Data paper en humanités numériques : Adressbuch 1854

Mareike König, Gérald Kembellec, Evan Virevialle

Data paper en humanités numériques : Adressbuch 1854

The Third Life of the Historic Adressbuch of German Migrants in Paris in 1854: A Data paper

Mareike König, mkoenig@dhi-paris.fr (1), Gérald Kembellec, gerald.kembellec@lecnam.net (1,2), Evan Virevialle, evirevialle@dhi-paris.fr (1,3)

(1) Institut Historique Allemand (2) Laboratoire Dicen-IdF, Cnam (EA 7339) (3) Université Panthéon-Sorbonne

Correspondant : gerald.kembellec@lecnam.net

Mots clés : Base de données, Histoire numérique, Allemands, Paris, Migration, XIXe siècle

Keywords : Database, Digital History, Germans, Paris, Migration, 19th century

Mareike König est directrice adjointe et directrice du département Humanités Numériques à l'Institut historique allemand (IHA). Elle est à l'origine du projet « *Adressbuch der Deutschen in Paris von 1854* ». Historienne de l'histoire franco-allemande du XIX^{e} siècle, ses recherches portent également sur l'Histoire à l'ère numérique et sur la communication scientifique dans les médias sociaux.

Gérald Kembellec est maître de conférences en sciences de l'information et de la communication au Cnam, il anime la thématique « Data, médiation, valorisation » du laboratoire « Dispositifs d'Information et de Communication à l'Ère du Numérique ». Il était détaché au département des humanités numériques de l'IHA en 2020 et 2021.

Evan Virevialle est étudiant en 2^{e} année du Master Communication du savoir, technologies de la connaissance et management de l'information à l'université Paris 1 Panthéon-Sorbonne. Il est également assistant de recherche à l'IHA, il y collabore notamment au projet « *Adressbuch der Deutschen in Paris von 1854* » sur les aspects de qualification et d'enrichissement des données, d'interface et de cartographie.

Résumé

Ce chapitre présente un *data paper* qui porte sur les jeux de données liés au projet « *Adressbuch der Deutschen in Paris von 1854* » au long cours d'histoire sur l'immigration des Allemands à Paris au XIX^{e} siècle réalisé à l'Institut historique allemand de Paris. Le projet a pour objectif de mettre à la disposition des chercheuses et chercheurs et généalogistes, dans une interface, les informations sur les individus et entreprises allemands installés à Paris en 1854. Ces informations sont issues d'un document historique, une sorte de bottin de commerce ou pages jaunes des Allemands à Paris.

Outre la description de méthodes de collecte des données, ce *data paper* présente la manière dont les jeux de données sont enrichis et valorisés dans le dispositif de consultation. Nous portons une attention particulière sur le respect des valeurs des humanités numériques et de la

science ouverte, en cohérence avec les spécificités des historiennes et historiens, ainsi que les généalogistes amenés à le consulter.

Abstract

This chapter presents a data paper on the datasets related to the project "Adressbuch der Deutschen in Paris von 1854", a project on German immigration to Paris in the 19th century carried out at the German Historical Institute in Paris. During the research project, a database was built with over 4700 addresses of German individuals and companies who settled in Paris and its suburbs in 1854. The information is derived from a historical document, a kind of business directory or yellow pages of Germans in Paris established in 1854.

In addition to the description of the data collection methods, this data paper presents the way in which the datasets are enriched, valorised, presented and visualised on historic city maps. Particular attention is paid to the respect of standards and values of digital humanities and open science alike, in coherence with the specific needs of the main users of the database (historians and genealogists).

