L'ouverture des données dans un projet de recherche en scienceshumaines et sociales : le cas d'Adressbuch à l'Institut Historique Allemand de Paris

Evan Virevialle

Université Paris 1 Panthéon-Sorbonne
UFR 09 Histoire
Master Communication du Savoir, Technologies de la connaissance et Management de l'information

Mémoire

Présenté par : **Evan Virevialle**

L'ouverture des données dans un projet de recherche en sciences humaines et sociales : le cas d'*Adressbuch* à l'Institut Historique Allemand de Paris

Mémoire dirigé par :
Gérald Kembellec

Remerciements

Je tiens à remercier mon directeur de mémoire et aussi professeur dans le master CTM, Gérald Kembellec, qui m'a prodigué de nombreux conseils sur la marche à suivre pour rédiger mon mémoire. Je le remercie pour sa pédagogie et pour le partage de son expérience qui m'ont énormément aidé.
Je le remercie de m'avoir fait découvrir le domaine des Humanités Numériques, de m'avoir formé aux nouvelles techniques d'information et de documentation dans le cadre de mon poste d'assistant de recherche à l'Institut Historique Allemand à Paris.

Je remercie Mareike König, responsable du département Histoire Numérique et directrice adjointe de l'Institut, de m'avoir donné l'opportunité de poursuivre mon expérience au sein de l'Institut Historique Allemand de Paris. Je la remercie pour sa pédagogie, son accueil ainsi que pour les conseils prodigués.

Je remercie Mareike König et Gérald Kembellec de m'avoir offert l'opportunité de participer à des événements comme le dh nord 2021 et le datathon à l'IHA qui m'ont permis de découvrir les événements qui existent dans les SHS et les Humanités Numériques.

Je remercie le personnel de l'Institut Historique Allemand pour son accueil, sa pédagogie et sa bienveillance qui m'ont permis de m'épanouir dans mon travail au sein de l'Institut.

Enfin, je remercie, Luc Grivel, pour m'avoir permis d'intégrer le master CTM et d'avoir eu, par conséquent, la possibilité de réaliser un stage conventionné qui me guidera peut-être vers ma future profession et de suivre des cours aussi enrichissants que professionnalisants.

Introduction

Selon les principes et lignes directrices de l'Organisation de Coopération et de Développement Économique (OCDE. 2007) pour l'accès aux données de la recherche financée sur fonds publics, les données de la recherche sont définies comme « des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.». L'OCDE considère que les données de la recherche sont des faits qui permettent de valider les résultats de recherche. Nous pouvons voir que les données regroupent un grand nombre de catégories ce qui en fait un champ très vaste voir assez flou. De manière générale, les données publiques sont de plus en plus partagées et ouvertes, la recherche n'échappe pas à cette tendance.: elle doit alors se transformer et s'adapter. Dans certains domaines de recherche, la transition vers le partage de données est déjà faite (Jacquemin, Schöpfel, Fabre 2019), tandis que d'autres appartenant notamment aux sciences humaines et sociales sont encore réfractaires à ces pratiques (Kaden 2021).

J'ai pu étudier le thème de l'open data dans le cadre de mes enseignements du master CTM[1] notamment sur les *smart-cities,* mais aussi dans le droit du numérique avec les droits sur les données et leurs partages, la conception d'un site web ou encore dans le management de l'information. Ces enseignements et ces projets du master CTM m'ont permis de constater que l'open data progresse dans les grandes métropoles et les grandes entreprises. Grâce aux données ouvertes, les villes et entreprises peuvent se développer, mieux comprendre les besoins des personnes et apporter de nouvelles technologies dans notre quotidien notamment l'Intelligence Artificielle qui consomment beaucoup de données pour pouvoir être fonctionnelle.

Cependant qu'en est t-il de l'open data dans la recherche ? Pierre Mounier avance que la pratique scientifique en sciences humaines est transformée par les méthodes numériques (Mounier, 2018). Les chercheurs et chercheuses sont encouragés à utiliser des services et plateformes numériques et de nouvelles formes et méthodes d'écriture pour publier, éditorialiser, partager et valoriser les données de leurs recherches que ce soit de manière institutionnelle avec les plans d'action nationaux ou européens, mais aussi dans le cadre d'innovations éditoriales disciplinaires en SHS (Jacquemin et al., 2018; Sauret, 2020,

[1] Communication du savoir, Technologies de la connaissance et Management de l'information

Kembellec et Le deuff, 2022; Clavert et al. . 2021; Vitali-Rosati et al., 2020). Cependant, les pratiques de partage et de valorisation restent encore minoritaires (Kaden, 2021).

La recherche en histoire et plus généralement en humanités numériques étant nouvelle pour moi, je n'imaginais pas qu'il serait aussi intéressant de réfléchir aux meilleures manières pour diffuser, valoriser et rendre accessibles les données d'un projet de recherche. Mon stage, ainsi que mon poste d'assistant de recherche au sein de l'IHA, m'ont permis de constater que la question de l'ouverture des données était centrale dans le projet *Adressbuch.* C'est avec ce projet que j'ai commencé à m'intéresser à la question des données et à l'intérêt de les partager. Ce projet s'inscrit également dans une tendance grandissante ces dernières années, celle de mettre à disposition des chercheuses et chercheurs les données de la recherche après des tâches de qualification et valorisation des données. Les chercheuses et chercheurs ne se « contentent» plus de publier les résultats de recherche, ils vont également publier les données pour que nous puissions vérifier leurs conclusions ou utiliser leurs données dans de nouveaux projets. Les discussions que j'ai eues avec les chercheurs au sein de cet institut ainsi que les formations que j'ai pu donner sur des logiciels de traitement des données m'ont interrogé également sur la question des normes et pratiques de l'ouverture des données. Cependant, j'ai compris que tous les chercheurs n'ont pas la même finalité concernant leurs projets. Certains publieront uniquement leurs résultats, car c'est leurs objectifs ou bien parce qu'ils n'ont pas les moyens à leurs dispositions pour créer des jeux de données ou des dispositifs de consultation. Tandis que d'autres chercheurs vont estimer qu'il est nécessaire de publier les données pour qu'elles puissent être utilisées de nouveau ou même corrigées (Kaden, 2021).

Cet essor grandissant de l'open data se fait également en respectant les principes FAIR[2]. C'est-à-dire que les données doivent être faciles à trouver, accessibles, interopérables et réutilisables. Sans ces quatre principes, les données ne sont pas ouvertes.

Ce mémoire s'interroge sur la question des méthodes pour l'ouverture des données, nous allons étudier comment les chercheurs peuvent partager les données. Nous allons nous intéresser aux théories de partage des données dans le cadre des sciences humaines sociales.

Dans le cadre du partage des données, la question éthique des données se pose également, la question de la nature des données et si nous avons le droit de les partager.

À ces questions s'ajoute celle de la pratique : un autre but de ce mémoire est de montrer également, au travers d'une étude de cas, comment les chercheuses et chercheurs peuvent nettoyer, publier, valoriser et partager leurs données en Sciences Humaines et Sociales.

[2] FAIR : findable, accessible, interoperable, reusable

Dans un premier temps, nous étudierons l'aspect méthodologique du partage des données avec une remise en contexte du projet *Adressbuch* et une brève présentation de l'Institut Historique Allemand de Paris. Ensuite, nous nous concentrerons sur la pratique du partage des données et comment cela a été réalisé dans le cadre du projet *Adressbuch* avec une présentation technique du projet, du processus de nettoyage et de valorisation des données. Enfin, je finirai sur le retour d'expérience au sein de l'Institut Historique Allemand en dressant une autocritique de mon travail réalisé du projet.

I) Le projet et l'ouverture des données appliquée aux Sciences Humaines et Sociales

I.1) Contexte du projet

L'institut

Ce mémoire s'inscrit dans le cadre de mon expérience professionnelle au sein de l'Institut historique de Paris (IHA). Anciennement centre Allemand de recherches historiques, l'IHA est fondé en 1958 par le médiéviste Eugen Ewig avec le soutien du chancelier Allemand de l'époque Konrad Adenauer. L'institut historique allemand est sous l'égide de la fondation de droit public Max Weber depuis 2002. Cette dernière est une fondation de recherche en sciences humaines et sociales financée par l'État fédéral allemand dont le siège est situé à Bonn. L'IHA étudie l'histoire de l'Europe de l'Ouest de l'Antiquité tardive à aujourd'hui. Le rôle des chercheuses et chercheurs est de publier des articles, organiser des manifestations et s'impliquer dans des projets de recherches financés par l'institut. Certains projets permettent la création de bases de données consultables. Les départements sont organisés par époque historique, c'est-à-dire histoire médiévale, moderne et contemporaine. L'institut dispose également d'un département d'histoire de l'Afrique et d'un département d'histoire numérique qui m'emploie.

Le département Histoire Numérique

Un nouveau département a vu le jour en 2015, il s'agissait du département Humanités Numériques, récemment renommé Histoire Numérique. Ce département est né de la volonté de l'institut de s'engager dans le libre accès des données dans les années 2010 sous l'impulsion de la directrice de l'époque Gudrun Gersmann.

Le département Histoire Numérique permet d'effectuer des recherches sur les méthodes numériques, sur la révolution numérique en sciences humaines et sociales. Le département va également utiliser ses méthodes numériques pour équiper les projets de recherche du département, mais aussi les projets de recherche des autres départements. Le but de ce département est également de promouvoir les méthodes numériques au sein de l'institut. Cette passe par la formation aux outils numériques pour les collaborateurs de l'institut, enfin le département histoire numérique rédige des articles pour promouvoir la culture numérique et

notamment sur le libre accès des données de recherche. Cette politique de libre accès des données permet au département d'histoire numérique de conserver à long terme les données en respectant le principe FAIR[3]. Le département suit également une politique d'Open Source puisque les applications créées dans le cadre des projets de recherche sont mises à la disposition de tous sur des plates-formes collaboratives comme GitHub[4]. Les thématiques de recherche portent sur les transformations digitales dans les sciences humaines et sociales, l'influence de ces transformations sur le travail de recherche des historiens, enfin, la communication scientifique dans les médias sociaux.

Les collaborateurs

Le département se compose tout d'abord de la responsable de celui-ci, Mareike Köenig. Chercheuse à l'IHA en histoire contemporaine et histoire numérique, ses recherches portent sur l'histoire à l'ère du numérique, la communication scientifique dans les médias sociaux, l'histoire franco-allemande au XIXe siècle et les immigrés allemands à Paris au XIXe siècle.
Puis, Gérald Kembellec, chercheur détaché à l'IHA jusque décembre 2021, ses recherches se concentrent notamment sur les données liées, l'importance du partage des données de la recherche : du contexte de leur production à leur réutilisation ou encore l'interdisciplinarité en contexte des humanités numériques.
Enfin, j'ai aussi pu collaborer avec deux étudiants, Hippolyte Souvay, étudiant en Master humanités numériques à l'École des Chartes et Eike Löhden, étudiant en Informatique à l'Université de Marburg en Allemagne.

Rétrospective du projet et les missions

Mon stage à l'IHA consistait à mettre à jour le projet *Adressbuch.* Ce projet se base sur un document unique : *Adressbuch der Deutschen von Paris für das jahr 1854*. Cet ouvrage est unique, par les informations qu'ils renferment, mais aussi par sa rareté. En effet, nous connaissons seulement 4 exemplaires de cet ouvrage, dont un conservé à la bibliothèque historique de la ville de Paris. C'est cet exemplaire que nous pouvons consulter sur le site *Adressbuch.*

[3] Cf note de bas de page 2
[4] Voir le GitHub de l'IHA : https://github.com/dhi-digital-humanities/

Cet annuaire de 248 pages[5] rassemble les adresses de 4 772 particuliers allemands ainsi que des entreprises dans Paris et sa banlieue. Cet ouvrage renferme de nombreuses informations. On y trouve des personnes de la classe populaire comme des nobles. On y trouve également des civils comme des militaires décorés de la Légion d'honneur. On dénombre, par exemple, 141 entreprises, 113 personnes issues de la noblesse ou encore 286 femmes. Ce document nous permet de nous faire une idée de la répartition de la population allemande à Paris au XIXe siècle. En effet, les 4 772 allemands figurant dans le livre font partie des 12 245 allemands reconnus officiellement dans le recensement de 1851. Ainsi, un tiers des allemands vivant à Paris en 1854 sont répertoriés dans cet ouvrage. En ce qui concerne l'auteur F. Kronauge, nous savons seulement qu'il était professeur de langue et qu'il vivait rue Richelieu à Paris.

Le projet a débuté au début des années 2000 et s'est finalisé en 2006. Il a notamment été financé par la fondation Gerda Henkel (Düsseldorf) et la société des amis du DHIP[6]. L'objectif principal était de créer une base de données rassemblant les informations contenues dans l'annuaire et de les mettre à disposition des chercheuses et chercheurs ainsi que des généalogistes. La première interface développée avec le logiciel FileMaker permettait de consulter le nom, prénom et adresses des personnes ainsi que leurs métiers ou encore leur statut social. Il était également possible de rechercher dans cette interface par le nom d'une personne, mais aussi par thématique, pour trouver les restaurateurs allemands par exemple, ou encore toutes les femmes dans *Adressbuch.* Cependant certaines fonctionnalités ne pouvaient pas être développées lors de la première phase. Il était trop complexe voir impossible de développer une cartographie interactive pour visualiser les Allemands sur une carte historique de Paris en diachronie avec une carte moderne. Les données brutes étaient placées sous copyright et n'étaient pas proposées en accès libre. Cette décision peut s'expliquer par le fait que dans les années 2000 peu de personnes se souciaient de la question de l'ouverture des données. Également, la question de la propriété intellectuelle se posait, les chercheurs pouvaient être réticents à laisser leurs données être utilisées par n'importe qui.

Aujourd'hui le projet fait l'objet de nombreuses consultations par des chercheurs et demandes d'informations de la part de généalogistes allemands qui mènent des recherches prosopographiques. Le projet *Adressbuch1854* a été relancé en 2020 dans le but de fournir une interface en Open Source et Open Data pour mettre à disposition des chercheuses et chercheurs les données du projet *Adressbuch.* Depuis 2021, le projet s'est enrichi d'un partenariat avec

[5] En réalité, une erreur de pagination amène cet ouvrage à 242 pages. Cette erreur figure dans les autres exemplaires.

[6] Deutsches Historisches Institut Paris : Institut Historique Allemand de Paris

l'*Institut für Digital Humanities zu Koeln* (IDH) qui s'occupe de l'hébergement du site web et de la base de données. Les données brutes sont déposées sur le serveur de données de recherche Zenodo et le code du nouveau site est disponible sur les plateformes Github de l'IDH et de l'IHA[7]. Github nous permet de faire du versionning pour le site et de voir les étapes de mise à jour du site.

Mon stage ainsi que mes missions s'inscrivent dans cette phase du projet. Cette deuxième phase avait pour but de réaliser ce qui n'avait pas pu l'être du fait des contraintes techniques lors de la première phase. Avec l'essor des plates-formes d'open data, mais également d'open source, les défis techniques du passé sont désormais réalisables aujourd'hui. En plus de ces missions, nous devions moderniser l'application web devenue obsolète et contraignante pour pouvoir partager les données aux utilisateurs. L'interface FileMaker d'origine ne correspondait plus aux besoins des utilisatrices et utilisateurs en termes d'accessibilité et d'ergonomie.

Il a fallu également penser la base de données en diachronie, car beaucoup de mutations administratives ou topographiques notamment ont eu lieu depuis la publication d'*Adressbuch* en 1854. Ainsi, il fallait retrouver les métiers modernes correspondant à ceux du XIXe siècle ou encore effectuer la correspondance avec les arrondissements qui ont été transformés en 1859[8]. De plus, Paris a subi de profondes mutations avec les travaux de Haussmann dans les années 1850/1860, il fallait trouver les correspondances entre les rues anciennes et nouvelles.

Tout d'abord, les données ont été nettoyées et enrichies avec le logiciel open source *OpenRefine*[9], ce dernier permet notamment la réconciliation des données avec des bases de données externes ou encore des services de notices d'autorité[10]. Cela permet à des usages plus larges selon les besoins des historiennes et historiens, mais aussi pour les généalogistes ou encore les visiteurs érudits. Le fichier tabulaire initial au format CSV[11] a été fragmenté en plusieurs fichiers au format SQL[12] pour remplir les tables de notre base de données. Ces derniers ont ensuite été insérés dans une base de données MySQL[13] gérée avec l'interface graphique phpMyAdmin[14] pour exploiter efficacement les données. Ensuite, le document primaire a été également mis à la disposition des utilisateurs sur la nouvelle plate-forme.

---

[7] Github: Adressbuch1854

[8] Loi du 16 juin 1859 (création des vingt arrondissements). – Décret impérial du 31 octobre 1859 (dénomination des nouveaux arrondissements)

[9] https://openrefine.org

[10] Par exemple : Wikidata, VIAF, GND

[11] Comma separated values

[12] Structured Query Language

[13] Système de gestion de bases de données relationnelles en SQL

[14] https://www.phpmyadmin.net/

La bibliothèque historique de la ville de Paris nous a permis de numériser le document original. Il est important pour les chercheuses et chercheurs de pouvoir consulter le document primaire et vérifier les données du projet. On ne peut pas seulement publier nos données, il faut qu'elles puissent être vérifiées avec la source d'origine.

Les données sont désormais consultables sur une interface web conçue avec le framework PHP[15] open source Cakephp qui nous permet de concevoir une interface *MVC : model view controller* pour interagir avec la base de données. Cakephp est maintenu par une communauté de développeurs et disponible en Open Source ce qui rend l'application pérenne. Cette dernière tente de répondre aux besoins des utilisateurs qui sont notamment des chercheurs et des généalogistes allemands. Une cartographie interactive est également disponible sur la nouvelle interface qui permet de visualiser la position des allemand en 1854 dans Paris. Cette dernière est réalisée en JavaScript à l'aide du logiciel open source Leaflet. Cette carte fonctionne également avec les données vecteurs du projet ALPAGE[16] pour les anciens arrondissements, les données vecteurs de Paris OpenData[17] pour les arrondissements nouveaux et la carte de l'état-major de 1822-1866 disponible sur l'IGN[18] avec le géoportail du gouvernement. Une autre nouveauté du projet est de proposer une recherche détaillée pour trouver rapidement les personnes figurant dans *Adressbuch*. Plusieurs champs sont disponibles pour cette recherche : le nom, prénom, sexe, rue, arrondissement ou encore rang dans la Légion d'honneur.

Une océrisation des pages d'Adressbuch a été réalisée pour permettre de consulter uniquement le texte des pages. Bien que les résultats comportent du « bruit », ils peuvent être consultés en appui aux numérisations.

Dans le cadre de ce projet, un datapaper a été réalisé, celui-ci nous a permis de parler des données, de leur structure, de leur partage et réutilisation. Il a été coécrit par Mareike Koenig, Gérald Kembellec et moi-même.

Un datathon a été aussi organisé en partenariat avec le DFK[19], cet événement de 3 jours, qui rassemble des historiens et des personnes travaillant dans le traitement des données, a permis de travailler de façon collaborative sur le projet Adressbuch et sur d'autres projets pour développer des solutions pour interpréter les jeux de données disponibles.

[15] PHP : Hypertext Processor
[16] AnaLyse diachronique de l'espace urbain PArisien : approche GEomatique
[17] Arrondissements — Paris Data
[18] https://www.geoportail.gouv.fr/donnees/carte-de-letat-major-1820-1866
[19] DFK : deutsches forum für kunstgeschichte / centre allemand d'histoire de l'art

La mise à disposition de la cartographie historique de Paris sur le site de l'IGN, les quartiers et arrondissements historiques de Paris récoltés par le projet ALPAGE géré par le LAMOP[20] de l'université Paris 1 Panthéon-Sorbonne a permis la création de la cartographie interactive sur la nouvelle plate-forme. L'open source y joue également un rôle prépondérant puisque sans le logiciel open source Leaflet, qui s'appuie sur OpenStreetMap l'alternative open-source de Google Map, la réalisation de la cartographie aurait été plus complexe.

Le projet Adressbuch s'inscrit dans une évolution d'Open Access qui est très intéressante, car elle soulève des questions pragmatiques sur les normes, encodages et formats qui sont sous-tendues par l'écosystème du libre, mais aussi des questions d'ergonomie et d'usabilité. En effet, le projet Adressbuch a fait le choix d'ouvrir ses données alors que nous aurions pu nous contenter d'une restauration de la plate-forme et laisser les données en licence privée. Plusieurs possibilités sont disponibles pour avoir accès aux données du projet Adressbuch. Tout d'abord, il est possible de télécharger l'ensemble des données soit les 4 772 personnes dans 4 formats : JSON[21], CSV, SQL et XML[22]. Nous avons fait le choix de ces formats pour permettre la meilleure compatibilité pour la réutilisation des données. Ensuite, il est possible de télécharger les données par index. Dans ces index, nous pouvons choisir de télécharger ces données par lot de 20, qui correspond au nombre de personnes affichées par page, et personne par personne. Les formats pour ces jeux de données sont le JSON et XML. Enfin, les données ont été déposées sur Zenodo avec les numérisations d'Adressbuch en haute définition et définition standard. Nous avons le dépôt initial contenant la première extraction de données provenant de la plate-forme sous FileMaker et un dépôt qui contient les données modélisées contenues dans la nouvelle plate-forme.

