scientist*, un *data analyst* ou un chercheur.

[6] Voir http://critiquesdart.univ-paris1.fr/

[7] https://www.nakala.fr/

[8] http://critiquesdart.univ-paris1.fr/opendata

[9] https://www.etalab.gouv.fr/licence-ouverte-open-licence et https://creativecommons.org/licenses/?lang=fr-FR

[10] Pour plus d'informations le droit des données de la recherche, voir également le blog de Lionel Maurel : https://scinfolex.com/2016/11/03/quel-statut-pour-les-donnees-de-la-recherche-apres-la-loi-numerique/

RÉFÉRENCES BIBLIOGRAPHIQUES

ABITEBOUL Serge & DOWEK Gilles (2017). *Le temps des algorithmes*. Le pommier, 2017. 192p.

ARRUABARRENA Béa, KEMBELLEC Gérald & CHARTRON Ghislaine (2019). Data littératie & SHS : développer des compétences pour l'analyse des données, In. *Colloque CODATA-France, Data Value Chain in Sciences & Territories* 14–15 mars 2019 – Paris Val d'Europe, p.135-142

BASTIN Gilles & TUBARO Paola (2018). Le moment big data des sciences sociales. *Revue française de sociologie*, 2018, vol. 59, n° 3, p.375-394

BERTIN Jacques (1967, rév. 1973), *Sémiologie graphique. Les diagrammes. Les réseaux*. Les cartes, Paris/La Haye, Mouton ; Paris, Gauthier-Villars-Mouton, 452p.

BERTIN Jacques (1970). La graphique. *Communications*, 1970, n°15, p.169-185.

BOWKER Geoffrey C. (2005). *Memory practices in the sciences*. Mit Press Cambridge, MA, 2005.

BUSA Roberto (1974). *Index Thomisticus Sancti Thomae Aquinatis Operum Omnium Indices Et Concordantiae in Quibus Verborum Omnium Et Singulorum Formae Et Lemmata Cum Suis Frequentiis Et Contextibus Variis Modis Referuntur*.

BUSA Roberto (1980). The annals of humanities computing: The index thomisticus. *Computers and the Humanities*, 1980, vol. 2, n°14, p.83–90.

CARDON Dominique (2015). *À quoi rêvent les algorithmes. Nos vies à l'heure du Big Data*. Paris : Le Seuil, 2015, 112p.

CASTETS-RENARD Céline & GANDON Nathalie (2016). Open data des données de la recherche publique : entre réformes législatives et retour d'expérience sur un guide pratique à destination des chercheurs. *LEGICOM*, 2016, vol.1, n° 56, p.67-75

CHUDNOV Daniel et al. (2006). Introducing UnAPI. In. *Ariadne*, (48). Consulté à l'adresse

DENIS Jérôme & GOËTA Samuel (2017). La fabrique des données brutes : Le travail en coulisses de l'open data. In C. Mabi, J.-C. Plantin, & L. Monnoyer-Smith (Éd.), *Ouvrir, partager, réutiliser : Regards critiques sur les données numérique*s.

DRUCKER Johanna (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, vol. 1, N°5, p.1–21.

EBERLE-SINATRA Michael E. & VITALI ROSATI Marcello (2014). Histoire des humanités numériques. In *Pratiques de l'édition numérique* (p.49-60).

FABRY Cécilia, et al. (2017). Dossier « Publier des données liées et ouvertes en sept étapes ». *I2D Information, données documents*, 2017, n° 54, p.12–14.

HERNDON Thomas, ASH Michael & POLLIN Robert (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics*, 2014, vol. 2, n°38, p.257–279.

HAMMA Ken (2004). Becoming digital. *Bulletin of the American Society for Information Science and Technology*, vol. 5, n° 30, 1p.1–13.

LE DEUFF Olivier (2015). Les humanités digitales précèdent-elle le numérique ? In. *H2PTM'15: Le numérique à l'ère de l'Internet des objets, de l'hypertexte à l'hyper-objet*, p.421-432.