Introduction

Dans la recherche historique, les bases de données sont souvent créées en accompagnement de projets de recherche. Leur création et leur indexation s'orientent la plupart du temps vers les thèmes centraux des projets. L'utilisation ultérieure des données par d'autres n'est pas envisagée, ou seulement tardivement. Pourtant, la plupart du temps, beaucoup de temps et d'argent ont été investis dans les projets, notamment par le biais de travaux manuels de saisie et d'indexation. Il en a été de même pour le projet « Annuaire des Allemands à Paris de 1854 », une base de données accessible en ligne depuis 2006. Des résultats essentiels du projet de recherche qui l'accompagnait avaient été publiés dans des publications classiques à l'époque. Ce qui manquait cependant, c'était une description de l'ensemble de données sur lequel repose la base de données. Le présent *data paper* vise à combler cette lacune. Il décrit et documente, sous forme de texte et non de simples métadonnées, le jeu de données ainsi que les conditions et les réflexions fondamentales qui ont présidé à sa création, à sa révision et à sa republication de 2020/2021. Le *data paper* accompagne ainsi la curation et la *fairification*[1] des données, tout comme la nouvelle version du site Web et la cartographie qui l'accompagne. Il doit permettre aux historiennes et historiens de travailler avec les données, de les comprendre, de les enrichir, de les évaluer et de les réutiliser. Nous souhaitons en même temps faire connaître le projet et notre démarche de curation des données et de remodélisation du dispositif de consultation. Le destin du projet et de ses données est intéressant parce qu'il devrait être typique de nombreux projets de numérisation du début des années 2000. La procédure peut servir d'inspiration, sans toutefois vouloir ou pouvoir s'imposer comme procédure de *best practice*. Pour l'écriture du *data paper* à 6 mains, nous n'avons pas utilisé un modèle, mais nous avons mélangé des éléments d'articles en histoire classique avec des éléments qui s'orientent davantage vers la description de données et de métadonnées, afin de rendre le *data paper* et les données qu'il décrit accessibles à tous.

[1] L'institut historique allemand souhaite aujourd'hui libérer et diffuser les données ainsi que le code sous une licence ouverte pour permettre et encourager leur réutilisation dans la recherche. Voir les principes FAIR pour rendre les données de la recherche faciles à trouver, accessibles, interopérables et réutilisables par l'homme et la machine : https://www.ouvrirlascience.fr/fair-principles/.

Présentation du projet - contexte historique

Au moment de la révolution de 1848, on estime que quelque 60 000 Allemands peuplaient les rues du Paris préhaussmannien et de ses banlieues (Grandjonc, 1974 ; Grandjonc, 1983 ; König, 2006 ; König, 2003). Par Allemands, il faut comprendre immigrants temporaires ou permanents, intellectuels, artisans, ouvriers — qualifiés ou non — et domestiques de culture et de langue germanique. Ils étaient originaires des différentes principautés composant alors la Confédération germanique, de l'Autriche, de la Suisse ou d'autres régions où l'on parlait l'allemand. C'est le sort des habitants de cette « colonie germanophone », ainsi perçus par leurs contemporains, que le projet proposait d'étudier avec une approche d'histoire sociale (König, 2003). Les immigrés germanophones constituaient pendant longtemps au XIX^{e} siècle, avant les Belges et les Italiens, le groupe d'immigrés le plus nombreux (Werner, 1995, p. 199-200 ; Noiriel, 1992). Ainsi vers 1848, les Allemands représentaient plus de 35 % des étrangers à Paris. Compagnons et ouvriers suivaient une partie de leur formation à Paris et souhaitaient souvent en même temps échapper aux mesures politiques des gouvernements restaurateurs des États allemands pour vivre plus librement sous la monarchie de Juillet. À Paris, ils se sont associés à des militants politiques, des réfugiés, des artistes et des écrivains germanophones pour fonder le premier mouvement d'ouvriers allemands (Schieder, 1963).

La plupart des compagnons, apprentis et ouvriers allemands vivotaient dans les quartiers pauvres de Paris, souvent à la limite du minimum vital. D'autres, comme les artisans et les commerçants allemands passaient par Paris pour y acquérir de nouveaux savoir-faire professionnels ou pour élargir leurs réseaux commerciaux. Nombre d'entre eux restaient, tout comme l'immigration bourgeoise et intellectuelle de l'époque, dans la capitale française pour profiter des possibilités économiques qu'elle offrait en tant que centre de l'art, de la mode et du savoir-vivre. Contrairement aux ouvriers non qualifiés et aux apprentis, ces riches artisans, commerçants et bourgeois allemands étaient en voie d'intégration. Ils apprenaient la langue du pays, épousaient des Françaises et demandaient parfois leur naturalisation (Dietrich et Varnier, 1987).