I.2) Contexte de recherche

[20] LAMOP : Laboratoire de Médiévistique Occidentale de Paris
[21] JavaScript Object Notation
[22] Extensible Markup Language

Théories sur le partage des données en SHS

Deux déclarations écrites au début du XXIe siècle ont permis la libération des données de la recherche dans le contexte institutionnel et scientifique. Tout d'abord la déclaration de Budapest en 2002, le *Budapest Open Access Initiative* permet de le libre accès aux publications de recherche. Le 1er et le 2 décembre se tient le congrès de Budapest où l'organisateur, George Soros, réunit les institutions les plus avancées dans le domaine des archives ouvertes. Ce congrès a eu pour but de réfléchir aux pratiques sur l'Open Access. Ils lancent un appel qui est publié dans la presse numérique le 14 février visant à inciter les chercheurs et chercheuses à mettre à disposition leurs publications scientifiques. Cette initiative est basée sur deux axes. Le premier est de proposer à l'auteur de mettre en ligne ses publications sur des archives ouvertes créées par des institutions. Le second est de proposer des revues alternatives en libre accès. On va alors soit archiver ses publications de recherches, soit les rééditer dans un format numérique en libre accès. La déclaration de Budapest est enrichie un an plus tard avec la déclaration de Berlin. La déclaration de Berlin sur le libre accès à la connaissance est un texte où les signataires réclament le libre accès de la littérature scientifique internationale, les données et les logiciels qui ont permis de produire cette connaissance. Presque 20 ans plus tard, en 2021, le texte a été ratifié par plus de 680 universités. La déclaration de Berlin dépasse la déclaration de Budapest qui demandait le partage des articles scientifiques. Les contributions au libre accès concernant les résultats de la recherche, de données brutes, de métadonnées, de documents sources et de sources multimédias.

Nous avons ici, deux déclarations fondatrices de l'ouverture des données dans la recherche. En France, l'accès libre aux données de la recherche scientifique fait partie des objectifs de la recherche selon l'article L112-1 du code de la recherche[23]. L'alinéa e de cet article stipule que « Les travaux de recherche menés dans le cadre de ces coopérations sont, en l'absence de clauses contraires, rendus publics et accessibles ». On comprend ici la volonté en France de rendre accessibles les résultats de la recherche publique. Nous avons aussi le plan d'action national de 2018-2020 avec l'engagement 18 sur la science ouverte qui vise à « construire un écosystème dans lequel la science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus intègre, plus rapide et d'accès plus universel ». Le but du gouvernement et plus particulièrement du ministère de l'Enseignement supérieur, de la recherche et de l'innovation est de construire un écosystème de la science ouverte dans lequel les données

[23] https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000027747800/

seront ouvertes, transparentes et accessibles. Ce n'est pas sans rappeler les principes FAIR qui représentent les bonnes pratiques de la recherche et sont encouragés pour le pilotage des données.

Les déclarations de Berlin et de Budapest définissent au niveau mondial l'ouverture des données. Le plan d'action du gouvernement ainsi que le code de la recherche définissent au niveau national ce que doit être le partage des données scientifiques de la recherche.

Au niveau des autorités européennes, certaines actions ont également été menées pour partager les données collectées par les pouvoirs publics. Ainsi en 2003, le parlement européen vote la directive *Public Sector Information* pour ouvrir les données institutionnelles afin qu'elles soient réutilisées. En 2011, en France, la création d'Etalab permet le partage et l'ouverture des données publiques issus des établissements publics. Enfin, Etalab est appuyé par la *Loi pour une République numérique* votée en 2016 qui permet la publication systématique des données collectées par les pouvoirs publics (Schöpfel, Jacquemin et Fabre. 2019).

Cependant, nous pouvons voir que les avis divergent selon les domaines dans lesquels l'ouverture des données a lieu. Dans des domaines de recherche tels que la biologie, la météorologie ou encore le médical, l'ouverture des données se fait depuis de nombreuses années et même est nécessaire pour faire avancer la recherche. Dans les sciences humaines et sociales, il est plus dur de briser la « tradition ». Il y est d'usage de ne pas partager les données de la recherche, mais de seulement partager les résultats de recherche. De manière générale, les chercheurs dans les sciences humaines souhaitent conserver cette liberté de choix de partager ou non des données. Néanmoins, si nous partons du principe que les chercheurs sont des fonctionnaires du secteur public et qu'ils effectuent des recherches à partir de financement publiques alors il devient évident que les données doivent être partagées (Jacquemin, Schöpfel, Fabre,2019). Selon l'article de Bernard Jacquemin, Joachim Schöpfel et Renaud Fabre : « Il est finalement nécessaire de prévoir les conditions de l'ouverture des données, qui devrait être systématique dès lors que rien ne s'y oppose ». Ici, l'avis est clairement tranché, si d'un point de vue moral et éthique, les données ne représentent pas de danger pour qui que ce soit alors les données doivent être partagées.

- Mais alors quelles sont ces données ?

Les données provenant des Humanités peuvent avoir différents formats, qu'ils soient textuels iconographiques, des animés, des graphismes, des jeux de métadonnées ou encore des notices bibliographiques. Le point commun de ces sources c'est qu'elles ont été collectées manuellement avant d'être transférées vers des processus automatiques pour un éventuel nettoyage ou enrichissement qui fournira par la suite un travail de qualité (Kembellec, 2019).

- Comment les données sont collectées ?

Les données en histoire ne sont pas numériques en règle générale puisque les historiens travaillent sur des sources historiques, comme des archives, des ouvrages, des manuscrits, des actes par exemple. Pour qu'ils puissent avoir un traitement informatique, ces documents doivent passer par une phase de numérisation des données. Les données sont transmises à la machine avec des instructions données par le chercheur ou/et l'ingénieur: Les données seront de qualités que si les instructions sont bonnes et si les données initiales sont de bonnes qualités.[24] Bien que l'automatisation des processus pour la qualification et le nettoyage des données se développe de plus en plus, nous ne pouvons le substituer en histoire totalement à un travail manuel, un travail humain de vérification et validation.

Le point crucial est alors de collecter les données de la bonne manière. Pour effectuer des transcriptions automatiques sur des manuscrits anciens, il faut les scanner dans un certain format comme le TIFF ou le JPEG dans une haute résolution comme du 600 dpi ce qui permet d'avoir de meilleurs résultats lors du traitement des données. Si nous souhaitons proposer ses scans à la consultation sur un site web, nous devons nous tourner vers les standards du web qui suggèrent un affichage en haute définition d'au moins 300 dpi ainsi que des prévisualisations d'au moins 72 dpi.

- Comment partager les données ?

Alors que les avis divergent sur le principe même de partager les fruits de ses recherches, la façon dont les données doivent être partagées est un point de vue également débattu. En effet, deux visions s'opposent, la première consiste à partager les données avec le maximum de documentation et de proposer des jeux de données clairs et concis avec une « normalisation » maximale pour permettre la reproductibilité des données. Cette structuration des données avant le partage permet de pouvoir les reproduire à l'identique comme sur l'interface de consultation d'origine, mais il faut rester prudent, car cette structuration pourrait empêcher une liberté totale sur la manipulation des données. Ce qui nous amène à une deuxième vision qui défend la «brutification» des données. C'est-à-dire le fait de minimiser la transformation des données avec peu ou aucune structure qui permettrait une ouverture maximale pour leurs manipulations. De mon point de vue, une donnée, lorsqu'elle est partagée, doit être documentée et structurée, car cela permet une meilleure visibilité, une meilleure compréhension de la donnée et par conséquent sa réutilisation sera plus simple. Dans les deux cas que nous venons de voir, il est

[24] Voir Garbage in Garbage out (déchêts en entrée déchêts en sortie) : https://journals.sagepub.com/doi/full/10.1177/1833358318774357

nécessaire que les données soient accompagnées de la source d'origine. Gérald Kembellec propose également 3 catégories de structuration des données dans lesquelles tout le monde pourrait trouver la structure qui lui convient (Kembellec 2019).

- Faiblement structurées

On y trouve des données qui peuvent être lues par un simple éditeur de texte, ce seront alors des formats ouverts et libres comme le CSV, TSV ou TXT. Les deux premiers disposent d'une mise en forme minimale avec des colonnes et un séparateur entre les données comme une tabulation ou une virgule tandis que le dernier ne présente aucune mise en forme, il est le plus brut de tous les formats.

- Peu structurées

Cette catégorie nécessite au moins un éditeur de code comme Sublime Texte ou Notepad++ pour être prise en charge et comprendre leur structure. Dans cette catégorie, on peut retrouver le JSON qui est un format encore assez simple, mais qui peut devenir très complexe selon la structure et le volume de données. Si on exporte, une base de données relationnelles au format JSON, le fichier de sortie peut devenir assez indigeste pour celui qui tente de comprendre son architecture.

- Fortement structurées

Ce seront les formats de données qui nécessitent un logiciel spécifique pour être lus et interrogeables. Si je souhaite lire un fichier SQL, j'aurais besoin d'une interface graphique de base de données relationnelles pour comprendre la base de données que j'ai téléchargée. Si je souhaite lire un fichier au format shapefile, j'aurais besoin d'un logiciel de cartographie comme QGIS pour comprendre de quoi il s'agit, par exemple.

Dans un projet comme *Adressbuch* à l'Institut Historique Allemand de Paris, peu importe à quel point les données sont structurées et documentées, les chercheurs et les généalogistes vérifieront toujours avec le document primaire. La véracité des données sera remise en cause et sera vérifiée avec le document primaire qui fait loi.

Les données doivent être signalées distinctement sur un portail dédié, sur une plateforme de dépôt, on peut citer Nakala du service Humanum ou encore Zenodo, ou sur le site du projet avec un onglet dédié au téléchargement des datasets par exemple. Cette dernière méthode n'est pas une bonne pratique, pour des raisons de pérennité et de maintenabilité (les données risquant alors de ne pas survivre à la durée de financement d'un projet et le site a des risques de ne pas être maintenu ainsi que les données qui y figurent). Le but étant ici de rendre les données «Findable» selon les principes FAIR.

Point de vue des chercheurs

Nous l'avons vu précédemment que le partage des données varie selon le domaine d'étude dans lequel les données sont collectées. De plus, des chercheurs ont dressé une liste des raisons pour lesquelles les chercheurs ne publient pas les données et que nous pouvons regrouper dans des grandes catégories (Kaden. 2019).

- Investissement

Les chercheurs souhaitent consacrer du temps à leurs recherches plutôt qu'à la façon de les publier, de les ouvrir et de les diffuser. De plus, la publication des données n'a pas été prise en compte dans la création du projet et il serait trop long de préparer un plan de gestion de données surtout si on ne dispose pas des ressources humaines nécessaires.

- Loi sur la protection des données

Les données de la recherche contiennent des données sensibles et ne peuvent pas être partagées.

- Équipements Institutionnels

Les institutions propres aux chercheurs n'offrent pas le soutien et les moyens pour partager les données. Je ne suis pas tout à fait d'accord avec la validité de cette raison pour ne pas partager les données. En effet, si notre institution n'est pas capable d'aider pour le partage des données, il existe de nombreux outils qui permettent de diffuser ses données sans aucune compensation financière. Nous pouvons citer par exemple les dépôts Zenodo, Nakala ou encore le site d'hébergement de gestion de versions, qui peut faire office de dépôt de données, Github.

- Conditions institutionnelles

Nous pouvons nous retrouver face à des contraintes institutionnelles qui nous empêchent de partager nos données ou alors le partage des données est permis seulement si des conditions strictes sont remplies.

- Possibilités et compétences de partage

Les chercheurs souffrent d'un manque de compétence en termes de littératie numérique et face à ce manque de compétence, ils ne partagent pas les données, car ils ne savent pas où les publier en toute sécurité et comment les rendre visibles. Ce n'est plus une question de volonté, mais bien une question de compétence et de formation qui empêche les chercheurs de publier les données. Cette absence de méthodologie et de formation est amplifiée avec l'exigence des normes scientifiques qui peut être complexe. La publication des données selon les normes scientifiques est une charge supplémentaire importante de travail en opposition aux potentiels

bénéfices. Ce déséquilibre entre le travail et le bénéfice n'incite pas les chercheurs à publier leurs données.

- Attitude personnelle

Ici, j'aimerais montrer que non seulement le domaine d'étude influence sur la décision de partager, mais surtout que c'est l'attitude personnelle propre à chaque chercheur qui fait que les données de la recherche. On relève notamment que certains chercheurs ne sont tout simplement pas intéressés par la science ouverte, les chercheurs souhaitent contrôler leurs données, la divulgation des données est considérée comme une atteinte à la liberté académique personnelle ou encore les chercheurs craignent que la faiblesse de l'analyse et de la collecte des données deviennent visibles (Kaden. 2019). Ce que nous pouvons relever avec l'étude de Ben Kaden c'est que les chercheurs ont surtout peur de se lancer dans la publication des données par manque de formations et de méthodologie.

- Droit d'Auteur et d'édition

Les chercheurs sont liés avec des éditeurs ou des partenaires commerciaux ce qui fait qu'ils ont des autorisations limitées pour le partage des données.

- Liberté académique

La science ouverte ne doit pas être perçue comme une contrainte ou même une obligation, les chercheurs doivent décider par eux-mêmes s'ils veulent rendre les données de recherche accessibles. On pourrait alors imaginer que les récentes décisions gouvernementales comme le plan d'action de 2018-2020 qui encourage l'ouverture des données de la recherche sont peut-être mal perçues par certains chercheurs, voire vues comme une obligation et alors interprétées comme une atteinte à la liberté de décision des chercheurs.

Cette étude montre que les chercheurs disposent de beaucoup de raisons, justifiées pour certaines et non justifiées pour d'autres, pour se montrer réfractaires à la publication et la diffusion des données de la recherche. La question qu'il faut se poser est de savoir comment remédier à cette appréhension face au partage des données pour que les chercheurs soient de plus en plus nombreux à diffuser librement leurs données de recherche.

Intéressons-nous maintenant aux entretiens semi-directifs réalisés par Violaine Rebouillat dans le cadre de sa thèse « Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs ». Tout d'abord, nous pouvons voir que la définition même de « donnée de la recherche » reste assez vague dans l'esprit des chercheurs. Nous pouvons y voir que le sens d'une donnée diffère dans le domaine où elle est utilisée. De plus, selon les résultats des entretiens menés sur 57 chercheurs, le terme « donnée » est plus répandu et plus utilisé dans les sciences techniques et la médecine que dans les sciences humaines et sociales.

Toujours selon les entretiens, il semblerait que la donnée renvoie toujours à l'idée de quantité et à un aspect numérique. Cette association entre la donnée, le quantitatif et le numérique peut expliquer que les chercheurs en sciences humaines ne trouvent pas forcément pertinent d'utiliser ce terme s'ils n'ont pas d'analyse statistique à produire.

Dans ces entretiens, il existe plusieurs raisons qui sont données pour justifier le partage des données dans la recherche.

Nous en avons parlé précédemment, plusieurs chercheurs interrogés lors de ces entretiens trouvent logique de partager leur données, car leurs recherches sont financées par des fonds publics et ils sont eux-mêmes des fonctionnaires (Jacquemin, Schöpfel, Fabre, 2019; Rebouillat, 2019)

Autre point intéressant de ces entretiens, ils révèlent que plusieurs chercheurs souhaiteraient publier leurs données pour lutter contre la fraude scientifique et la justesse de ces dernières. Cela permet au lecteur de vérifier par lui-même que les données sont exactes. On pourrait exiger lors de la publication de l'article le lien vers le dépôt de données qui ont servi de matériau à l'article. Cependant une personne mal intentionnée pourrait trafiquer également le dépôt et si ce n'est de reconstituer le jeu de données à partir des sources : c'est la procédure que les articles de type « data paper » est censé décrire, il n'existerait pas de réels moyens pour déceler la fraude.

La question des normes et standards dans la recherche

La relation entre la recherche et la technique est essentielle dans le processus d'un projet de recherche pour ouvrir les données. J'ai eu l'occasion d'avoir justement cette relation où je me retrouvais avec les besoins des historiens, que je connais un peu, car je suis issu d'une formation historique, et la question de la faisabilité technique pour réaliser ces besoins. Cette expérience m'a permis d'abord de voir que nous devons nous adapter aux besoins des historiens qui peuvent être spécifiques. Je prends pour exemple le carnet d'adresses des Allemands à Paris où je pensais que pour la nouvelle interface de consultation il suffisait de créer des index pour consulter les personnes recensées dans Adressbuch. La demande qui m'a été faite dans ce projet est de proposer les numérisations d'origine en plus de la consultation de la base de données pour permettre aux utilisateurs de consulter une version scannée du document d'origine et ainsi de s'assurer de la véracité des informations contenues dans la base de données. Ici nous devons alors composer avec les besoins des historiens à savoir le besoin de consulter le document d'origine. Nous pouvons aussi prendre pour exemple la base de données du projet

Adressbuch où non seulement elle doit pouvoir fonctionner de manière efficace d'un point de vue technique avec le dispositif de consultation, mais elle doit aussi répondre aux enjeux historiques du projet, notamment l'étude diachronique des Allemands. C'est-à-dire que notre projet étudie les Allemands à la période où le bottin a été publié, mais également à notre époque avec la correspondance des rues et des arrondissements de Paris. Il faut alors réfléchir avec les chercheurs et les techniciens à la disposition des données dans la base avec les tables et les tables de jointure pour pouvoir ensuite interagir efficacement avec nos données depuis notre dispositif de consultation. À l'inverse, l'historien doit comprendre que tout ne peut pas être réalisé, du moins certaines contraintes peuvent apparaître et empêcher de réaliser un dispositif répondant à tous les besoins des chercheurs. Nous pouvons citer trois grandes contraintes :

- Le temps qui est une des grandes contraintes des projets de recherche, si certaines tâches sont complexes à réaliser et que le temps disponible est trop court alors elle ne sera pas réalisée.
- L'argent, on dit que l'argent est le nerf de la guerre, il pourrait être également celui de la recherche, car sans financement le projet ne peut pas aboutir et alors certaines demandent techniques ne peuvent pas aboutir surtout si les chercheurs font appel à des prestataires, ils seront obligés de faire des concessions afin de faire baisser la facture.
- Les contraintes techniques. Tout est possible, néanmoins, les contraintes imposées par un projet de recherche limitent les réalisations techniques.

Je peux prendre pour exemple le projet Adressbuch : lors de sa première phase au début des années 2000 si nous avions eu à réaliser une carte interactive cela aurait été trop compliqué au niveau technique, car des outils comme Leaflet ou OpenStreetMap n'existaient pas encore. Ici, nous avons alors cet angle où les chercheurs sont obligés à leurs tours de se plier aux différentes contraintes qu'elles soient d'ordre technique, financier ou temporel pour faire aboutir le projet à son terme.

Nous avons relevé que la relation entre technicien et chercheur peut-être complexe lors de la réalisation technique du projet de recherche et celle-ci peut s'accentuer sur le choix des langages, des formats et des standards pour mener à bien le projet. Nous pouvons notamment sur les appels à projets, sur les offres d'emplois où les chercheurs cherchent des ingénieurs d'étude avec des compétences spécifiques. Ainsi, on retrouve souvent sur les offres récentes qu'un ingénieur d'études doit être capable de maîtriser des frameworks comme la suite Omeka, il doit connaître des langages d'analyses de données comme le Python ou il doit connaître des langages d'encodage comme le XML-TEI. Alors, c'est une liste non exhaustive des compétences demandées pour un ingénieur d'études dans la recherche et elle dépend

évidemment du projet de recherche sur lequel il est censé travailler. Cependant, ce que je veux montrer ici c'est que parfois les chercheurs vont choisir des langages comme étant la référence alors qu'ils en existent des multitudes: Pour l'analyse des données, j'ai cité le langage Python, mais le langage R est un bon langage pour l'analyse des données, Omeka est un bon framework pour la gestion de bibliothèques numériques, mais ils existent de très bons frameworks en PHP qui peuvent faire aussi bien que le logiciel Omeka, certains chercheurs ne jurent que par Microsoft Excel pour nettoyer les données alors que le logiciel open source Open Refine est plus puissant pour nettoyer les grandes masses de données et conçu pour cela. Ce sera alors aux ingénieurs d'études par la suite de proposer plusieurs solutions aux chercheuses et chercheurs pour réaliser le projet de recherche et démontrer pourquoi tel logiciel ou tel format est le meilleur pour ce projet. Nous avons alors des projets en Sciences Humaines et Sociales qui font la jonction avec les Sciences de l'Information de la Communication comme le projet ALPAGE où plusieurs laboratoires de SIC sont intervenus sur ce projet pour développer la cartographie de Paris du 13e au 19e siècle.

I.3) L'éthique des données

La question de l'éthique des données dans les SHS

Lorsque nous partageons les données d'un projet de recherche, il faut réfléchir à la dimension éthique des données de la recherche. Alors que la recherche française encourage une gestion FAIR des données,[25] nous pouvons nous poser la question de la dimension éthique dans la gestion FAIR des données. Le CNRS avertit dès 2015 sur la gestion FAIR des données : « toutes ces consignes générales peuvent paraître en opposition avec les restrictions légales formulées au nom du respect de la vie privée, du droit d'auteur, de l'obligation de secret ou de la sécurité » (Comets, 2015; Schöpfel, Jacquemin, Chaudiron, Kergosien. 2018). Lors de l'ensemble du cycle de vie des données il faut penser à cet aspect éthique des données, car la recherche scientifique peut-être en conflit avec cette dimension éthique des données. Schöpfel, Jacquemin, Chaudiron et Kergosien déterminent six facettes pour identifier la relation entre les données de la recherche et l'éthique.

Les six facettes qui correspondent aux points les plus importants sont :

- Le plan de gestion

[25] Voir plan d'action national 2018-2020, engagement 18

On va intégrer la dimension éthique dans le plan de gestion des données. Ainsi dans le data management plan actuel du programme européen H2020 évoque la dimension éthique des données dans la gestion FAIR des données. On va alors s'intéresser aux aspects éthiques et justifier s'il existe des problèmes éthiques pouvant impacter sur le partage des données ou encore justifier l'absence de la libre diffusion des données par des raisons éthiques ou bien décrire les mesures de sécurité mises en place pour partager les données sans toutefois évoquer leurs natures. Les plans de gestion peuvent être considérés comme des compléments aux comités éthiques pour démontrer la prise des chercheuses et chercheurs sur comment la gestion des données va intégrer la dimension éthique (Schöpfel, Jacquemin, Chaudiron, Kergosien. 2018).

- Les données à caractère personnelles

Selon la définition de la CNIL[26], les données à caractère personnel sont « toute information relative à une personne physique susceptible d'être identifiée, directement ou indirectement »[27]. Les données personnelles sont un des problèmes principaux qui empêchent les chercheurs de diffuser librement les données avec la réglementation juridique et éthique qui n'incitent pas les chercheurs à trouver des solutions pour partager librement les données et ne vont pas au bout de leurs démarches d'open data. Le problème des données personnelles ne se pose pas dans tous les domaines de recherche, ce problème sera notamment présent dans les données biomédicales par exemple, car le fait que ce soit des données sur la santé pourrait amener à l'identification d'une personne. Dans les sciences humaines et sociales comme l'histoire, ce problème se pose moins, car on a affaire à des archives, des documents historiques dont nous avons les autorisations pour l'exploitation et la diffusion, exception faite à l'histoire moderne et contemporaine où les données peuvent être relatives à des personnes en vie.