GANDON Fabien, CORBY Olivier & FARON-ZUCKER Catherine (2012). *Le Web sémantique : Comment lier les données et les schémas sur le Web* ? Paris : Dunod, 2012, 224p.

GINGRAS Yves (2018). Les transformations de la production du savoir : de l'unité de connaissance à l'unité comptable. *Zilsel*, vol. 2, n°4, p.139-152.

KEMBELLEC Gérald (2012). *Bibliographies scientifiques : de la recherche d'informations à la production de documents normés*. Lieu de soutenance : Université Paris VIII Vincennes-Saint Denis, décembre 2012, 422p.

KEMBELLEC Gérald (2017), Méthodes d'exploration systématiques et visuelles de corpus : enjeux et méthodes [Vidéo]. In *Séminaire de recherche Numerev*, Montpellier, Consulté à l'adresse

KEMBELLEC Gérald, FOURNIER Raphaël & CUBAUD Pierre-Henri (2018). L'histoire du Cédric : penser un dispositif archivistique en histoire des sciences. *Cahiers d'histoire du Cnam*, 2018, n° 7-8, p.133-153.

KEMBELLEC Gérald & BOTTINI Thomas (2017). Réflexions sur le fragment dans les pratiques scientifiques en ligne : entre matérialité documentaire et péricope. In *20° Colloque International sur le Document Numérique : CiDE. 20*.

KEMBELLEC Gérald, DESFRICHES-Doria Orélie & GISPERT Marie (2019, à paraître). Bibliographies de Critiques d'art francophones. *Ingénierie des systèmes d'information*, 2019, vol. 1, n° 24. Paris, Hermés Lavoisier.

LEMERCIER Claire & ZALC Claire (2008). *Introduction*. In : LEMERCIER Claire (dir.). Méthodes quantitatives pour l'historien (p. 3-7). Paris : La Découverte.

MAURE Lionel (2018). La réutilisation des données de la recherche après la loi pour une République numérique. In *La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques*. Presses Universitaires de Provence.

MERZEAU Louise (2009). Du signe à la trace. *Médium*, 2009, n°1, p.21–36.

MULLER Julien, SCHIAVON Martina & ROLLET Laurent (à paraître). Les procès-verbaux du Bureau des longitudes : un patrimoine numérisé (1795-1932). In REBUSCHI Manuel & BENZITOUN Christophe (dir.). *Les corpus en sciences humaines et sociales*. Nancy, 2019, Presses universitaires Nancy (Éditions universitaires de Lorraine).

PAVEL Ilarion & SERRIS Jacques (2016). *Modalités de régulation des algorithmes de traitement des contenus* (No. 2015/36/CGE/SG) (p. 63). Conseil général de l'économie, de l'industrie, de l'énergie et des technologies.

PRIME-CLAVERIE Camille & KEMBELLEC Gérald (2016). Dossier « Web de données et création de valeurs : le champ des possibles ». *I2D – Information, données & documents*, 2016, vol. 2, n°, p.28-69.

GITELMAN Lisa (dir.) (2013). *«Raw Data» is an Oxymoron*. Cambridge, MIT Press, 2013, 192 p.

REINHART Carmen M. & ROGOFF Kenneth S. (2010). Growth in a Time of Debt. In *American Economic Review*, 2013, vol. 2, n°100, p.573–78.

REINHART Carmen M. & ROGOFF Kenneth S. (2013). Growth in a Time of Debt: Errata. *Harvard University*.

REYMONET Nathalie et al. (2018). *Réaliser un plan de gestion de données « FAIR » : modèle*.

SCHNEIDER René (2013). Research data literacy. *European Conference on Information Literacy*, 134–140. Springer.

WEINBERG Sandy (2003). The Future. *In* WEINBERG Sandy (dir.). *Good Laboratory Practice Regulations*, Third Edition, Revised and Expanded (p.227–238). CRC Press.

WILKINSON Mark D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In *Scientific Data*, n°3., Nature Publishing Group

WYTHE Deborah. (2007). New technologies and the convergence of libraries, archives, and museums. In. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 2007, n°8, p.51–55.