La loi « pour une république numérique » de 2016[28] affirme que les données de la recherche sont libres de diffusion si ces dernières ne sont pas protégées par une réglementation ou un droit particulier. Ainsi, pour gérer un projet de recherche, les données posent des problèmes. Il faut être en conformité avec la loi Informatique et Libertés de 1978[29], l'autorisation et la publication des données, il faut être en conformité avec la réglementation sur les données liées à la santé puis il faut une autorisation pour la réutilisation des données. Autrement dit, autant d'obstacles qui peuvent freiner la libre diffusion des données dans certains domaines de

[26] Commission Nationale Informatique et Libertés
[27] https://www.cnil.fr
[28] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000033202746/
[29] https://www.cnil.fr/fr/la-loi-informatique-et-libertes

recherche. Cependant, il est important d'encadrer la circulation des données personnelles pour le respect de la vie privée de tous.

Dans le contexte de notre projet de recherche, il faut prendre en compte non seulement les données de la recherche, mais également les données des utilisatrices et utilisateurs qui consultent notre dispositif de consultation. Bien que notre nouveau site ne possède pas de cookies pour fonctionner, notre interface doit tout de même comporter des mentions légales ainsi qu'une page pour la politique de confidentialité conformément à la Réglementation Générale sur la Protection des Données (RGPD) de 2018[30]. Nos données de recherche ne contiennent pas de données personnelles, mais nous devons garder à l'esprit les normes et pratiques pour être en conformité avec le RGPD et la CNIL.

- Le respect des personnes

Alors que nous avons décrit plutôt l'importance de la protection des données, ce qui nous intéresse maintenant c'est la légitimité des données c'est-à-dire qu'on va justifier la collecte de nos données pour montrer qu'elles sont essentielles à nos recherches. Ainsi lorsqu'on va utiliser des méthodes invasives (surtout dans les données biomédicales) pour collecter les données alors on va avoir en France un comité national d'éthique pour valider la légitimité des actions envisagées en fonction des hypothèses émises (Schöpfel, Jacquemin, Chaudiron, Kergosien. 2018). Ces comités vont alors étudier ces méthodes de collectes, mais aussi étudier les données elles-mêmes et notamment les données sensibles comme l'appartenance religieuse qui doit être strictement encadrée. Un peu comme notre précédente partie les comités éthiques vont surtout intervenir dans des domaines proches de la santé, dans les sciences humaines et sociales ce seront alors des domaines proches de la santé comme la sociologie, la psychologie au sein desquelles les données peuvent être sensibles. Enfin, c'est par la communication concernant le devenir des données entre les chercheurs et les personnes concernées que l'ouverture des données va se faire. Cependant, comme le précise Schöpfel, Jacquemin, Chaudiron et Kergosien, l'échange ne doit pas avoir de conflits d'intérêts où une des deux parties tirerait un avantage de la collecte des données au détriment de l'autre partie. Puis l'ouverture des données peut se faire au détriment des personnes impliquées.

Nous pouvons demander une évaluation éthique de notre protocole de recherche en saisissant la fédération des CER (comités éthiques de recherche)[31] ce qui va nous permettre de savoir si nous sommes dans les limites de l'éthique concernant nos données et nos méthodes de collecte

[30] https://www.economie.gouv.fr/entreprises/reglement-general-sur-protection-des-donnees-rgpd
[31] https://www.univ-reims.fr/recherche-innovation-et-valorisation/comite-d-ethique-pour-la-recherche/comite-d-ethique-pour-la-recherche,25954,42763.html

des données. Il est important de savoir que les comités d'éthique de la recherche vont être sollicités pour les disciplines en Sciences Humaines ou d'autres. En revanche, pour le biomédical, la loi Jardé de 2016 donne un cadre pour formaliser un avis sur l'éthique des recherches sur l'être humain. Les comités examinent alors la dimension éthique des recherches qui ne sont pas concernées par la loi Jardé.

- la crédibilité des données

L'éthique des données va passer par la crédibilité des données, c'est-à-dire la confiance qu'on va accorder au service des données (Schöpfel, Jacquemin, Chaudiron, Kergosien. 2018).

Pour cela, on va examiner trois axes, le premier, la qualité des données, on va vérifier si les chercheurs ont fourni des données de qualité. Pour le projet Adressbuch, les données ont été nettoyées, enrichies, transformées : les utilisateurs peuvent savoir comment nous avons enrichi les données, avec quelle base de données nous avons travaillé. Cela permet aux chercheurs de vérifier également la véracité des données, ce qui accentue la crédibilité des données et la confiance envers les utilisateurs.

Puis on va avoir la qualité du dispositif c'est-à-dire on va s'intéresser à la documentation, la conservation, l'interopérabilité des données sur la plateforme de diffusion. Ces enjeux se rapprochent de l'éthique des données concernant la finalisation des données et les promesses de la plateforme concernant la réutilisation des données. Alors, il existe de nombreuses façons de conserver les données dans le temps, pour pérenniser les jeux de données et la consultation des données elles-mêmes on fait le choix de dispositifs Open Source qui sont plus simples à maintenir que des logiciels propriétaires, des logiciels Open Source où les mises à niveaux seront plus simples et qui sont maintenus par des communautés de développeurs assez importante. Ensuite, mettre les jeux de données sur des sites d'hébergement facilement accessibles avec un numéro DOI[32] qui va permettre de le retrouver facilement également. On propose les données sous plusieurs formats ouverts et simples à manipuler pour la navigation et la réutilisation des données tels que le JSON, CSV. On peut également proposer le code du dispositif en Open Source pour favoriser la reproductibilité du site et des jeux de données. Enfin, on met en place des fonctions pour favoriser l'interopérabilité des données avec certains standards du web comme rendre les données interopérables et compatibles avec schema.org qui permet de documenter le jeu de données et récupérer facilement les métadonnées.

[32] Digital Object Identifier

Le dernier axe est la qualité de la recherche va concentrer la crédibilité et l'intégrité du chercheur et de son équipe qui est mise en parallèle avec la qualité des données qui contribuent à rendre les données crédibles et favorisent alors leurs diffusions.

- la sécurité des données

Ici, nous allons nous intéresser à la question de conservation et de consultation des données en toute sécurité. Une mauvaise conservation et sécurisation des données pourrait entraîner des problèmes éthiques. En effet, l'intégrité des données repose sur une bonne conservation de ces dernières sans qu'elles soient altérées pendant la période de conservation et de disponibilité sur le dispositif de consultation. Il peut avoir des problèmes éthiques des données notamment si des accidents techniques apparaissent lors de la consultation des données, car cela ne permettrait pas un accès équitable aux données de la recherche. De plus, la sécurisation des données peut être fragilisée par l'erreur humaine qui est toujours présente comme le soulignent Schöpfel, Jacquemin, Chaudiron et Kergosien, mais aussi la malveillance avec des personnes mal intentionnées qui pourraient interagir avec les données et affecter leurs intégrités.

- la propriété intellectuelle

On pourrait également s'intéresser à la propriété intellectuelle des données de la recherche. Le problème actuellement lorsqu'on partage les données de la recherche c'est qu'ils sont parmi de nombreux contenus gratuits et accessibles ce qui fait que certains utilisateurs vont utiliser ces données sans même citer les créateurs des données ou/et s'en servir à mauvais escient. Les chercheurs ont alors peur de partager les données et de les voir utiliser pour des actions dont ils n'auraient pas connaissance sans bénéfice et sans morale avec aucune citation (Schöpfel, Jacquemin, Chaudiron, Kergosien. 2018).

II) Méthodes et pratiques utilisées dans le cadre du projet

II.1) Présentation technique du projet

Conception de la base de données

Pour réaliser le nouveau dispositif de consultation et pour permettre l'ouverture des données pour les utilisateurs, il était important de repenser la structure des données afin que ces derniers soient facilement consultables. Dans un premier temps, les données ont été extraites de l'ancienne interface réalisée avec Filemaker 7. À partir de ces données, nous devions réfléchir

à une nouvelle structure pour stocker, puis consulter et partager les données de manière efficace. L'autre objectif de cette migration est d'ouvrir les données initialement protégées par le droit d'auteur en les partageant sous une licence libre afin de permettre l'ouverture de ces données conformément aux bonnes pratiques en matière de données de la recherche.

La modélisation de la base de données a été réalisée par Gérald Kembellec et Alina Ostrowski. Cette structure a été conçue avec le logiciel en ligne Mocodo qui permet de construire une base de données relationnelles en MySQL. Notre but était de construire une base de données prosopographique[33] qui réponde aux besoins des chercheuses et chercheurs, mais aussi qui soit réalisable par les ingénieurs d'études. La structure de la base de données a été pensée en diachronie avec deux temporalités : Paris aujourd'hui et Paris au moment de la publication du bottin du commerce en 1854. C'était important d'avoir cette temporalité, car beaucoup de changements ont eu lieu à Paris depuis 1854, on peut évoquer les travaux Hausmann dans les années 1860 avec le percement des grands boulevards et par conséquent la disparition de certaines rues de Paris. Puis nous avons également, la transformation des arrondissements en 1859 passant de 12 à 20 arrondissements. La base de données contient 21 tables. Les tables courantes contiennent les données du projet tandis que les tables associatives assurent la jonction entre les différentes tables courantes. Enfin certaines jonctions sont simplement réalisées entre les tables courantes par une clé étrangère. Les tables courantes sont : Persons, Companies, Streets, Addresses, Arrondissement, External_References, Original_References, Ldh_rank, Military_statuses, Social_statuses, Prof_categories, Reference_types, Occupation_statuses.

Nous allons à présent décrire précisément ce que contiennent les tables courantes de la base de données.

Table Persons :

Nom	**Type**	**Description**
id	integer	identifiant de chaque personne dans la base de données
surname	varchar	Nom
first_name	varchar	Prénom (souvent une initiale)

[33] Dans ce contexte, réaliser une base de données contenant les informations de chaque personne provenant d'Adressbuch, pour étudier la communauté allemande à Paris au XIXe siècle.

gender	enum (‘M’,’F’)	Genre choix entre ‘M’ Homme ou ‘F’ Femme
title	varchar	Titre qui accompagne le nom et prénom comme un titre de noblesse (‘*Graf*’ par ex. pour comte)
name_predicate	varchar	prédicat du nom
specification_verbatim	varchar	spécification sur la personne par rapport à sa position, sa profession ou autre
profession_verbatim	varchar	profession telle qu’elle est écrite dans le bottin du commerce
profession_unified	varchar	modernisation de l’écriture de la profession si nécessaire
de_l_institut	tinyint	Si la personne fait partie d’un institut
notable_commercant	tinyint	Si la personne est indiquée comme notable
bold	tinyint	Si la personne a participé à la publication du bottin, elle est indiquée en gras dans le bottin
advert	tinyint	Si la personne dispose d’une publicité (pour son entreprise par exemple)
ldh_rank_id	integer	clé étrangère assure la relation avec la table ldh_rank
military_status_id	integer	clé étrangère assure la relation avec la table military_status
social_status_id	integer	clé étrangère assure la relation avec la table social_status
occupation_status_id	integer	clé étrangère assure la relation avec la table occupation_status

prof_category_id	integer	clé étrangère assure la relation avec la table prof_category

La table Persons est la table principale du projet, celle autour de laquelle s'organise le projet *Adressbuch,* car c'est elle qui contient toutes les données contenues dans le bottin du commerce. Avec ses 4772 lignes, elle recense toutes les personnes figurant dans le bottin.

Table companies:

Nom	**Type**	**Description**
id	integer	identifiant de l'entreprise
name	varchar	nom de l'entreprise
specification_verbatim	varchar	spécification supplémentaire sur l'entreprise
profession_verbatim	varchar	profession telle qu'elle est écrite dans le bottin
profession_unified	varchar	modernisation de l'écriture de la profession si nécessaire
prof_category_id	integer	clé étrangère assure la jonction avec prof_categories
notable_commercant	tinyint	si l'entreprise est indiquée comme notable
bold	tinyint	si l'entreprise a participé à la publication du bottin, elle figure en gras
advert	tinyint	si l'entreprise dispose d'une publicité

La table Companies contient les 141 entreprises qui figurent dans le bottin du commerce, elle constitue une table principale puisqu'elle fait partie avec la Table Persons des index consultables dans le nouveau dispositif de consultation.

Table Streets :

Nom	Type	Description
id	integer	identifiant pour chaque rue
name_old_verbatim	varchar	nom de la rue tel qu'il est écrit dans le bottin du commerce
name_old_clean	varchar	nom de la rue normalisé
name_new	varchar	nom de rue actuel s'il a changé depuis 1854
geo_long	float	longitude
geo_lat	float	latitude

La table Streets contient les 971 rues qui figurent dans le bottin du commerce. Cette table est importante, car elle permet de placer sur une carte historique et moderne de Paris les personnes figurant dans le bottin avec les colonnes geo_long et geo_lat. Elle permet également une étude diachronique des rues de Paris puisque nous avons les noms de rues anciens et actuels de Paris.

Table Arrondissements :

Nom	Type	Description
id	integer	identifiant de chaque arrondissement
no	smallint	numéro d'arrondissement
insee_citycode	mediumint	code INSEE de chaque arrondissement
type	varchar	le type de chaque arrondissement si c'est un arrondissement avant ou après 1860 ou bien des communes annexées
postcode	mediumint	code postal de chaque arrondissement

La table Arrondissement permet de recenser chaque arrondissement avant et après 1860, car ils ont subi des mutations. Elle permet de réaliser une étude diachronique de Paris pour voir dans

quels arrondissements étaient les allemands avant et après 1860. Cette table a pu être réalisée grâce aux archives de la ville de Paris[34] qui nous permet d'établir la correspondance entre les anciens et nouveaux arrondissements de Paris.

Nouvelle interface avec cakephp

Le nouveau dispositif de consultation permet l'ouverture des données. Pour le réaliser, nous avons utilisé le framework cakePHP. Un framework est un ensemble de composants qui permettent d'avoir la structure de base d'une application. Ici, cakePHP nous permet d'avoir une base d'application fonctionnelle qui peut interagir avec une base de données. cakePHP est un framework en PHP qui développé en open-source. Il est maintenu par une grande communauté de développeurs, ce qui permet aux applications d'être pérennes. cakePHP est disponible sur Github et a également un dépôt github pour télécharger le code source[35]. L'installation requiert d'avoir PHP installé sur sa machine dans la version 7.2 minimum pour la dernière version de cakePHP (4x). Il faut également certaines extensions de PHP comme intl, mbstring ou encore pdo_sql qui permet la connexion à une base de données et le bon fonctionnement de cakePHP.

Ce sont sur ces fonctions que se base le framework cakePHP, si elles ne sont pas installées sur le serveur de développement et de production, le framework ne pourra pas fonctionner. Il suffit de les installer sur les serveurs (si cela n'est pas déjà fait) et de décommenter les lignes correspondantes dans le fichier de configuration PHP : php.ini à la rubrique extension. L'extension PHP pdo_sql est une extension sur laquelle se base cakePHP pour créer ses configurations entre son application et la base de données, si l'extension n'est pas installée cakePHP ne peut pas avoir accès à la base de données.

Enfin, l'installation nécessite d'avoir Composer d'installé. Composer est un gestionnaire de dépendances PHP. Une fois que celui-ci est installé, nous pouvons créer notre application en ligne de commande dans le répertoire de notre serveur de développement. Cakephp ne nécessite pas de serveur de développement web, en effet, il suffit de lancer son propre serveur en ligne de commande. Ce serveur est uniquement utilisable pour les phases de tests et développement, mais en aucun cas ce n'est un serveur web de production.

[34] https://archives.paris.fr/_depot_ad75/_depot_arko/articles/47/correspondance-entre-les-arrondissements-anciens-et-nouveaux_doc.pdf
[35] https://github.com/cakephp/cakephp

Un des avantages de cakephp c'est qu'on peut l'utiliser uniquement ou presque en ligne de commande que ce soit pour créer l'application ou pour créer les fichiers qui vont composer notre structure, c'est-à-dire les Models, Controllers ou les Templates (index, view).
cakePHP suit une architecture MVC (Model, View, Controller).

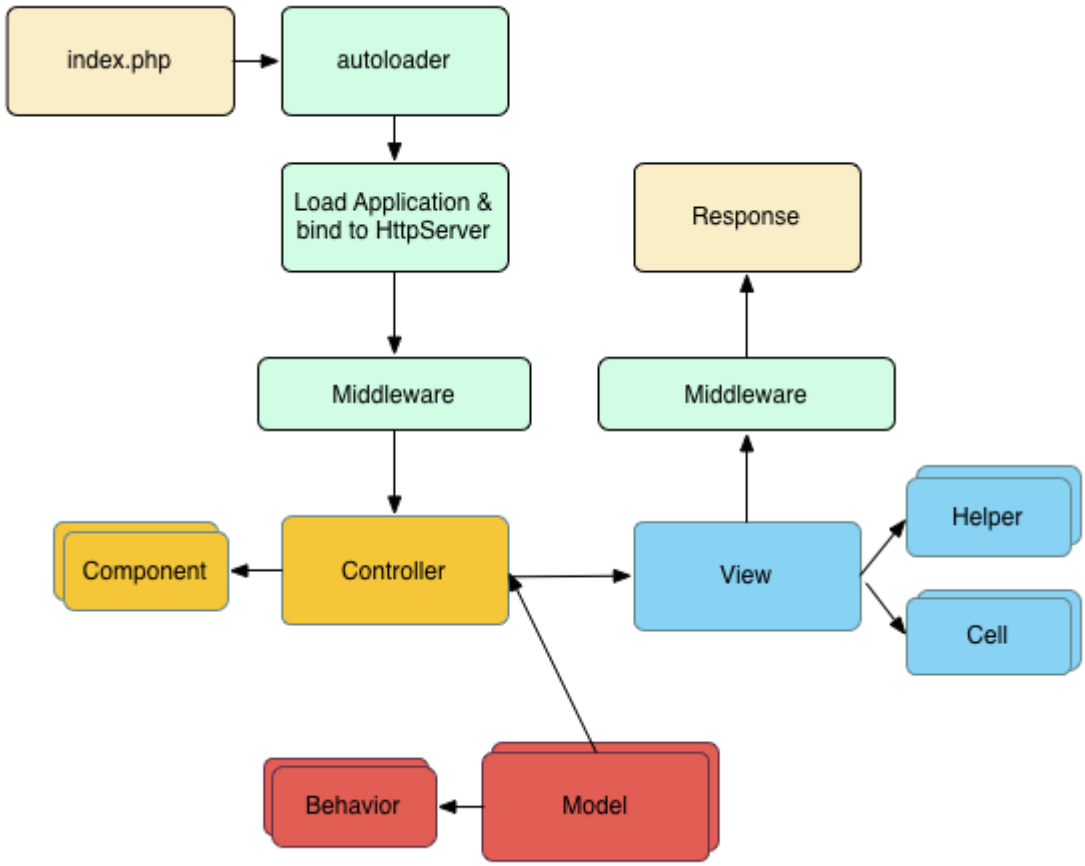


Fig 1. Schéma illustrant l'architecture MVC de cakePHP , https://book.cakephp.org/4/fr/intro.html

Le 'Model' contient les données de la base de données que nous pouvons créer avec la commande : `cake bake model`. Dans notre application, les fichiers de Models sont égaux au nombre de tables dans la base de données c'est-à-dire 21. Le controller permet de faire la relation entre la View et le Model, c'est dans le controller qu'on va créer les fonctions d'index, de vue unique ou encore d'export, le controller permet de gérer les actions des utilisateurs. Les controllers disponibles dans l'application sont Persons, Companies, Streets, Arrondissement qui sont les quatre onglets de notre dispositif à pouvoir être consultés et Search qui permet la recherche avancée et la barre de recherche à l'accueil pour une recherche simple. Il peut être créé avec la commande : `cake bake controller`. Enfin, la View qui permet de réaliser l'interface graphique et d'afficher les données. Elle peut être créée avec la commande : `cake`

`bake template`. Les templates[36] sont identiques aux controllers c'est-à-dire Persons, Companies, Streets, Arrondissement et Search. Chaque template comporte un fichier index qui nous permet de voir un tableau rassemblant les données de la base et un fichier view qui permet d'avoir une vue spécifique pour chaque personne qui figure dans le bottin du commerce. Enfin, j'ai créé un template fonction qui me permet d'écrire toutes les fonctions en PHP nécessaires au bon fonctionnement du site.

Ces fonctions permettent notamment de générer des balises span[37] en HTML pour enregistrer chaque personne, chaque image du site dans une bibliothèque Zotero. La norme NISO Z39.88 ou Context Object In Span (COinS) permet d'exposer des métadonnées bibliographiques ou catalographiques et les rend détectables par des outils de navigation "augmentée" liés à un outil de gestion de référence bibliographique (Kembellec, 2013).

Cela permet une meilleure interopérabilité avec les autres plates-formes.

Enfin, cakePHP permet une modernisation de l'interface par rapport au web companion FileMaker. Une nouveauté est la recherche avancée qui permet de rechercher efficacement dans l'interface avec de nombreux champs de recherche. Ainsi, nous pouvons chercher une personne selon son nom, prenom, sexe, adresse, rang dans la Légion d'honneur ou encore la catégorie professionnelle et la catégorie sociale. Cependant, la nouveauté la plus importante concerne l'ouverture des données avec les différents exports proposés. Les utilisateurs ont le choix entre quatre formats pour télécharger la base de données entière CSV, JSON, SQL, XML. Le CSV est le format le plus polyvalent, car il ne nécessite pas de logiciels propriétaires pour être manipulé. De plus, c'est un format avec une structure simple où les données sont simplement séparées par des virgules, il peut alors être facilement manipulé. Le JSON et le XML sont des formats de données structurées qui nécessitent le plus souvent un éditeur de code pour être lus, ils sont facilement interrogeables en utilisant les champs définis dans le fichier. Enfin le SQL est le format pour recréer la base de données à l'identique par les utilisateurs. Toutes les requêtes pour créer la base et insérer les données sont dans le fichier, il suffit de l'importer dans son système de gestion de base de données MySQL ou dans l'interface graphique qui gère les bases de données (phpMyAdmin) pour ensuite utiliser le fichier.

[36] Les templates sont les fichiers contenant la logique pour récupérer les données depuis le Controller et l'afficher pour les utilisateurs.

[37] Une balise span est un conteneur de support pour CSS ou JS, voir : https://www.w3schools.com/tags/tag_span.asp

Fig 2. La nouvelle interface réalisée avec cakePHP

La nouvelle interface se comporte de plusieurs onglets. Les onglets Persons, *Unternehmen* (Compagnies), *Strassen* (Rues) et Arrondissements dépendent directement de la base de données et chacun de leurs Template et Controller spécifique dans l'application. C'est une nouveauté par rapport à l'ancien dispositif de pouvoir consulter les données du bottin du commerce selon une thématique précise. Les onglets *Adressbuch* (Carnet d'adresses), *Datenbank* (base de données), *Karte* (Carte), *Digitalisate* (Numérisations), Partner und Team, *Hilfe* (aide) et *Download* (téléchargements) sont gérés par le PageController qui est un contrôle qui assure la gestion des pages qui ne dépendent pas directement de la base de données. Les onglets précédemment évoqués sont enregistrés dans le Template Pages. Ces onglets ont pour objectif de documenter le projet Adressbuch pour que chaque utilisatrice et utilisateur connaisse l'histoire du projet ainsi que le fonctionnement de la base de données ou encore de la cartographie interactive. Les utilisateurs disposent également d'une aide pour naviguer dans la base de données ainsi que la possibilité d'enregistrer leurs recherches et de les retrouver dans l'onglet de téléchargement pour pouvoir exporter leurs sauvegardes à tout moment. Enfin, l'onglet *Digitalisate* permet de découvrir le document primaire, c'est-à-dire qu'il est possible de consulter les numérisations du carnet d'adresses sous la forme d'un livre numérique.

Numérisation du document

Pour pouvoir consulter le document primaire et réaliser les OCRisations, nous devions d'abord numériser le bottin du commerce. Le carnet d'adresses qui est disponible sur le site est conservé à la bibliothèque historique de la ville de Paris. La bibliothèque historique de la ville de Paris a accepté de nous prêter le bottin du commerce une journée pour pouvoir le numériser.

Le bottin du commerce est un annuaire d'adresses recensant les Allemands vivant à Paris en 1854. Il a été publié à l'initiative de F. A Kronauge qui vivait rue Richelieu et travaillait pour un institut de langues. Numériser l'ouvrage permet aux utilisateurs du site de consulter un *fac-similé* de qualité du document d'origine et d'attester de la véracité de nos données. Cela apporte de la transparence concernant la source de nos données. Le livre est consultable depuis l'onglet *digitalisate* du site web, nous avons d'abord une prévisualisation dans une définition standard qui permet de ne pas ralentir le chargement de la page puis lorsque nous sélectionnons une page en particulier nous pouvons apprécier chaque détail de la page dans une haute définition en 300dpi, résolution recommandée selon les standards du web.

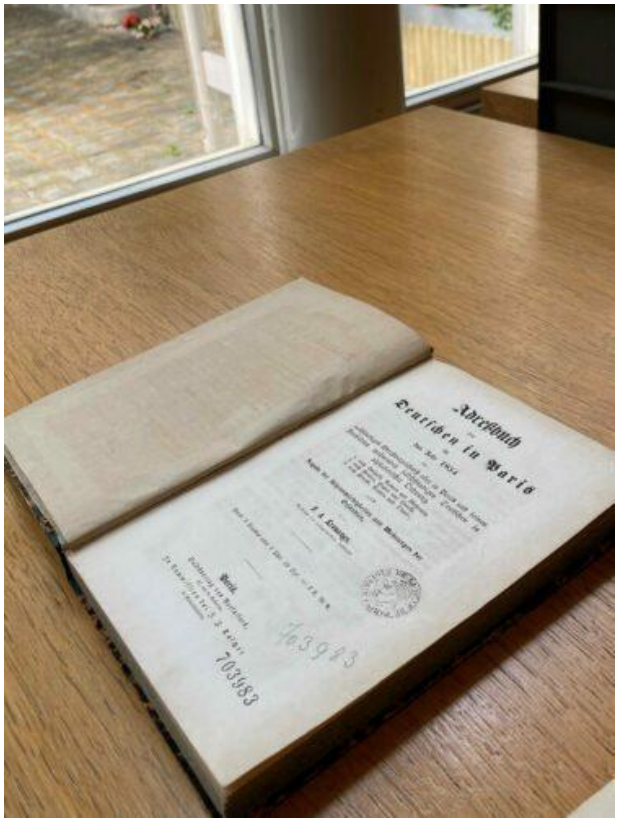
Pour effectuer la numérisation, nous avons utilisé le scanner de la bibliothèque de l'institut historique allemand de Paris qui est conçu pour la numérisation d'anciens documents. Un de ces avantages est de pouvoir compenser de part et d'autre du livre le volume de pages ce qui permet de ne pas abîmer la reliure du livre. Ce sont 247 pages du livre que j'ai scannées puis transférées sur une clé USB afin de pouvoir les inclure dans le site. Avant cette étape, j'ai effectué l'extraction des métadonnées et le redimensionnement de la page. Concernant l'extraction des métadonnées, j'ai utilisé dans une boucle 'do / done' la commande shell 'file' en redirigeant la sortie standard vers un fichier csv au moyen de '>' :

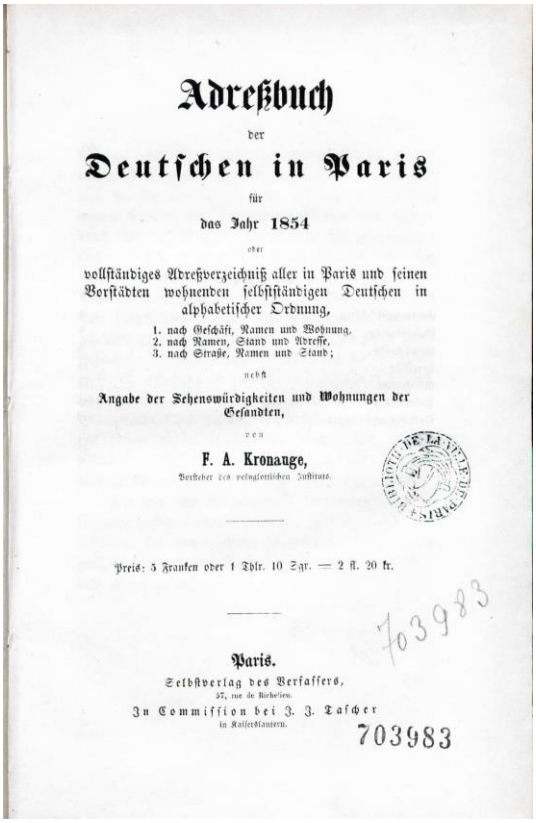
`for i in *jpg do file $i > $i.csv done;` dans un interpréteur de commande Linux. J'ai ensuite enregistré les informations de chaque numérisation dans un fichier CSV disponible sur le dépôt Zenodo. Ces informations, comme la résolution, la taille de l'image, le nom ou encore le format sont également disponibles sur le site web. Puis, pour

Fig 3. Du document primaire à la numérisation

le redimensionnement de chaque image j'ai utilisé le package imagemagick[38] toujours dans le shell Linux avec la commande :

[38] https://imagemagick.org/





Adreßbuch
der
Deutschen in Paris
für
das Jahr 1854
oder
vollständiges Adreßverzeichniß aller in Paris und seinen Vorstädten wohnenden selbstständigen Deutschen in alphabetischer Ordnung,
1. nach Geschäft, Namen und Wohnung,
2. nach Namen, Stand und Adresse,
3. nach Straße, Namen und Stand;
nebst
Angabe der Sehenswürdigkeiten und Wohnungen der Gesandten,
von
F. A. Kronauge,
Vorsteher des pädagogischen Instituts.
Preis: 5 Franken oder 1 Thlr. 10 Sgr. = 2 fl. 20 kr.
Paris.
Selbstverlag des Verfassers,
37, rue de Richelieu.
In Commission bei J. J. Tascher
in Kaiserslautern.
703983

`mogrify -units pixelsperinch -density 72x72 resize 400x800 -path chemin_pour_les_nouvelles_images` qui permet le traitement d'images en lot. Ainsi, j'ai pu redimensionner chaque image dans une qualité standard c'est-à-dire en 72dpi avec une taille moyenne de 400x800 pour permettre une prévisualisation des numérisations sans demander trop d'efforts au nouveau dispositif de consultation puisque la résolution est réduite.

OCRisation du document

L'OCRisation du document primaire était un appui intéressant, car il permet d'avoir uniquement le contenu textuel du carnet d'adresses, il est alors plus facile de lire une page sans sa typographie d'origine qui peut-être difficile à lire ou à cause de la conservation du document en lui-même, car les couleurs peuvent changer en 170 ans.

Réaliser un OCR sur un document historique nécessite un logiciel. Dans le cas du carnet d'adresses, nous avons un document imprimé du XIXe siècle. Les numérisations que nous avons faites dans une résolution en 300 dpi nous permettent d'avoir des OCR de bonne qualité. On considère qu'un OCR a fait une bonne performance lorsque son taux d'erreurs est inférieur à 10% (Abaynarch, El Fadili, Zenkouar. 2015). On peut ensuite envisager de l'exploiter en corrigeant les erreurs restantes manuellement.

Cependant, notre document primaire a plusieurs paramètres à prendre en compte. Tout d'abord la plupart des pages sont sur deux colonnes il faut donc prendre en compte la segmentation des pages pour que le taux de réussite soit plus élevé. Autre élément à prendre en compte pour la réussite des OCR : la typographie. En effet, dans le carnet d'adresses nous avons du Fraktur[39] qui est un type d'écriture gothique cela peut engendrer des erreurs si on ne le prend pas en compte. Enfin, dans le bottin du commerce nous avons deux langues le français et l'allemand, il faut également indiquer les deux langues lors du processus d'OCR pour qu'il ne renvoie pas d'erreurs.

Pour réaliser cette mission, nous avons utilisé Tesseract-ocr qui est un moteur d'OCR open-source et gratuit. C'est également le moteur OCR open-source considéré comme le plus puissant en libre accès. Cependant, Tesseract-OCR a des limites: on ne peut faire des OCR que des documents imprimés. J'ai réalisé une première tentative d'OCR qui a donné des résultats mitigés. En effet, il y a eu beaucoup d'erreurs du fait que je n'avais pas pris en compte la typographie du carnet d'adresse, la segmentation et les langues. J'avais exécuté un script shell de traitement par lot avec tesseract pour procéder à un OCR en lot des numérisations.

```
for i in *jpg do tesseract $i $i.txt
```

Cette commande permet de traiter toutes les numérisations au format JPEG pour obtenir une OCRisation au format texte en sortie.

J'ai par la suite effectué un second processus d'OCR qui a donné de meilleurs résultats pour plusieurs raisons. La première est que j'ai réalisé un pre-processing des images, c'est-à-dire que j'ai cherché à réduire les bruits qui peuvent être présents dans documents anciens comme les tâches sur les pages. J'ai ensuite augmenté le contraste pour avoir une typographie plus noire et un fond plus blanc ce qui améliore les résultats de l'OCRisation. Enfin j'ai ajouté des paramètres pour prendre en compte les difficultés que j'ai rencontrées lors du précédent OCR. On obtient le script shell qui suit :

```
for i in *jpg do tesseract $i $i.txt –psm 3 -oem 1 -l script/Fraktur+fra
```

Avec cette commande j'utilise le paramètre de segmentation -psm 3 qui permet de détecter automatiquement les colonnes sur chaque numérisation, puis j'utilise pour les langues le français et un script le German Fraktur qui a été utilisé dans un projet en Allemagne par

[39] https://fr.wikipedia.org/wiki/Fraktur

l'université de Mannheim pour effectuer des OCRisations sur des journaux allemands et prussiens du XIXe et XXe siècle[40].

Les résultats ont été optimisés, mais ils demeurent imparfaits, c'est pourquoi j'ai procédé à une correction manuelle de chaque page pour corriger les erreurs résiduelles. C'est une étape fastidieuse, mais elle permet d'avoir des fichiers textes quasi-parfaits et cela permet une meilleure visibilité pour notre document primaire ainsi qu'une meilleure accessibilité pour les usagers de la nouvelle plate-forme. Elle nécessite cependant une connaissance de la période historique, du contexte socioprofessionnel et des deux langues traitées. Ce travail de qualification est donc hybride.

Versioning avec Github et Cologne

Pour le nouveau dispositif de consultation, l'Institut Historique Allemand de Paris a formé un partenariat avec l'Institut d'Humanités Numériques à Cologne qui s'occupe de l'hébergement de l'application et de la base de données. Comme le développement de l'application s'est réalisé au moment de la pandémie de COVID-19, il était impossible pour nous de nous déplacer à Cologne et il fallait réfléchir à une solution pour développer notre plate-forme à l'institut puis l'envoyer à l'Institut de Cologne pour qu'elle soit hébergée. La solution qui a été retenue est d'utiliser Github. Github est un service web d'hébergement et de gestion de développement de logiciels. Cette plateforme utilise le logiciel de gestion de version Git développé par Linus Torvalds. Github permet de gérer les versions de notre nouvelle plate-forme et de la mettre à jour dès que nous en avons besoin. Pour réaliser ses mises à jour, l'institut de Cologne a créé un « repository », c'est-à-dire un dossier sur leur compte Github dans lequel se trouve notre nouvelle plateforme. Nous avons ensuite effectué un « fork », une copie de ce dossier sur le compte Github de l'Institut Historique Allemand de Paris, ce qui nous permet de participer à ce projet et de suggérer des modifications.

Lors du développement de l'application, j'ai téléchargé le logiciel Github desktop client qui permet d'avoir l'interface graphique Github sur un ordinateur Windows et de copier ses repositories localement sur l'ordinateur. Une fois que j'avais effectué des changements sur la nouvelle plateforme sur mon serveur local, j'effectuais les mises à jour d'abord sur la copie locale de mon repository. Une fois les changements pris en compte, j'effectuais un « push » (envoi) des mises à jour vers le repository en ligne, c'est-à-dire que je pousse les mises à jour du repository local vers mon repository distant. Enfin, la dernière étape, j'effectuais une «pull

[40] voir : https://digi.bib.uni-mannheim.de/periodika/en/imperial-gazette/ocr/

request» vers le github de l'Institut de Cologne pour qu'ils puissent à leur tour effectuer les mises à jour. On vient suggérer les modifications en détaillant quels changements ont été faits sur l'application.

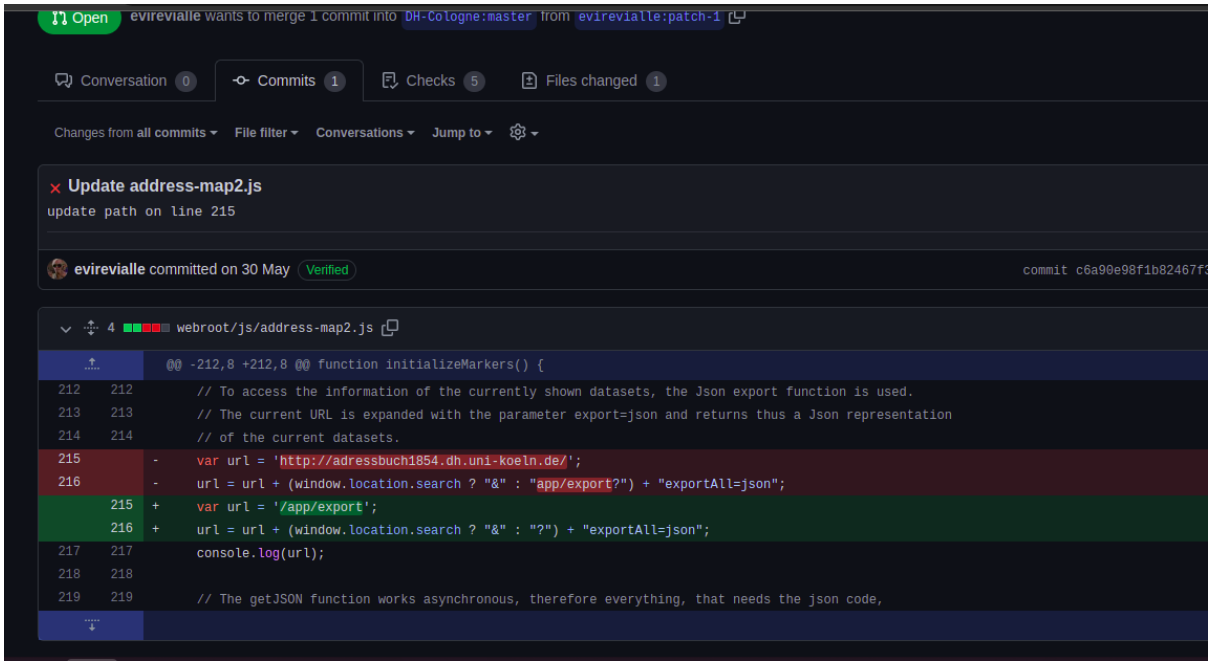


Fig 4. Exemple de Pull Request sur Github

Je ne pouvais pas effectuer les changements moi-même sur le Github de l'Institut de Cologne, car je n'ai pas les droits sur le repository pour accepter les changements, il faut alors attendre que l'administrateur du Github accepte ou refuse la « pull request ».

Bien que cette expérience fût enrichissante, car j'ai pu appréhender un nouvel outil Github, j'ai également appris à m'en servir en ligne de commande, je trouve qu'il reste difficile de travailler sur la gestion des versions d'un site web avec une autre équipe lorsque celle-ci ne se trouve pas dans la même entreprise. Il peut avoir des difficultés pour les disponibilités des deux équipes et les changements ne seront pas immédiatement pris en compte.

Toutefois, Github est une plateforme très connue des développeurs, ce qui apporte de la visibilité à notre projet, de plus, il est possible de reproduire la même plateforme si les usagers du site le souhaite car le projet est Open-Source. Il est alors tout à fait possible de récupérer l'intégralité du code que nous avons créé.

II.2) La qualification des données

OpenRefine : pour le nettoyage et l'enrichissement des données

Avant de pouvoir partager les données sur le nouveau dispositif de consultation, il est important de les nettoyer et éventuellement de les enrichir pour obtenir des datasets prêts à être utilisés. Nous l'avons expliqué plus haut, les données ont besoin d'être structurées et documentées pour en permettre une (ré)utilisation optimale. Nous intervenons en amont sur les données avec une étape de qualification. Pour qualifier les données, j'ai utilisé le logiciel OpenRefine conçu pour le nettoyage et l'enrichissement des données. C'est un logiciel développé à l'origine par Metaweb, puis par Google qui avait racheté Metaweb. OpenRefine est à présent maintenu par une communauté de développeurs et est disponible en open-source. OpenRefine nécessite l'installation de Java pour fonctionner ou alors il faut sélectionner le package *with embedded java* qui permet de faire fonctionner OpenRefine sans installer Java sur son ordinateur. Il suffit ensuite de sélectionner le fichier exécutable pour lancer le logiciel. OpenRefine est compatible avec de nombreux formats de fichier ce qui rend son utilisation plus simple, de plus, il est possible de se connecter directement avec une base de données pour travailler sur certaines données. Ainsi OpenRefine accepte le SQL, XML, JSON, CSV, TSV ou encore Excel comme formats d'entrée.

Tout d'abord, nous avons utilisé OpenRefine pour le nettoyage des données. En effet, une fois avoir récupéré le fichier tabulaire initial de l'ancien dispositif nous pouvons commencer le nettoyage des données. Une première étape a été la normalisation des données. Pour ce faire, nous avons utilisé les options de facettes d'OpenRefine. Les facettes permettent de voir les choix qui existent par colonne dans notre tableau et ensuite nous pouvons les regrouper si nous le souhaitons. Nous avons les facettes textuelles qui permettent de trier les données texte, puis les facettes numériques qui permettent de trier les nombres de notre tableau. Enfin, nous avons les facettes chronologiques qui permettent de trier les dates. Dans le cas d'Adressbuch, ce sont surtout les facettes textuelles qui nous ont servi puisque notre fichier contient essentiellement des chaînes de caractères. Dans OpenRefine, il faut sélectionner **nom de colonne>facette>facette textuelle**, il faut ensuite sélectionner **groupe** pour pouvoir normaliser nos données. Cette action va permettre d'avoir la même écriture pour les données qui ont le même sens. Par exemple, nous rassemblons « Mechaniker » et « MECHANIKER » sous la même graphie afin d'avoir les données les plus uniformisées possibles. Lorsque les données ont plus de différence, nous pouvons utiliser la distance de Levenshtein[41]. Cette distance entre

[41] https://fr.wikipedia.org/wiki/Distance_de_Levenshtein

deux mots est le nombre minimum de modifications d'un seul caractère. La distance de Levenshtein entre « Mechaniker » et « Mekaniker » est de 2, car le nombre minimum de modifications est de 2 pour passer d'un mot à l'autre. Cette fonction pour grouper les mots similaires sur OpenRefine nous permet de normaliser toutes les écritures, cependant pour les professions nous avons gardé également le nom des professions telles qu'elles sont écrites dans l'annuaire afin que l'utilisateur dispose des informations telles qu'elles sont indiquées dans Adressbuch.

Ensuite, nous avons fait des transformations dites courantes dans OpenRefine qui permet le nettoyage de nos données, à l'aide de l'option **transformations courantes** nous avons supprimé les espaces superflus, mis nos données en initiales majuscules dans l'optique de les normaliser. Pour des transformations plus spécifiques, j'ai utilisé le langage **GREL**[42] qui est un langage d'expressions régulières qui permet de nettoyer les données. Le langage GREL nous a permis de supprimer des caractères spécifiques comme les tirets par exemple. Sur une colonne, nous appliquons une formule pour enlever des caractères ou les remplacer. Si nous devons remplacer les «-» par des «_» dans notre colonne, alors nous appliquerons la formule : `value.replace('-','_')`.

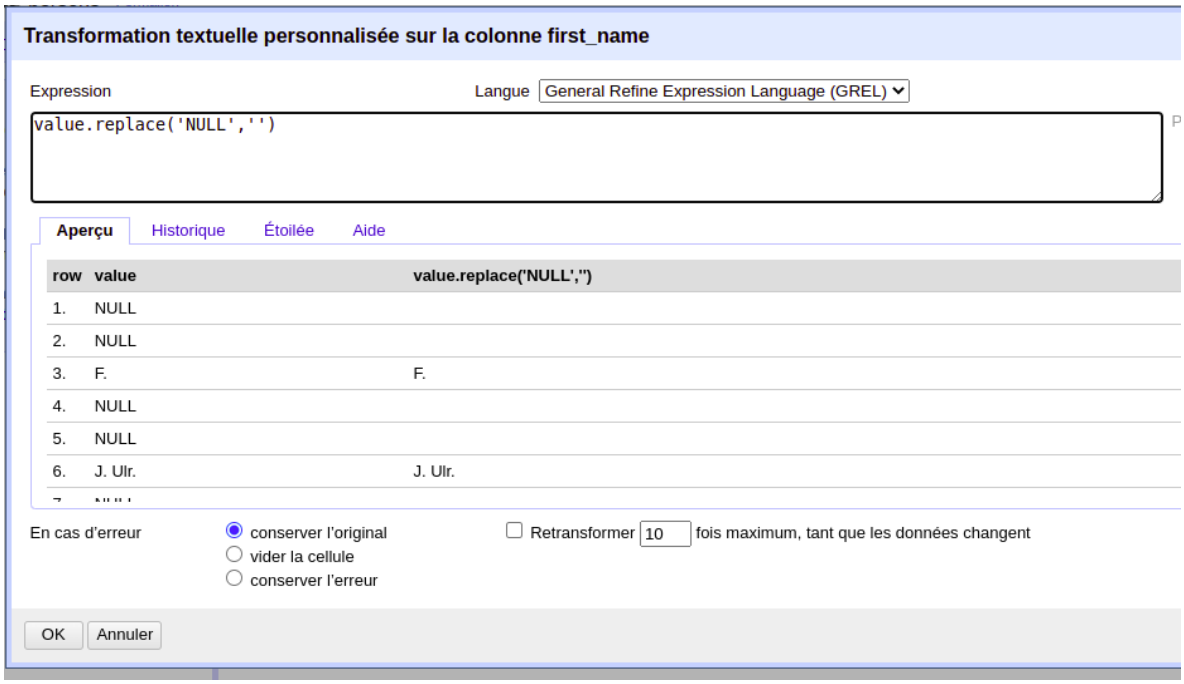


Fig 5. Fenêtre pour utiliser le GREL dans OpenRefine

Après cette étape de normalisation et de nettoyage de données, j'ai utilisé OpenRefine pour fragmenter le fichier tabulaire initial avec la fonction **export tabulaire personnalisé** pour créer

[42] Google Refine Expression Language

les tables qui vont constituer la base de données. Une fois que cela a été fait, j'ai pu créer des exports en SQL. Plusieurs options sont disponibles lorsqu'on fait un export SQL, on peut avoir la possibilité de créer sa table en choisissant les attributs qui correspondent à chaque colonne. Dans notre cas, la base de données était déjà créée, nous avions besoin uniquement de la requête d'insertion SQL. Cela nous fait gagner un temps considérable puisque nous avons juste besoin d'importer nos fichiers dans l'interface graphique phpMyAdmin pour remplir notre base de données.

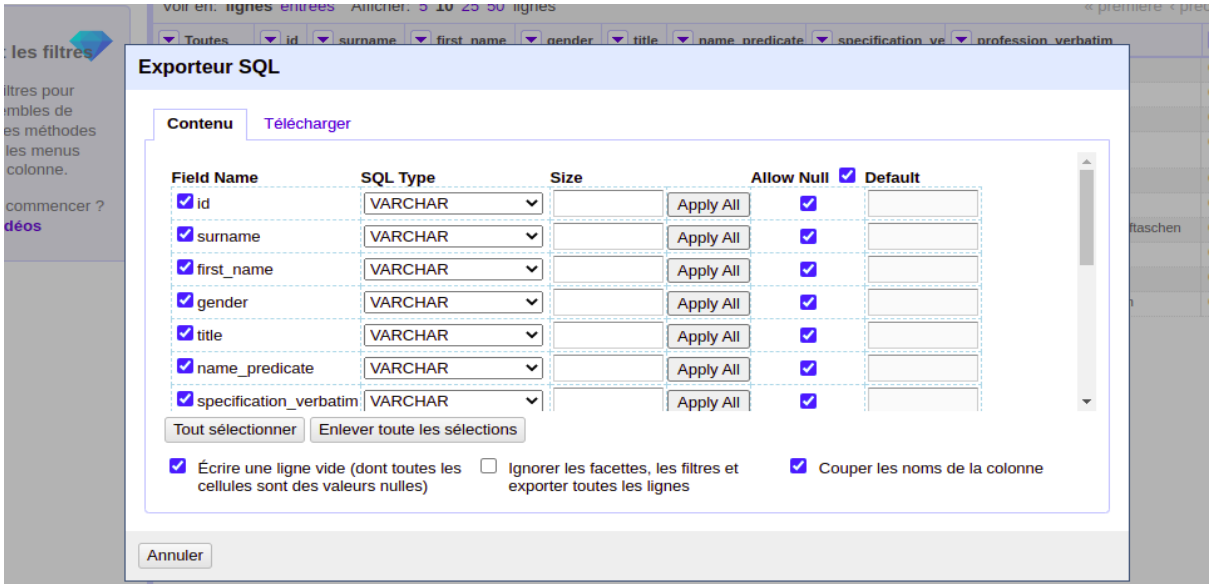


Fig 6. Fenêtre d'export d'OpenRefine pour le SQL

Wikidata : réconciliation des rues pour la cartographie interactive

Une autre spécificité d'OpenRefine est de pouvoir enrichir ses données avec des services de données externes comme Wikidata. Pour effectuer l'enrichissement des données, on va choisir l'option **Démarrer la réconciliation** dans la colonne qu'on souhaite enrichir à partir de là on va choisir le service externe dans lequel on va chercher les connaissances dont nous avons besoin. Pour enregistrer un nouveau service, il faut se rendre sur le site qui répertorie les API[43] de réconciliation compatibles avec OpenRefine[44]. Parmi ces sources, nous avons Wikidata qui regroupe les connaissances de Wikipédia dans plusieurs langues et centralise les données des projets de Wikimedia, VIAF[45] qui contient les connaissances de plusieurs bibliothèques nationales ou encore Geonames qui est une base de données géographique.

[43] API : Application Programming Interface (Interface de Programmation)
[44] https://reconciliation-api.github.io/testbench/#/
[45] Virtual International Authority File

Une fois que le service est choisi, nous allons avoir une détection automatique qui permet de faire correspondre la colonne que nous avons choisie avec les données que nous souhaitons réconcilier. Si je choisis de réconcilier la colonne « Adresse » et que je choisis Wikidata comme base de connaissances, OpenRefine va chercher automatiquement dans l'ontologie de Wikidata la propriété qui correspond le mieux à la colonne « Adresse ». OpenRefine va me suggérer de réconcilier ma colonne avec la propriété « rue » de Wikidata qui correspond à la propriété Q79007 dans l'ontologie de Wikidata. On peut ensuite lancer la réconciliation. À partir de ce moment, OpenRefine va chercher les correspondances pour chaque cellule avec les articles du Wikidata dont la propriété est « rue ». Quand la réconciliation est terminée, certaines cellules sont associées automatiquement avec Wikidata, ce sont généralement les cellules où un seul choix est possible selon OpenRefine. En revanche, les cellules où plusieurs choix sont possibles, OpenRefine nous demande de choisir l'entité qui correspond le mieux aux données de notre fichier. Nous pouvons choisir **apparier cette cellule uniquement** lorsque l'entité correspondra uniquement à cette cellule ou nous pouvons sélectionner **apparier les cellules correspondantes** pour associer l'entité de Wikidata avec plusieurs cellules qui contiennent la même donnée.

Dans notre projet, il fallait prendre en compte une donnée importante, les rues qui sont inscrites dans le bottin du commerce sont les noms de rues de 1854. Il est important de vérifier chaque nom rue pour être certain que ce soit bien le même nom de rue au XIXe siècle. Quand le nom de la rue a changé, il fallait alors chercher le nom actuel de la rue. Malheureusement, certaines rues n'ont pas pu être retrouvées, car elles ont été détruites avec les travaux Haussmann dans les années 1860 lorsque les grands boulevards de Paris ont été percés. D'autres rues n'ont pas de correspondance, car nous avons trop peu d'informations sur ces dernières dans le carnet d'adresses.

Quand cette première étape est terminée, nous avons associé chaque cellule de notre colonne avec une entité correspondante dans le Wikidata, maintenant nous voulons enrichir notre fichier avec de nouvelles colonnes contenant les données de Wikidata. En choisissant **Éditer la colonne>ajouter des colonnes à partir des valeurs réconciliées**, nous pouvons enrichir notre fichier tabulaire initial avec les données de Wikidata. La fenêtre de commande d'OpenRefine nous propose ensuite des propriétés qui sont associées aux articles Wikidata. Notre colonne « Adresse » est associée à l'entité « rue », OpenRefine nous propose les informations contenues dans l'entité « rue », nous pouvons alors avoir le code postal, la ville, les arrondissements ou encore les coordonnées GPS. Ce dernier point est essentiel dans notre projet puisque ce sont ces coordonnées GPS qui nous ont permis d'intégrer les personnes de

l'Adressbuch sur une carte interactive. Une fois que nous avons sélectionné les informations que nous souhaitons, il suffit de l'ajouter dans notre fichier pour avoir de nouvelles colonnes. Nous pouvons voir ici l'intérêt du partage des données. En effet, le fait que Wikidata nous mette à disposition la base de connaissances de Wikipédia nous permet d'enrichir nos données et également de créer de nouvelles fonctions comme la cartographie. Cela nous permet de réaliser une étude diachronique, car les utilisateurs du site vont désormais pouvoir connaître le nom de la rue en 1854, mais également le nom de la rue actuelle s'il a changé. Sans OpenRefine et l'API Wikidata, il aurait été plus fastidieux de rechercher pour chaque rue (il y'en a 971 dans le carnet d'adresses) les coordonnées GPS pour créer notre carte.

Utiliser les données partagées en SHS pour faire parler nos données (ALPAGE, Paris Open Data)

L'intérêt de partager les données des projets en Sciences Humaines Sociales est de permettre que ces données puissent être utilisées dans d'autres projets de recherche. Le but est que les espaces de partage des données de la recherche ne deviennent pas des cimetières de données inutilisables. Lorsqu'on fait un projet de recherche, on gagne du temps à s'appuyer sur les projets existants pour les enrichir. Actuellement, il serait fastidieux et inutile de créer par nous-mêmes toutes les données dont nous avons besoin alors qu'il existe des projets de recherche qui ont déjà produit ces données. En plus des projets de recherche nous pouvons trouver aussi des projets Open Data lancés par les régions, par les villes et par l'État pour proposer des données exploitables.

Dans le cadre du projet Adressbuch, nous nous sommes servis de projets de recherche en Histoire, mais aussi des sites gouvernementaux qui permettent l'utilisation des données.

Tout d'abord, nous utilisé le projet ALPAGE[46] du LaMOP[47] de Paris 1 Panthéon-Sorbonne. Ce projet a eu pour visée de créer une plateforme géohistorique de Paris de l'époque médiévale jusqu'au XIXe siècle. Cette plateforme a été créée par des historiens, des géomaticiens et des informaticiens afin de concevoir un système d'information géographique (SIG). Les données disponibles sont une cartographie interactive disponible pour tous et les données sont disponibles en téléchargement libre. La plupart des données peuvent être téléchargées au format shapefile[48] qui sont très facilement exploitables pour créer des cartes interactives. Ainsi

[46] https://alpage.huma-num.fr/
[47] Laboratoire de Médiévistique Occidentale de Paris
[48] format de fichier pour les systèmes d'informations géographiques.

on peut obtenir la voirie de Paris en 1380, les lieux de savoirs à l'époque médiévale ou encore les arrondissements de Paris avant 1860. Les données sont libres d'accès sous réserve de mentionner le projet comme une source dans notre projet de recherche et dans les articles qui en découlent.

Grâce à l'interopérabilité des données du projet ALPAGE, nous avons pu mener à bien notre cartographie interactive. Si nous avions eu à nous occuper de la cartographie avec le géoréférencement des arrondissements et des quartiers de Paris, la carte interactive ne serait pas disponible actuellement, car ce travail aurait nécessité un temps supplémentaire considérable.

Puis, pour les arrondissements actuels, nous avons utilisé la plate-forme Paris OpenData[49] qui les propose en libre accès. Paris Open Data est une plate-forme qui a vu le jour en 2011 sous l'impulsion du maire de Paris Bertrand Delanoë et de son adjoint chargé de l'innovation Jean Louis Missika. Cette démarche s'inscrit dans le constat que les métropoles, les collectivités et les entreprises amassent de gros volumes de données. Ces données doivent être gérées et désormais les données doivent être partagées, exploitables et accessibles par tous. C'est dans cette optique que la maire de Paris, Anne Hidalgo, a réaffirmé sa volonté de partager les données en mettant en place une clause Open Data dans les marchés publics en 2014. Les données sont disponibles sous la forme de carte, de tableau et également d'API[50]. Ces APIs permettent d'inclure directement dans les projets de recherche les données provenant de Paris Open Data. Pour intégrer plus facilement les données de Paris Open Data et pour les mettre à la disposition de tous les usagers, nous avons récupéré les données de Paris Open Data, à savoir les arrondissements du Paris moderne, au format shapefile de la même manière que les données du projet ALPAGE.

[49] https://opendata.paris.fr/pages/home/

[50] cf. note de bas de page 23

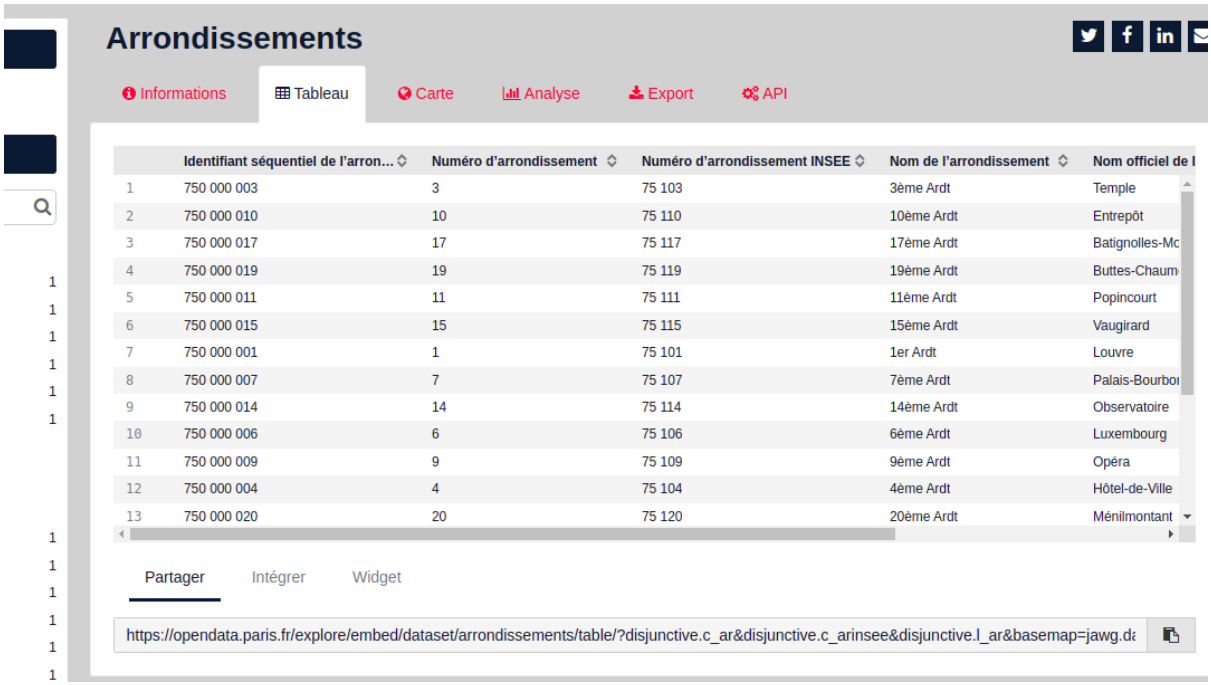


Fig 7. Interface du site Paris Open Data avec les différentes visualisations

Enfin, pour la carte historique nous avons utilisé le géoportail[51] du gouvernement. Le géoportail s'appuie sur les référentiels de l'IGN, il s'est construit sur une logique de partage et d'interopérabilité des données, le géoportail s'enrichit de données publiques. Il propose ses données géographiques au public depuis 2006. Ce qui nous intéressait sur cette plate-forme, ce sont les cartes historiques qui sont disponibles en accès libre. Bien qu'obtenir les cartes en passant par le service WMTS[52] soit une tâche assez fastidieuse, il reste assez simple d'intégrer une carte historique à notre plate-forme notamment avec la documentation proposée par le géoportail. Le service WMTS permet d'ajouter une carte soit sur un logiciel SIG tel que QGIS ou une bibliothèque de cartographie telle que Leaflet.

II.3) Valorisation des données

La valorisation des données est un aspect non négligeable d'un projet de recherche. Il est important de réfléchir à la valorisation de nos données afin que ceux-ci disposent d'une meilleure visibilité et d'une meilleure diffusion. Je présente dans cette partie, les différentes possibilités pour valoriser les données d'un projet de recherche. La valorisation des données

[51] https://www.geoportail.gouv.fr/
[52] WMTS : Web Map Tiles Service, service qui permet d'obtenir des cartes géoréférencées et tuilées depuis un serveur de données externe

s'inscrit également dans la diffusion des données, car si les données sont correctement mises en avant, elles seront alors diffusées de manière efficace que ce soit socialement, mais également par les moteurs de recherche traditionnels par la multiplication des sources secondaires qui les mobilisent.

La cartographie : nouvel outil pour explorer Adressbuch

Après avoir nettoyé les données, puis les avoir insérées dans notre nouvelle base de données, nous pouvons les rechercher et visualiser dans notre nouveau dispositif de consultation. À ce stade, le nouveau site avait les mêmes fonctions que l'ancien dispositif réalisé avec FileMaker à savoir la consultation des personnes qui figurent dans la base de données sous forme d'index. Une des nouvelles fonctionnalités que nous avons souhaité rajouter et qui a été une des missions de mon stage est la cartographie interactive. Cette cartographie doit permettre la consultation de l'ensemble des personnes du carnet d'adresses sur une carte de Paris. De plus, cette carte doit permettre une étude diachronique de Paris, c'est-à-dire étudier Paris selon deux temporalités, au moment où le bottin a été publié en 1854 et Paris actuellement. La personne chargée du développement de la carte avant moi, Alina Ostrowski, avait commencé à développer une carte avec Leaflet[53]. Cependant, il restait l'ajout des éléments historiques de la carte à faire ainsi que l'affichage des personnes.

Leaflet est une bibliothèque JavaScript développée en 2011 par l'ukrainien Volodymyr Agafonkin. Cette bibliothèque permet le développement de cartes interactives et de les ajouter facilement à des pages web. Il suffit d'importer les fichiers JavaScript et css initiaux pour créer sa carte interactive. Leaflet est maintenu par une grande communauté de développeurs qui permettent la création de nombreux plug-ins et est disponible en open source.

Notre carte interagit avec nos données avec un fichier JSON[54] qui contient l'ensemble des personnes de l'Adressbuch. À partir de ces fichiers, notre script va extraire la latitude et la longitude de chaque personne que nous avons réconciliées grâce à OpenRefine et à la base de connaissance Wikidata pour les afficher sur la carte interactive. Des marqueurs personnalisés ont été créés pour distinguer les personnes des entreprises qui figurent dans le carnet d'adresses, des *pop-ups* ont également été créés pour afficher le nom, prénom des personnes ainsi que leurs adresses et un lien qui renvoie à la page web de chaque personne avec ses informations détaillées. À ce stade, rien ne nous permettait d'effectuer une étude diachronique de Paris, j'ai

[53] https://leafletjs.com/

[54] cf note de bas de page 21

alors réfléchi à quel outil était le plus approprié pour réaliser cette carte historique de Paris. J'ai d'abord pensé au logiciel QGIS[55], qui est un logiciel Open Source SIG[56] permettant de réaliser de nombreuses cartes à partir de plusieurs sources. Depuis QGIS, j'ai alors importé une carte de l'état-major provenant du Géoportail qui représente la France en 1820-1866. J'ai pu importer cette carte grâce au Web Map Tiles Service (WMTS) que propose le Géoportail et qui permet d'utiliser ses cartes en établissant une connexion entre leurs cartes et le logiciel QGIS. Concrètement, j'établis la connexion entre le serveur de données et QGIS pour récupérer des cartes géoréférencées tuilées.

Pour établir cette connexion, j'ai utilisé la documentation complète du géoportail[57] sur le WMTS et j'ai ensuite écrit l'URL correspondant à mes besoins à savoir : http://wxs.ign.fr/cartes/geoportail/wmts?SERVICE=WMTS&REQUEST=GetCapabilities

Cette URL me permet de connaître les ressources disponibles concernant les cartes et de les ajouter si j'en ai besoin. Une fois que j'ai construit ma carte sur QGIS je peux ensuite l'exporter pour créer une carte web interactive avec le plug-in QGIS2Web, ce plug-in me permet de créer une carte interactive soit avec OpenStreetMap soit avec Leaflet et de l'exporter pour obtenir une carte en ligne. Cependant je n'ai pas retenu l'option de QGIS pour créer ma carte historique, car il est complexe d'intégrer le dossier QGIS qui contient la carte interactive dans un autre projet web. Je me suis alors demandé comment je pouvais ajouter ma carte historique sur ma carte déjà existante dans le projet en utilisant le WMTS. Il s'est avéré que c'était simple d'ajouter une carte historique, il suffit d'ajouter un nouveau layer au fichier JavaScript en rentrant les paramètres propres à la carte historique.

Dans le fichier cela se traduit comme ceci :

[55] https://www.qgis.org/fr/site/

[56] SIG : système d'informations géographiques

[57] https://geoservices.ign.fr/documentation/services/api-et-services-ogc/images-tuilees-wmts-ogc

```
var cartohisto = L.tileLayer(
      "https://wxs.ign.fr/cartes/geoportail/wmts?" +
          "&REQUEST=GetTile&SERVICE=WMTS&VERSION=1.0.0" +
          "&STYLE=normal" +
          "&TILEMATRIXSET=PM" +
          "&FORMAT=image/jpeg" +
          "&LAYER=GEOGRAPHICALGRIDSYSTEMS.ETATMAJOR40" +
          "&TILEMATRIX={z}" +
          "&TILEROW={y}" +
          "&TILECOL={x}"
     {
          Zoom: 12,
          maxZoom: 16,
          attribution:'<a target="blank"
          href="https://www.geoportail.gouv.fr/donnees/carte-
          de-letat-major-1820-1866">IGN-F/Geoportail</a>',
          tileSize: 256,
          transparent: true,
      }
  ).addTo(leafletMap);
```

Ici, nous retrouvons la même URL que pour QGIS sauf que nous allons chercher une carte en particulier celle de l'état-major entre 1820-1866. Il suffit ensuite de l'ajouter à notre carte pour intervertir ensuite entre la carte historique et la carte moderne dans notre projet.

J'ai ensuite rajouté par-dessus un plug-in de Leaflet : Opacity Layer[58] qui permet la superposition de couches en contrôlant l'opacité d'une couche ou d'une autre. Pour notre projet la superposition des couches est très intéressante cela permet de voir les changements de Paris effectués en 170 ans. Pour obtenir un plug-in Leaflet, il suffit de se rendre sur la page du projet qui est la plupart du temps sur GitHub et de télécharger le projet, on l'ajoute ensuite à notre code pour utiliser le plug-in.

[58] https://github.com/lizardtechblog/Leaflet.OpacityControls

Puis j'ai ajouté les fichiers vecteurs au format shapefile provenant du projet ALPAGE et de Paris Open Data. À la différence des cartes qui sont des fichiers rasters[59], les arrondissements et les quartiers de Paris sont fichiers vecteurs, c'est-à-dire des dessins mathématiques composés de forme géométrique et des données qui lui sont associées. Pour les ajouter, j'ai utilisé le plug-in Leaflet-shapefile[60] qui me permet d'ajouter simplement les fichiers shapefile comme n'importe quel autre vecteur et de les utiliser avec un contrôleur de couche que j'ai ajouté à la carte. Ces vecteurs permettent aussi de voir les changements de Paris depuis 1860 avec les nouveaux arrondissements notamment.

En plus de la consultation par index et par vue personnalisée, nous avons intégré un nouvel outil de visualisation qui est dans ma perspective l'élément le plus important du nouveau dispositif. La cartographie permet de voir où les Allemands se situaient dans le Paris du XIXe siècle, elle permet d'avoir un élément interactif qui nous autorise à naviguer parmi les 4772 personnes, la navigation dans l'interface est alors plus ludique, dynamique et aussi plus simple. Enfin, utiliser des bibliothèques Open Source comme Leaflet permet d'anticiper la pérennité de la carte sur le site tant que Leaflet est maintenu par sa communauté et cela permet aussi la reproductibilité de la carte par les usagers, car toutes les données sont accessibles ainsi que le code JavaScript.

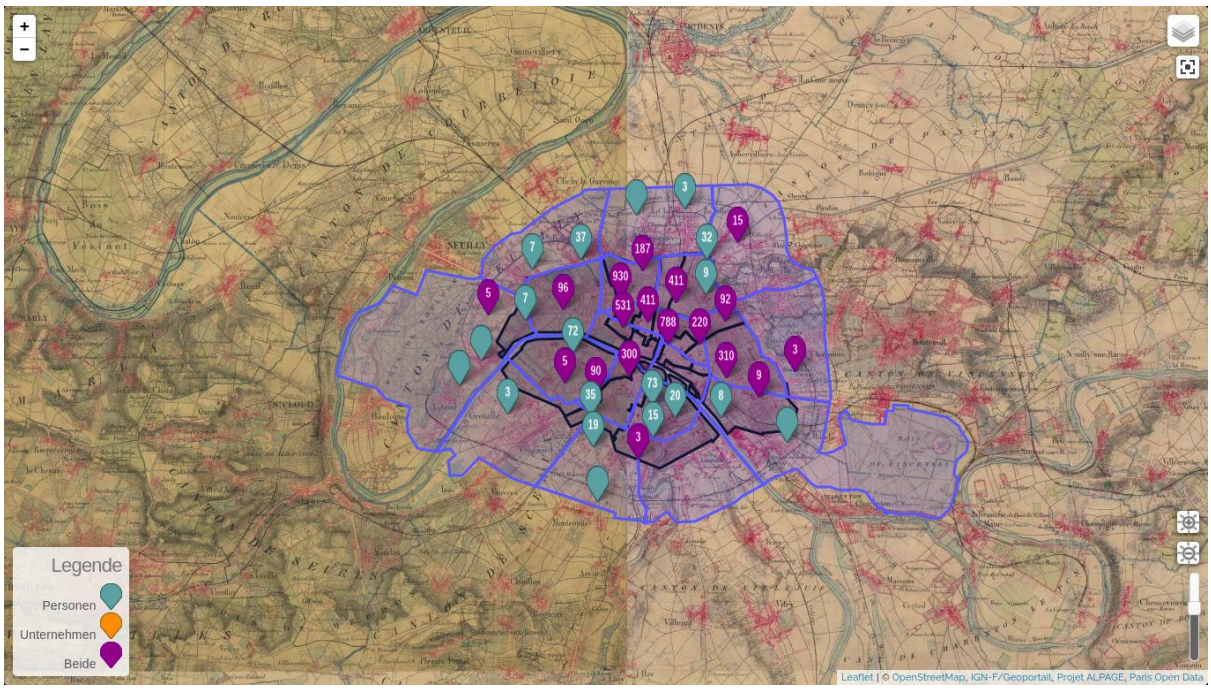


Fig 8. La carte disponible sur le nouveau dispositif

[59] Le fichier raster se compose de milliers de pixels contenant des valeurs représentant des informations. Une carte géoréférencée est un fichier raster.
[60] https://github.com/calvinmetcalf/leaflet.shapefile

Le datapaper : nouvel outil de publication des chercheurs en SHS

Le projet DoRANum (Apprentissage Numérique sur les Données de la Recherche), porté par l'Inist-CNRS et le réseau des Urfist, définit les data papers comme « des articles à part entière suivant le même processus éditorial que les articles scientifiques. Ils ont pour but de rendre des jeux de données accessibles, interprétables et réutilisables » (Doranum, 2017; Kembellec et Le Deuff, 2022). On pourrait également reprendre la définition de Victor Gay : « le datapaper constitue un outil qui peut permettre aux producteurs de données de faire reconnaître leur contribution scientifique en rendant leurs données facilement citables, mais aussi en améliorant la pertinence ainsi que le périmètre de la réutilisation de leurs données ». Les data papers sont définis comme des articles scientifiques, ils ne sont pas là pour présenter les résultats de recherche, mais plutôt pour répondre à des questions, ce que Gérald Kembellec et Olivier le Deuff ont nommé le « Quintilien du data paper » dans leur dossier en 2022. On doit répondre à quoi, qui, comment, pourquoi et où concernant les données que nous présentons dans le data paper. Le data paper est une nouvelle forme d'écriture scientifique qui nécessite des normes, des règles et des formes pour être mené à bien. Dans les sciences techniques et médicales, le data paper reste court et n'excède pas souvent les 10 pages c'est un court descriptif qui vise plus à décrire qu'à expliquer et interpréter les données. Ce data paper s'est développé par la suite dans sciences humaines et sociales pour créer sa propre version de ce que doit être un data paper à la croisée des chemins entre l'article scientifique en respectant sa forme littéraire tout en gardant les codes du data paper en sciences techniques médicales. C'est un écrit d'appui qui vise à expliquer les données et le projet qui l'accompagne (Kembellec et Le Deuf 2022). Malgré certaines consignes de mise en forme, la data paper reste un format d'article scientifique assez libre. Il peut rester dans une forme assez classique d'un article scientifique en Sciences Humaines et Sociales où on va rajouter une partie explicative sur les données de la recherche ou bien on peut produire un article plus technique où on va expliquer les données et faire parler les données avec des graphiques, des cartes, des statistiques à la manière d'un jupyter notebook[61] où le texte, le code et les résultats fusionnent pour produire un data paper.

Pourquoi écrire un datapaper en sciences humaines et sociales ?

Durant les dernières années, les sciences humaines et sociales ont collecté de plus en plus de données grâce notamment à la production et au développement des statistiques. Alors la production des statistiques se développe plus dans des disciplines comme la sociologie, mais

[61] https://jupyter.org/

nous pouvons dire que de manière générale les données sont de plus en plus nombreuses dans toutes les disciplines que comportent les SHS. De plus, en histoire - discipline dans laquelle j'ai co-écrit un data paper, nous trouvons de plus en plus d'articles qui sont dotés d'un graphique ou d'un tableau (Gay. 2021). Autre fait qui peut pousser les chercheuses et les chercheurs à écrire un data paper, c'est améliorer la diffusion et la reproductibilité des données des travaux de recherche. Cette reproductibilité, on y parvient avec les principes de FAIR, car les données reproductibles requièrent d'être trouvables, accessibles, interopérables et réutilisables (Gay, 2021).

Quel est le public d'un data paper ?

Il existe encore peu de revues dédiées aux data papers en SHS, en revanche, il existe un peu plus de revues qui proposent le format data papers aux chercheurs et aux lecteurs. En histoire, le *journal of digital history* propose des articles sous la forme de notebook jupyter. Le producteur du data paper publie pour que les utilisateurs potentiels utilisent les données, mais aussi pour un public plus large, car l'offre de revues est restreinte, par conséquent, on s'adresse à un public qui sort de notre domaine de recherche. Le producteur de data paper doit alors composer avec ce lectorat qui n'est pas spécialiste de sa discipline et adapter son écriture scientifique pour être compris par tous.

Comment construire un data paper ?

Pour un plan générique, nous pouvons choisir le plan proposé par data Scientific Data qui est composé de :

- Contexte et résumé : description des données ainsi que du contexte scientifique de la recherche
- Méthode : comment les données de recherche ont-elles été produites ?
- Fichiers de données : on décrit les données associées au data paper et on fournit les liens pour faciliter l'accès
- Validité des données décrites : Ici on montre comment on a fait pour vérifier les données avec d'autres sources de données, des connexions avec des bases existantes comme lors de l'enrichissement des données
- note d'usage : pour montrer comment on peut se servir des données dans d'autres dataset
- disponibilité du code : on donne la marche à suivre pour trouver le code, comment le reproduire

Le producteur de données doit faire face à la mise à disposition des données qui à la différence d'un article scientifique doivent être présentes lors de la soumission de l'article avec les procédures d'accès qui doivent être explicitement indiquées dans le data paper.
Nous pouvons voir que ce plan de data paper est également conçu pour respecter les principes FAIR qui sont indissociables d'un data paper ou d'un projet de recherche de mon point de vue qui s'est forgé sur l'observation des chercheurs. Le data paper permet de rendre les données faciles à trouver, accessibles, interopérables et reproductibles.
J'ai eu l'occasion de co-écrire un data paper avec mes responsables Mareike Koenig et Gérald Kembellec dans le cadre du projet Adressbuch à la la convention dh nord 2021[62] consacrée aux data papers et à leurs enjeux. Ce data paper était l'opportunité de présenter le projet, ses données ainsi que les différentes phases qui ont composé le projet. En plus de la manière de décrire les données, nous présentons la manière dont les jeux de données sont enrichis et valorisés. Une focale est portée sur les respect des valeurs en humanités numériques, en cohérence avec les demandes spécifiques des historiens et historiennes, mais aussi d'autres métiers comme les généalogistes (Koenig, Kembellec, Virevialle 2021). Nous sommes restés sur la base d'un article scientifique classique, mais nous avons consacré une grande partie de notre article à la partie technique du projet, notamment sur la cartographie qui est un des éléments vedettes de ce nouveau dispositif de consultation. Ce data paper nous a permis de promouvoir le projet en présentant son histoire avec le contexte historique du projet, les premières « vies », comme nous aimons le dire, du projet Adressbuch. Puis, nous présentons les données, c'est-à-dire, comment nous les avons nettoyées en utilisant le logiciel Open Refine, comment nous avons créé la base de données MySQL et insérer les données dans l'interface phpmyadmin. Ensuite, nous décrivons la technique du projet avec la présentation du nouveau dispositif de consultation sous le framework cakePHP, l'hébergement avec notre partenaire : l'*Institut zu Digital Humanities* à Cologne, puis, les nouvelles possibilités de d'affichage des personnes ainsi que les possibilités de téléchargement. Enfin, nous indiquons où les données peuvent être téléchargées via notre compte Github ou encore le dépôt de données Zenodo. Nous donnons alors la possibilité aux chercheuses et chercheurs d'utiliser nos données et même de les reproduire en les proposant dans des formats ouverts tels que le CSV ou le JSON. Il permet aussi dans le cadre du dh nord 2021[63] qui s'intitule : Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux de présenter un exemple

[62] https://www.meshs.fr/page/dhnord2021
[63] https://www.meshs.fr/page/dhnord2021

de data paper en Sciences Humaines et Sociales, comment on l'imaginait. Cet exercice fut très enrichissant surtout lorsque j'ai pu participer à la table ronde pour parler des enjeux du data paper et découvrir les différents data papers qui ont été produits et connaître les projets qui existent aussi pour produire des data papers. Je pense notamment au Journal of Digital History[64] où certains articles de type notebook intègrent des éléments de datapaper ce qui en fait un bon exemple dans la pour la réalisation d'articles interactifs.

Fig 9. Présentation d'un datapaper sur le Journal of Digital History

Nous avons également les nouvelles formes d'écritures qui se développement comme l'outil Stylo[65], projet dirigé par la chaire de recherche du Canada sur les écritures numériques et hébergé par l'infrastructure Huma-Num[66], qui permet l'écriture basée sur le Markdown, permet des sorties sous de multiples formats et la production de métadonnées riches ainsi que des bibliographies structurées. C'est aussi une écriture simplifiée si on la compare aux langages du web balisés tel que le HTML. De plus, avec ses multiples formats il est facile avec Stylo de mettre ses articles, ses data papers en ligne afin de permettre aux usagers de les consulter. Nous pourrions aussi citer l'outil Authorea qui permet l'écriture collaborative d'article scientifique, l'avantage d'Authorea c'est qu'on peut insérer des métadonnées ainsi que des blocs de code, ce qui rend l'article un peu technique et il se rapproche de cette manière du data paper.

[64] https://www.journalofdigitalhistory.org/en
[65] http://stylo-doc.ecrituresnumeriques.ca/fr_FR/#!index.md
[66] https://www.huma-num.fr/

Cependant, j'ai eu des problèmes avec Authorea, car nous avons essayé de rédiger le data paper avec cet outil, mais de nombreux bugs sont apparus et nous avons dû nous tourner vers une solution plus classique.

Le datathon : la collaboration des chercheurs autour d'un projet et d'un événement

Organiser des événements autour des projets de recherche reste un moyen efficace de les promouvoir et permet une meilleure visibilité pour la diffusion des données associées. Afin de promouvoir notre projet Adressbuch, un datathon a été organisé en coordination avec le DKF : le centre allemand d'histoire de l'art à Paris. Un datathon est un événement organisé sur deux ou trois jours généralement qui rassemble des participants travaillant sur des jeux de données afin de trouver des solutions pour les interpréter, enrichir et valoriser les données en question. Le datathon est donc différent du hackathon qui se concentre plutôt sur la programmation ou l'amélioration des logiciels afin de trouver des solutions à certains besoins. Ce datathon a permis de promouvoir également le projet sur les correspondances de Constance de Salm[67] et les projets du DFK à savoir la correspondance Henri Fantin-Latour - Otto Scholderer : une édition numérique d'une correspondance franco-allemande (1858-1906)[68] et le projet *Deutsch-französische Kunstvermittlung* : collection des matériaux avec 6000 articles environ[69] parus dans les revues d'art allemandes et françaises de 1870-1960.

Le datathon s'est tenu en novembre 2021 et a duré 2 jours. L'objectif de ce datathon est de permettre aux participantes et participants d'utiliser les données des projets afin de créer des solutions innovantes et abouties pour les interpréter, valoriser et enrichir. Les participants sont accompagnés d'experts qui vont également faire des sessions pour former sur certains logiciels et des *data buddies* comme moi qui sont capables de parler des données des projets (Adressbuch dans mon cas) pour répondre aux questions des participantes et participants. À la fin du datathon les projets sont présentés ainsi que les résultats, les participants peuvent se baser sur des hypothèses existantes du projet de recherche ou peuvent choisir d'étudier un nouvel angle, une nouvelle hypothèse.

- Comment s'organise le datathon ?

[67] https://constance-de-salm.de/fr/home-francais/
[68] https://quellen.perspectivia.net/fr/fantin-scholderer/start
[69] https://dfk-paris.org/fr/research-project/curation-de-donnees-lexemple-de-la-base-de-donnees-deutsch-franzoesische

Le premier jour, les participants se rencontrent et les *data buddies* présentent les projets de recherche aux participants. Nous avions proposé de former des groupes selon les niveaux :
Un groupe débutant pour découvrir les méthodes numériques dans l'analyse, le traitement et l'enrichissement des données avec des problèmes proposés réalisables en deux jours.
Puis, un groupe initié / confirmé qui dispose d'expérience dans le traitement des données en sciences sociales et qui résout une question scientifique par les organisateurs.
Les groupes ont été formés la première journée, puis ils ont fait connaissance entre eux ainsi qu'avec les données des projets. Ce fut une étape stressante pour moi, car c'est le moment où les participants utilisent en condition réelle pour la première fois le nouveau dispositif de consultation du projet Adressbuch. C'est l'occasion de vérifier si les données sont réellement accessibles, faciles à trouver, interopérables et réutilisables. Si les participants ont des questions, nous sommes à leur disposition pour les aider au mieux sur les jeux de données.
Nous avions fourni les liens pour télécharger les données et ouvert un repository Github et un framapad pour que les participantes et participants puissent mettre à jour les avancées de leurs projets et qu'ils publient à leur tour les résultats de leur collaboration.
Les participantes et participants se familiarisent avec les données et déterminent ensemble sur quels jeux de données ils souhaitent travailler, c'est ensuite à nous de dire si leur projet est réalisable en 2 jours.
Le datathon est aussi l'occasion de promouvoir des projets de recherche similaires qui ne participent pas au datathon mais qui sont présentés à l'occasion d'une conférence. Ce fut le cas avec le projet Quartier Richelieu[70] qui a été présenté la première journée et qui similaire au projet Adressbuch, de par la nature des sources qui sont des bottins du XIXe siècle et de par leur objectif de projet comme la réconciliation des données avec des sources de données externes et également par la volonté de créer une cartographie historique interactive représentant les personnes figurant dans les bottins du commerce.
Le datathon permet également d'échanger avec ces chercheurs et d'échanger sur la façon dont nous avons géré le projet.
Nous sommes d'ailleurs invités à participer à un séminaire en décembre pour présenter les Allemands qui étaient présents dans le quartier Richelieu.
Le deuxième jour, les participantes et participants se concentrent sur leurs projets.
Cependant, il est possible d'assister à des sessions en rapport avec les humanités numériques.

[70] https://www.inha.fr/fr/recherche/le-departement-des-etudes-et-de-la-recherche/domaines-de-recherche/histoire-des-collections-histoire-des-institutions-artistiques-et-culturelles-economie-de-l-art/richelieu-histoire-du-quartier.html?search-keywords=quartier%20Richelieu

Les sessions se répartissent sur trois catégories :

- Logiciels
    - OpenRefine
    - Gephi , logiciel d'analyse et de visualisation de réseaux
    - Omeka S, framework pour la gestion de bibliothèques numériques
    - Palladio
- Sémantique
    - Wikipédia, Wiki Commons, Wikidata
    - Web des données liées en contexte scientifique
    - Travailler avec des référentiels
    - Jupyter notebooks enrichis avec schema.org
- Méthodes numériques
    - éditer des correspondances (TEI/XML)
    - Introduction à l'analyse de réseau
    - le géocodage des données
    - Bloguer sur un *work in progress*
    - Médiation de la recherche via blogs et twitter

Étant donné que ce datathon était le premier, je pense que nous avions planifié beaucoup de sessions et elles n'ont pas eu toutes lieu. L'objectif est de profiter de ces deux jours pour améliorer la littératie numérique des personnes novices dans ces méthodes et leur montrer comment se servir de ces outils pour faire parler leurs données de recherche.
Les sessions qui ont eu lieu sont celles qui furent nécessaires à la réalisation des projets des participants du datathon comme Gephi, OpenRefine, par exemple.
Le datathon est une excellente occasion de rencontrer des experts qui parlent de leurs expériences sur ces logiciels ou/et thématiques. Je peux prendre pour exemple Martin Grandjean, assistant d'histoire contemporaine à l'Université de Lausanne qui est spécialiste de la visualisation des données et l'analyse de réseau, qui a fait une session sur l'utilisation de Gephi très enrichissant.
La dernière journée nous avons essayé de terminer les projets de chaque groupe puis chacun des groupes a effectué une présentation de son travail devant les participants et les experts. Parmi les projets qui m'ont marqué en utilisant les jeux de données provenant d'Adressbuch, un groupe avait réalisé des triplets RDF à partir d'un export des données et ils avaient ensuite enrichi les données avec la base de connaissance Wikidata. Je trouvais que c'était un projet très

abouti en deux jours et surtout cela offrait de nouvelles perspectives pour le projet Adressbuch en termes d'interopérabilité des données. Un autre projet qui m'a marqué est celui auquel j'ai participé avec un groupe sur les données d'Adressbuch également. Ce projet a mis en lumière de nouvelles visualisations cartographiques des données du projet grâce au logiciel QGIS. Ce dernier nous a permis d'analyser les données du projet par catégorie de métier, par sexe ou encore analyser la densité de personnes selon les quartiers et les arrondissements de Paris. Ce projet de deux jours a également mis en lumière de nouvelles interprétations de la répartition des Allemands à Paris au XIXe siècle, il a permis de montrer qu'il existe de potentiels liens entre la répartition des Allemands dans Paris selon les métiers.

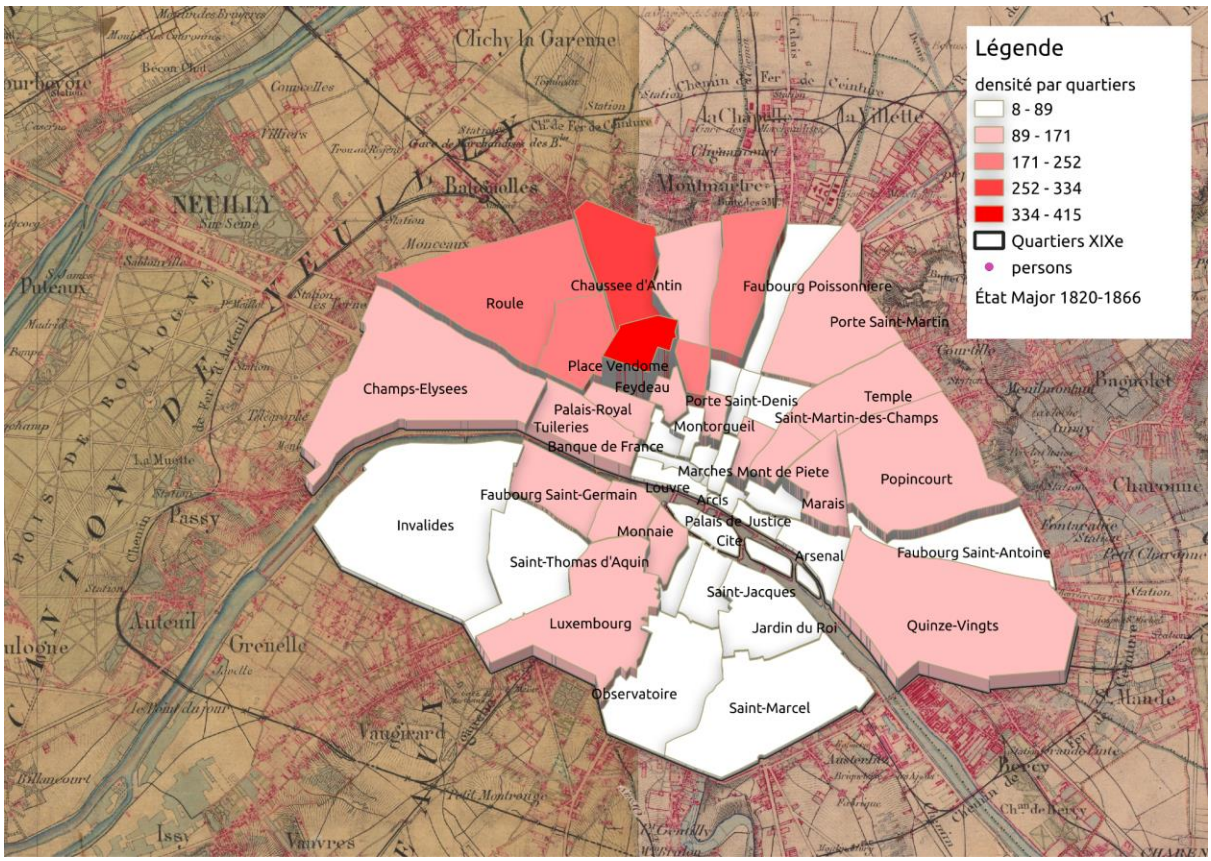


Fig 10. La répartition des allemands selon les quartiers de Paris aux XIXe (réalisation : Evan Virevialle. 2021)

Pour conclure, le datathon est un excellent moyen de promouvoir les données de la recherche, il permet aussi de tester les projets que nous avons créés pour révéler d'éventuels erreurs ou bugs sur les plateformes. Il permet de nous rendre compte à quel point les données sont accessibles, réutilisables ce qui est très important pour un datathon.

Enfin, le datathon permet d'avoir de nouvelles visions au sujet de nos données, on peut voir des interprétations auxquelles nous n'avions pas pensé et également de nouvelles suggestions pour partager nos données comme le triplet RDF par exemple.

Les jupyter notebooks : outils de visualisation et d'interprétation de datasets

Dans la continuité de nos interrogations sur la façon la plus efficace de partager, publier et valoriser les données de la recherche, je souhaitais fixer mon attention sur les Jupyter notebooks. Ce sont des articles exécutables qui permettent d'interagir dynamiquement avec des données et de voir comment ces derniers ont été interprétés. Le projet Jupyter[71] un projet open-source à but non-lucratif, né du projet ipython en 2014. Les Jupyter notebook permettent de créer des articles exécutables dans différentes langues de programmation, les langages de programmation pris en charge de base sont le Julia[72], le Python[73] et le R[74]. Jupyter Notebook est un projet open-source et qui a vocation, selon son équipe, à rester libre et gratuit. Le jupyter notebook est une forme d'écriture qui se développe dans les Humanités Numériques et les Sciences Humaines et Sociales, il permet le *semantic publishing* . Le *semantic publishing* est un mode de publication qui permet la mise en œuvre de documents numériques auto-décrits. On y ajoute des métadonnées qui peuvent être interprétées comme des fragments d'informations par les machines qui comprennent alors la structure et le contexte de cette information (Kembellec, 2019). Avec le Jupyter notebook, on décrit nos données, on les rend accessibles aussi grâce à une adresse de ressource unique (URI), on permet également la reproductibilité entière du notebook grâce à l'endroit où nous l'avons déposé comme Github et les cellules de code qui sont apparentes dans le notebook et donc reproductibles.

- Comment utilise-t-on un jupyter notebook ?

Il existe plusieurs manières d'utiliser un Jupyter notebook. Tout d'abord, à distance via des services qui proposent d'utiliser des notebook sur des machines virtuelles. Dans les sciences Humaines et Sociales, Huma Num propose son Gitlab[75] pour ceux qui ont un identifiant Huma Num pour partager ses notebooks, Google propose également de créer des notebooks via sa

[71] https://jupyter.org/
[72] Le Julia est un langage de programmation pour créer des applications
[73] Le Python est un langage de programmation orienté objet performant pour l'analyse de données
[74] Le R est un langage de programmation similaire au Python pour l'analyse de données statistiques
[75] https://gitlab.huma-num.fr/users/sign_in

plateforme Colab[76]. L'Institut Historique Allemand de Paris nous permet d'utiliser des jupyter notebook via les services de GWDG[77].

La deuxième méthode consiste à installer localement le projet Jupyter sur machine. Il est possible de le faire via Conda[78] qui est un gestionnaire de packages notamment pour le python ou encore sans Conda avec pip[79]. Je suis partisan du deuxième mode d'installation, je trouve que cela est mieux pour gérer les librairies du notebook sur lequel on travaille, car lorsque l'on travaille sur des machines virtuelles à distance comme Colab ou la GWDG, les installations des librairies sont temporaires et doivent être effectuées à chaque fois que nous relançons le noyau du notebook.

- Comment fonctionne le Jupyter Notebook ?

Un Jupyter Notebook est composé de deux types de cellules, la première est une cellule de code dans le langage de programmation que l'on souhaite, ce code peut-être visible, c'est l'option par défaut et c'est également le but des jupyter notebooks de voir comment ils sont conçus. Ces cellules de code sont suivies par des cellules qui interprètent nos codes et nous affichent un résultat qui peut également être caché. Ces deux cellules peuvent être entourées par de la documentation qui explique notre démarche sur l'analyse des données. Cette documentation est réalisée en Markdown, langage simple à prendre en main. Il peut également être accompagné d'éléments HTML[80] comme des balises images, des balises span pour notices bibliographiques ou encore des balises meta pour les métadonnées.

Le notebook peut être sauvegardé au format IPYNB[81] qui est le format des jupyter notebook et ne peut-être lu que par des plates-formes qui sachent lire ce format comme nb viewer[82]. Il peut également être exporté au format HTML ce qui rend sa publication plus simple en ligne, le rendu sera une page web qui affiche nôtre jupyter notebook. Pour ma part, je rends disponible mes Jupyter notebooks sous le format IPYNB dans un repository sur Github, ce qui permet à tout le monde de l'enregistrer et de lire avec son service Jupyter préféré. Je crée également un lien hypertexte entre le notebook sur mon Github et la plateforme nb viewer, ce qui permet à

[76] https://colab.research.google.com/
[77] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen : https://www.gwdg.de/
[78] https://docs.conda.io/en/latest/
[79] Pip Installs Packages : https://pypi.org/project/pip/
[80] HyperText Markup Language
[81] IPython NoteBook
[82] https://nbviewer.org/

ceux qui n'ont pas de service Jupyter de consulter mon notebook et de voir les possibilités en matière d'analyse des données.

- Quelles utilisations du Jupyter Notebook ?

Selon mon avis, le notebook jupyter peut avoir plusieurs utilités. D'abord, on peut utiliser un notebook jupyter pour réaliser un data paper. Je pense que le notebook jupyter qu'on peut également qualifier d'*executable paper* est lié au data paper. Bien que le data paper ait pour but de montrer les données, de les décrire, de donner les accès pour les proposer aux utilisateurs, il peut aller plus loin avec le Jupyter notebook, en interprétant les données, en montrant comment les faire parler et permettant de reproduire leurs résultats. Je renvoie à nouveau vers le *Journal of Digital History* et notamment la présentation de Frédéric Clavert[83] de cette revue lors du dh nord 2021, car c'est pour moi le meilleur exemple d'un article de type notebook qui a su allier la couche des données avec la couche d'explication scientifique.
Une autre utilisation du Jupyter notebook permet de montrer comment faire parler nos données ou même un jeu données plus réduit issu du dataset initial. Avec les différentes librairies, il est possible de montrer des analyses graphiques et statistiques des données afin de vérifier des hypothèses.
Enfin, une dernière utilité des Jupyter notebook est de créer des petits scripts pour récupérer des données provenant de bases de connaissances comme DBpedia, Wikidata ou même des archives pour aider les chercheurs dans leurs collectes des données pendant leurs recherches.

- Exemples de notebook jupyter

J'ai eu l'occasion de co-construire un notebook avec Gérald Kembellec pour une conférence sur les *semantics notebooks* comme nouvelle forme de publication scientifique[84]. L'objectif était de montrer comment réaliser des jupyter notebook, mais surtout comment à partir d'un jeu de données. Nous pouvons réussir à créer des analyses statistiques et les enrichir avec une base de connaissances. Ce notebook a été réalisé avec un jeu de données représentant les nobles allemands vivant à Paris et figurant dans le carnet d'adresses de 1854. Avec les librairies

[83] https://www.meshs.fr/page/decouvrir_le_journal_of_digital_history
[84] https://dhiha.hypotheses.org/tag/semantic-publishing

comme urllib[85] ou SPARQLWrapper[86], les données ont été enrichies avec la base de connaissances DBpedia[87] qui nous a permis de rassembler plus d'informations sur les grades de la Légion d'honneur dont sont pourvus la plupart des nobles de notre jeu de données. Ensuite des graphiques ont été réalisés pour montrer la répartition des différents titres de noblesse dans notre jeu de données, par exemple. J'ai réalisé des éléments de cartographie à partir de ce jeu de données pour montrer la répartition des nobles dans Paris avec la librairie Folium[88].

À partir de ce premier notebook, j'en ai réalisé un deuxième entièrement consacré à la cartographie où je souhaitais montrer les différentes possibilités qu'offre le langage Python avec diverses librairies pour créer des cartes interactives et pour interpréter les données. Avec ce notebook, j'ai utilisé deux jeux de données, le premier représente les nobles allemands à Paris en 1854 et le second toutes les personnes figurant dans le bottin du commerce de 1854. Avec la librairie Geopandas, j'ai pu utiliser des fichiers au format shapefile pour ajouter les arrondissements de Paris ainsi que ces quartiers par-dessus ma carte interactive. J'ai également utilisé les plug-ins de Folium pour créer des cartes synchronisées, rajouter des cartes historiques sur ma carte moderne ou encore créer des clusters de personnes pour rendre ma carte de Paris plus lisible. Ce notebook est accessible en ligne et reproductible[89]. Il est à noter que ce notebook intégrait également des données et métadonnées issues du Web sémantique, de dbPedia en particulier et qui étaient mises en valeur au moyen de schema.org. Cette réalisation suit les préconisations du consortium des moteurs de recherche pour la valorisation de contenus qualifiés et documentés : Google et autres moteurs de recherche sont ainsi plus à même de les valoriser via son knoledgegrap dans les réponses aux questions des usagers : il s'agit d'une bonne pratique diffusionnelle issue des techniques de SEO.

[85] https://docs.python.org/3/library/urllib.html
[86] https://pypi.org/project/SPARQLWrapper/
[87] https://www.dbpedia.org/
[88] https://python-visualization.github.io/folium/
[89] https://github.com/evirevialle/notebook_python_folium

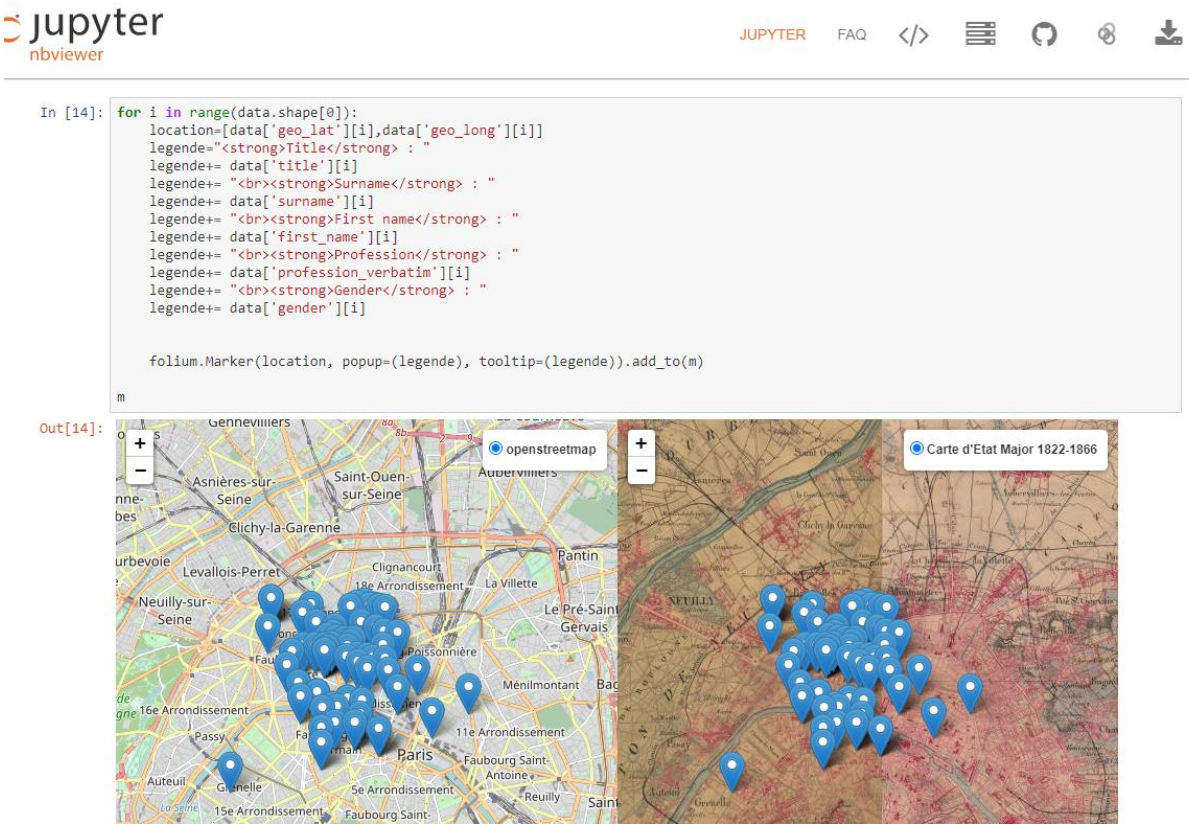


Fig 11. Le notebook cartographique depuis nb viewer

Enfin, un dernier Jupyter notebook que j'ai pu réaliser pour aider une collègue médiéviste à l'Institut Historique Allemand consistait à automatiser le téléchargement d'un manuscrit de l'époque médiévale afin pour analyse. Ainsi avec la librairie Pillow[90] et urllib, j'ai pu télécharger les centaines de pages que contient ce manuscrit en lançant mon script alors qu'il aurait fallu les télécharger une par une sans cette solution.

En conclusion, les jupyter notebook sont d'excellents moyens de partager et valoriser les données de la recherche. Avec ces notebooks, nous pouvons montrer comment interpréter les jeux de données, cela permet d'avoir un côté ludique, car nous pouvons indiquer aux utilisateurs comment nous avons procédé grâce aux cellules de code. Grâce à ces cellules de code, nous pouvons améliorer la littératie numérique des chercheuses et chercheurs, mais aussi faciliter la reproductibilité des notebook. Bien que les chercheurs français en SHS franchiront peut-être plus tard le pas du jupyter notebook comme data paper, ils peuvent s'en servir dès maintenant pour simplifier et automatiser leurs recherches notamment lors de la collecte et du traitement des données. Le sujet est encore pour l'instant en soi une question de recherche.

[90] https://pypi.org/project/Pillow/

Conclusion

Mon mémoire a été l'occasion d'une analyse réflexive concernant la méthodologie que j'ai pu employer au sein du projet Adressbuch. Il me reste à dresser une auto-critique. Il est évident que si je devais recommencer le projet certaines choses seraient plus simples et d'autres plus rapides. Je pourrais prendre pour exemple la cartographie, j'ai maintenant de meilleures connaissances pour réaliser des cartographies interactives qui auraient été utiles lorsque je cherchais comment créer une carte historique. Ce que je retiens surtout du projet Adressbuch, c'était et c'est toujours un projet qui me permet d'apprendre continuellement et de progresser dans les domaines qui m'intéressent comme l'analyse des données, la création d'interfaces ou encore la cartographie. Cette expérience sur le long terme m'a permis d'acquérir de solides connaissances qui me seront utiles dans le secteur où je souhaite travailler, celui du développement d'applications web. J'ajouterais que les résultats produits par ce projet m'ont permis de prolonger mon contrat jusqu'à fin décembre 2022. Grâce à l'Institut historique allemand, je disposerais d'une expérience solide de presque 2 ans. J'ai pu également constituer un réseau avec cette expérience dans le monde de la recherche et obtenir un contrat pour rénover un site web d'un autre projet de recherche en tant qu'autoentrepreneur en conseil sur les systèmes et logiciels informatiques. Cette double expérience sera à mettre en valeur lorsque je serai en recherche d'un autre emploi.

L'ouverture des données dans le domaine de la recherche en sciences humaines et sociales se développe de plus en plus. Les actes au niveau européen et national comme les déclarations de Budapest et de Berlin ainsi que le plan national du gouvernement 2018-2020 montrent la volonté des institutions de favoriser le partage des données dans tous les domaines de recherche. Ce partage des données se fait dans le respect des principes de FAIR. C'est-à-dire que pour avoir une ouverture des données de la recherche optimale, il faut que les données soient faciles à trouver, accessibles, interopérables et réutilisables. Les données de la recherche proviennent de documents, pour les sciences humaines et sociales, ce sont souvent des documents iconographiques, des manuscrits ou autres. Il faut alors que le chercheur dispose déjà de compétences numériques pour collecter ces données, les numériser puis les traiter, les nettoyer, les enrichir puis les partager. Nous avons pu voir que certains domaines de la recherche comme le domaine médical sont déjà dans une logique de partage des données depuis de nombreuses années, car ce partage est essentiel pour faire avancer la recherche.

Bien que le partage des données se développe dans les sciences humaines et sociales, il reste marginal, car ce n'est pas un réflexe automatique de la part des chercheurs de partager les données. Les entretiens et les avis collectés par les enquêtes que nous avons vus dans ce mémoire (Kaden, 2019; Rebouillat, 2019) montrent que les chercheurs en sciences humaines et sociales qui partagent les données de leur recherche sont peu nombreux. Plusieurs raisons subsistent comme la propriété intellectuelle des données, le manque de littératie numérique ou encore des conditions institutionnelles et économiques ne favorisant pas le partage des données dans les SHS. Certaines de ces raisons sont assez simples à résoudre telle que la littératie numérique qui peut-être développée avec des formations ou des outils numériques en appui à la recherche. Cependant, le temps consacré à la recherche ainsi que les moyens financiers sont des maux plus profonds, plus difficiles à modifier sur le long terme. Je pense que ce sont ces derniers maux qui vont causer le plus de tort dans le partage des données de la recherche.

À ces problèmes qui empêchent le partage des données, il faut rajouter la question de la relation entre les chercheurs et les techniciens qui mettent en place les dispositifs de partage de données. Je n'ai pas eu ce problème avec les chercheurs avec qui j'ai travaillé, mais cela ne veut pas dire qu'il n'existe pas. Les chercheurs ont des attentes concernant les plates-formes qu'ils souhaitent mettre en place pour partager les données, ils souhaitent, parfois, certaines normes, langages ou framework. Les techniciens ne peuvent pas tout le temps répondre à ces attentes pour des questions de faisabilité. Il faut alors pouvoir trouver des compromis, discuter entre ces deux corps de métier afin de trouver des solutions. Ces discussions peuvent prendre du temps et cette complexe relation entre le chercheur et le technicien peut dissuader d'entamer des démarches pour construire des plates-formes et *in fine* partager les données.

Enfin, le dernier problème que nous avons soulevé est la question de l'éthique des données dans la recherche. Bien que les sciences humaines et sociales ne soient pas le domaine le plus concerné par l'éthique des données, car nous traitons rarement des données sensibles soumises au RGPD, nous pouvons indirectement faire face à cette question de l'éthique des données. En dehors de la question des données à caractère sensible, nous avons pu voir que la sécurité des données, la crédibilité des données ou encore la propriété intellectuelle sont des problèmes qui doivent être pris en compte lorsqu'on que nous décidons d'ouvrir les données et toujours dans une dimension FAIR. Ces problèmes peuvent également dissuader les chercheurs à partager leurs données surtout lorsqu'ils ne savent pas à quelles fins certaines personnes vont utiliser leurs données. Si mon mémoire s'est beaucoup concentré sur les raisons pour lesquelles les chercheurs ne partagent pas leurs données, il faut aussi insister que de plus en plus de chercheurs partagent leurs données et grâce à cette poignée de chercheurs, les données peuvent

être utilisées afin d'enrichir d'autres projets de recherche et qu'ils soient plus aboutis. C'est l'objectif de l'ouverture des données dans la recherche : faire en sorte que d'autres chercheuses et chercheurs puissent s'en servir afin de l'aider dans leurs recherches.

Ma première partie étant concentrée sur la théorie du partage des données, je souhaitais passer à la pratique dans la deuxième partie en prenant comme cas d'étude le projet Adressbuch sur lequel je travaille depuis plus d'un an. Ce projet historique montre les méthodes et pratiques que nous pouvons adopter pour partager, diffuser les données dans une dimension FAIR.

J'ai d'abord décrit comment nous avons construit la base de données ainsi que le dispositif de consultation pour permettre la facilité à trouver les données, l'accessibilité sur la plateforme pour les données ainsi que l'interopérabilité des données au moyen de fonctions PHP. Ces fonctions permettent que les données soient identifiables par les machines avec des standards définis par schema.org ou encore Zotero. Ensuite, je montre comment nous avons numérisé le document et l'avons rendu accessible selon les standards du web avec la bonne résolution pour la prévisualisation et la visualisation de haute qualité. Le document primaire est primordial, car il permet aux chercheurs de consulter la source d'origine et vérifier par lui-même la véracité des informations présentes sur le dispositif de consultation. Enfin, nous avons montré comment nous gérons les versions de notre dispositif en coopération avec l'institut à Cologne. Bien que j'ai évoqué des problèmes avec l'institut à Cologne sur la gestion des versions du site, Github reste la meilleure solution pour gérer les versions d'une application à distance avec une autre institution. De plus, cela permet de valoriser le code de l'application et le rendre open source, non seulement nous ouvrons les données, mais nous ouvrons aussi le code pour qu'il soit reproductible.

J'ai ensuite voulu montrer comment il est possible de qualifier et valoriser les données dans un projet de recherche. Pour ma part, ce sont des étapes essentielles pour ouvrir les données, elles permettent de fournir des données propres et qu'il sera ensuite simple de récupérer et d'utiliser. Cette étape de qualification des données doit passer par un nettoyage des données où nous allons supprimer le bruit des données, ce sont des caractères qui vont entraver la bonne utilisation des données et qui ne sont pas nécessaires à leur compréhension. Une deuxième étape passera par l'enrichissement des données par des bases de connaissances externes qui vont permettre d'ajouter de la valeur à nos données et qui peut améliorer l'analyse de ces derniers. Une fois cette qualification effectuée nous allons pouvoir les valoriser.

Cette valorisation peut passer tout d'abord par des outils.

Nous avons vu la cartographie comme éléments de valorisation des données. La cartographie améliore la compréhension des données par sa visualisation avec l'analyse temporelle sur

plusieurs périodes qu'elle peut offrir. Elle montre également l'intérêt de l'interopérabilité des données puisque nous utilisons les résultats de recherche d'autres projets historiques tels que le projet ALPAGE. Sans ces données notre carte disposerait de moins d'éléments pour permettre une analyse spatiale des personnes figurant dans Adressbuch.

Les jupyter notebooks sont de fabuleux outils qui permettent de valoriser les données. En utilisant les différentes bibliothèques en python, nous pouvons faire parler nos données et obtenir des interprétations graphiques et géographiques. De plus, ils ont un aspect ludique avec les cellules de codes qui permettent de les rendre reproductibles et d'améliorer la littératie numérique des utilisatrices et utilisateurs qui découvrent le notebook et le python.

La valorisation des données de la recherche peut également passer par l'organisation d'évènements visant à promouvoir les projets et les données.

Le datathon est un excellent moyen de promouvoir des projets de recherche ainsi que leurs données. Il permet aussi de faire un test auprès des participantes et participants pour voir si nos données et le projet atteignent les objectifs attendus. Cela permet d'avoir un retour constructif s'il y a des problèmes à signaler et aussi de voir ce que les personnes peuvent faire avec nos données.

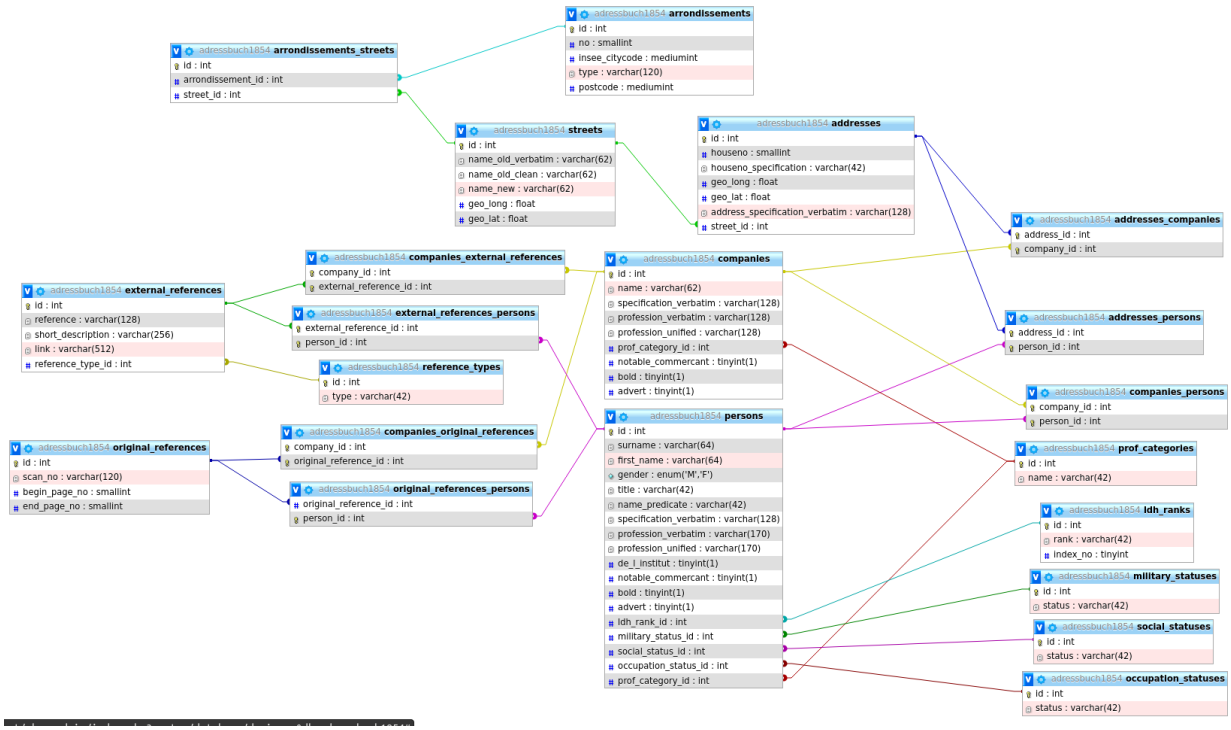
Enfin, le datapaper est un article scientifique qui permet de promouvoir le projet de recherche et ses données, en faisant une description détaillée de ces dernières. Il va permettre également d'expliquer notre démarche et de fournir les liens pour accéder aux données du projet. La promotion passant par les articles et les événements demeurent essentiels pour la visibilité d'un projet de recherche. Je rajouterai que l'ouverture des données dans un projet de recherche fournit une meilleure visibilité que si le projet ne partageait pas les données. Partager ses données est aussi une forme de promotion du projet.

Ce mémoire a pour visée de montrer les points forts, mais aussi les difficultés à ouvrir les données dans un projet de recherche. L'objectif est de montrer aux chercheurs réfractaires à l'Open data, qu'il y a plus d'avantages à partager les données de leurs projets que de les fermer. Cependant, cela suscite des interrogations. Comment pouvons-nous convaincre les chercheurs d'ouvrir leurs données ? Comment améliorer la littératie numérique des chercheurs afin qu'ils s'ouvrent aux méthodes numériques et *in fine* à l'open data ? Quels sont les besoins des chercheurs concernant les méthodes numériques ? Les chercheurs ont-ils besoin de formation d'accompagnement ou bien d'outils pour s'ouvrir aux méthodes numériques ?

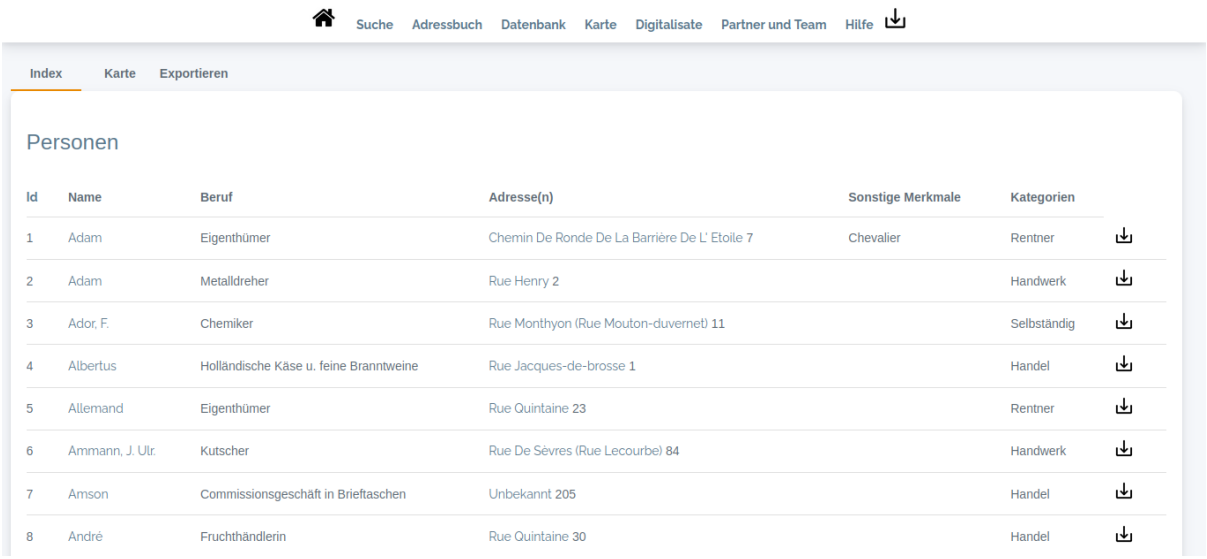
Je pense que ces questions sont légitimes et méritent d'être étudiées afin de faire de l'ouverture des données dans un projet de recherche une norme et non une exception.

Annexes

Annexe 1 : Schéma conceptuel de la base de données du projet Adressbuch



Annexe 2 : Nouveau dispositif de consultation du projet Adressbuch



Annexe 3 : Export JSON depuis le nouveau dispositif


```
{
    "persons": [
        {
            "id": 1,
            "surname": "Adam",
            "first_name": null,
            "gender": "M",
            "title": null,
            "name_predicate": null,
            "specification_verbatim": null,
            "profession_verbatim": "Eigenth\u00fcmer",
            "profession_unified": "Eigenth\u00fcmer",
            "de_l_institut": false,
            "notable_commercant": false,
            "bold": false,
            "advert": null,
            "ldh_rank_id": 1,
            "military_status_id": 2,
            "social_status_id": 1,
            "occupation_status_id": 3,
            "prof_category_id": 8,
            "prof_category": {
                "id": 8,
                "name": "Rentner"
            },
            "occupation_status": {
                "id": 3,
                "status": "Annuitant"
            },
            "social_status": {
                "id": 1,
                "status": "Commoner"
            },
            "military_status": {
```

Annexe 4 : Extrait du document source du projet Adressbuch

Adreßliste

oder

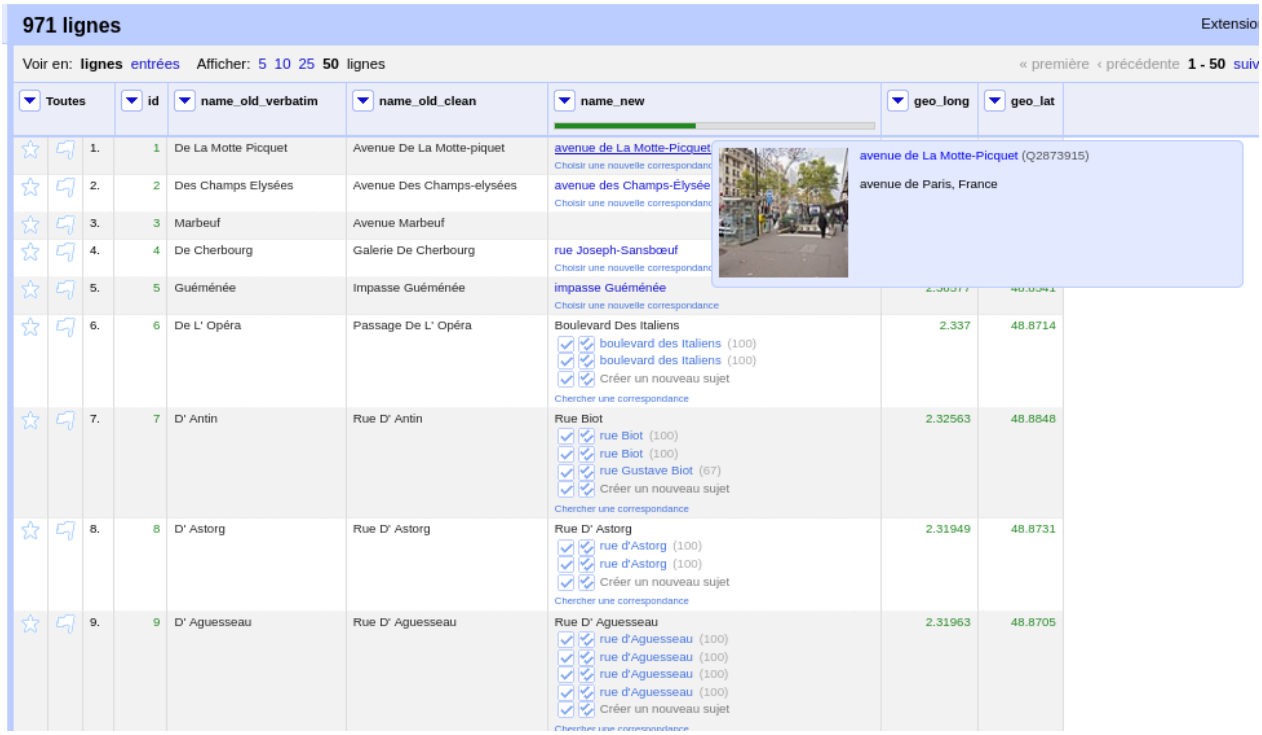
alphabetisches Verzeichniß aller in Paris und seinen Vorstädten wohnenden selbstständigen Deutschen.

Aaron, Bronzehändler, passage Choiseul, 72.
Aaron, Speisewirth, rue Marie-Stuart, 3.
Aaron, Mich. Porzellanfabrik. Bondy, 30.
Abel, Messerschmied, Paradis-Poissonnière, 59.
Abel, coke et escarbilles, Austerlitz-St-Marcel, 46.
Abel. Möblirtes Haus. Avenues des Champs-Elysées, 108, und Chaillot, 111.
Abel. Möbel. Cannettes, 20.
Abel, Orgelspieler, rue Fontaine-St-Georges, 34.
Abel, rue de la Tour-d'Auvergne, 38.
Abert, Bäcker, Ecole-de-Médecine, 64.
Abler, Damenschuhmacher, Neuve-des-Petits-Champs, 91.
Abraham, E. Stickereien. Ste-Apolline, 7.
Abraham, Michael. Mützenfabrik. Temple, 55.
Abraham, L., Arzt, Ponthieu, 14.
Abraham, Buchbinder, Parchemmerie, 2.
Abraham, Domänenverificator, Martyrs, 63.
Abraham-Dubois *, Referendar am Rechnungshofe, Cassette, 7.
Abraham-Cliver, Eduard. Kravatten und Foulards. Vendôme, 7.
Abraham-Levy, Bankier, rue Grange-Batelière, 13.
Abrahams junior, J. D. Feine Steine. Valois-Palais-Royal, 43.
Achtsnit, Steinschneider, rue des Enfants-Rouges, 1.
Acker, Papierhändler, rue Neuve-des-Petits-Champs, 29.
Ackermann, Speisewirth, rue du Pont-Louis-Philippe, 2.
Ackermann, Sattler, Faub.-St-Denis, 104.
Adam, Alfred, Baumeister, Lille, 1.
Adam *, Advocat, Faub.-St-Denis, 41.
Adam, G. A., Advocat, Chaussée-d'Antin, 19.
Adam, Emil, Advocat 1ster Instanz, place St-Germain-l'Auxerrois, 41.
Adam. Vergoldete Schmuckfachen. Temple, 176.

Annexe 5 : Extraction du texte provenant des numérisations

```
20

Adam. Fabrik vergolv. Schmuck
fachen. Notre-Dame-de-Na-
zareth. 6

Adam. HofentrÃ¤ger.
Augustins. 67.

Adam. sen.. Brofchirer, Grand-
Chantier, 3.

Adam, Schuhmacher, Petit-
Carreau. FE

Adam, Ã€, u. Comp.  Com-
miffiong- er Speditionsger
fcpÃ¤ft. Pl. des Vietoires, 3.

Adam, Branntwein - Abzieh er,
Poissonniere. 26.

Adam, 0. *, Nathsherr am
Nechnungehofe, St-Domini-
que, 50.
```

Annexe 6 : Enrichissement des données avec Open Refine depuis Wikidata



Bibliographie

Arruabarrena Béa, « Datavisualisation : principes, enjeux et perspectives pour des utilisateurs non experts », dans : Évelyne Broudoux éd., *Big Data - Open Data : Quelles valeurs ? Quels enjeux ?Actes du colloque « Document numérique et société », Rabat, 2015*. Louvain-la-Neuve, De Boeck Supérieur, « Information et stratégie », 2015, p. 151-163. DOI : 10.3917/dbu.chron.2015.01.0151. URL : https://www-cairn-info.ezpaarse.univ-paris1.fr/---page-151.htm

Rebouillat Violaine, Ouverture des données de la recherche : de la vision politique aux pratiques des chercheurs. Sciences de l'information et de la communication. Conservatoire national des arts et métiers - CNAM, 2019. Français. ⟨NNT : 2019CNAM1254⟩. ⟨tel-02447653⟩

Clavert, Frédéric. *et al.*, « Découvrir le *journal of digital history* », dans dhnord2021, « publier et partager les données de la recherche », 2021. https://www.meshs.fr/page/decouvrir_le_journal_of_digital_history

Fickers, Andreas., et Clavert, Frédéric. (2021). « On pyramids, prisms, and scalable reading.», *Journal of Digital History*. https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb

Jacquemin Bernard, Schöpfel Joachim et Fabre Renaud, « Libre accès et données de recherche. De l'utopie à l'idéal réaliste », *Études de communication* [En ligne], 52 | 2019, mis en ligne le 15 décembre 2020, consulté le 15 février 2022. URL : http://journals.openedition.org.ezpaarse.univ-paris1.fr/edc/8468 ; DOI : https://doi-org.ezpaarse.univ-paris1.fr/10.4000/edc.8468

Jacquemin Bernard, Schöpfel Joachim, Chaudiron Stéphane *et al.*, « L'éthique des données de la recherche en sciences humaines et sociales », dans : Laurence Balicco éd., *L'éthique en contexte info-communicationnel numérique. Déontologie, régulation, algorithme, espace public*. Louvain-la-Neuve, De Boeck Supérieur, « Information et stratégie », 2018, p. 71-86. DOI : 10.3917/dbu.balic.2018.01.0071. URL : https://www-cairn-info.ezpaarse.univ-paris1.fr/---page-71.htm

Kaden Ben, « Pourquoi les données de recherche ne sont-elles pas publiées ? », *Études de communication* [En ligne], 52 | 2019, mis en ligne le 01 janvier 2021, consulté le 22 février

2022. URL : http://journals.openedition.org.ezpaarse.univ-paris1.fr/edc/8783 ; DOI : https://doi-org.ezpaarse.univ-paris1.fr/10.4000/edc.8783

Kembellec, Gérald. « Produire, analyser et partager des données ouvertes en Humanités Numériques : quelques bonnes pratiques. », *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances?*, Oct 2019, Montpellier, France. ⟨hal-02306958⟩, https://hal.archives-ouvertes.fr/hal-02306958/

Kembellec, Gérald. et Le deuff Olivier., « Poétique et ingénierie des data papers », *Revue française des sciences de l'information et de la communication*, n°24, janvier 2022, http://journals.openedition.org/rfsic/12938 ; DOI : https://doi.org/10.4000/rfsic.12938

Kembellec Gérald. (2019). « Semantic publishing, la sémantique dans la sémiotique des codes sources d'écrits d'écran scientifiques dans Les Enjeux de l'information et de la communication » numéro 20/2, p. 55-72. https://www.cairn.info/article.php?ID_ARTICLE=ENIC_027_0055&contenu=article

Kembellec Gérald. Bibliographies scientifiques : de la recherche d'informations à la production de documents normés. Sciences de l'information et de la communication. Université Paris VIII Vincennes-Saint Denis, 2012. Français. ⟨tel-01578217⟩

OCDE (2007), *Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics*, Éditions OCDE, Paris, https://doi.org/10.1787/9789264034020-en-fr.

Reymonet Nathalie. *et al.* (2018). « Réaliser un plan de gestion de données "FAIR" : modèle », https://archivesic.ccsd.cnrs.fr/sic_01690547v2/