Rétrospective du projet et présentation de la source

Le projet de recherche sur l'immigration germanophone à Paris au XIX^{e} siècle lancé par l'Institut historique allemand de Paris en 2003, était accompagné par la mise en données d'un document unique à bien des égards, cofinancé à l'époque par la fondation Gerda Henkel : *l'Adressbuch der Deutschen in Paris für das Jahr 1854*[2] (Kronauge, 1854). Il ne reste aujourd'hui que quatre exemplaires de cet ouvrage, dont un conservé à la Bibliothèque historique de la Ville de Paris[3].

Cet annuaire compile sur 248 pages[4] un total de 4 772 adresses de particuliers et d'entreprises germaniques domiciliés à Paris et dans sa banlieue en 1854. On y trouve notamment les membres des classes moyennes et de la bourgeoisie, avec un éventail des professions et

[2] Traduction du titre en français : Annuaire des Allemands à Paris de 1854.

[3] Selon l'OCLC l'ouvrage n'est accessible qu'à Paris (à la BnF et à la BHVP), Londres (*British Library*) et *Bonn Bibliothek der Friedrich Ebert Stiftung*. Voici l'url ark de la notice du document archivé sous la cote 703983 à la Bibliothèque historique de la ville de Paris.

https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0000884072

[4] Une erreur de pagination l'amène en réalité à 242 pages. L'erreur se trouve dans les trois exemplaires qui ont été vérifiés au cours du projet : deux à Paris et à un Londres.

métiers extrêmement large : surtout des artisans comme des ébénistes, menuisiers, tailleurs, orfèvres et imprimeurs, mais également des entrepreneurs et des négociants, ainsi que des libraires, des banquiers, des professions libérales comme des médecins, écrivains, architectes, artistes et musiciens. En compilant les adresses, l'initiateur, un certain F.A. Kronauge, dont nous savons seulement qu'il était professeur de langue et qu'il habitait rue de Richelieu, entendait « offrir à nos compatriotes un moyen de se retrouver et de se faire connaître les uns des autres » (Kronauge, 1854). La source est d'une grande importance pour la recherche historique : on y trouve 4 772 particuliers des 12 245 immigrés allemands recensés officiellement en 1851[5]. Parmi eux, notons des personnes célèbres comme le poète Heinrich Heine, l'architecte de la Gare du nord Jakob Ignaz Hittorff, le libraire Klincksieck et quelques membres de la famille de Rothschild. Mais encore l'importance de la source réside dans le fait qu'y sont répertoriées des personnes inconnues, qui représentent un échantillon de l'immigration allemande aisée. Ces noms peuvent constituer le point de départ pour une recherche sur des parcours migratoires individuels, en croisant les noms avec d'autres sources comme les registres des mariages, des baptêmes et de l'état civil, des actes de notoriétés ou les actes de naturalisation et d'autres sources encore (König, 2003).

Ce projet, initié au début des années 2000, s'est achevé, dans sa version initiale, en 2006. Il s'agissait à l'époque d'un travail historique dans une mouvance innovante, pionnière même, si l'on considère que les humanités numériques en tant que concept problématisé, datent de 2004 (Schreibman, Siemens et Unsworth, 2004). L'objectif était de créer une base de données et de la publier en ligne pour ainsi mettre à disposition des chercheuses et chercheurs, généalogistes et internautes, les noms et adresses des immigrés allemands compilés dans l'*Adressbuch.*

Les informations enregistrées manuellement par l'équipe de recherche dans les années 2000 ont été versées dans une base de données de type FileMaker. Il s'agissait de fiches mentionnant les noms, les prénoms, les adresses, les professions, les noms de rues et les arrondissements (d'avant 1860 et d'aujourd'hui[6]). Malheureusement, visualiser de manière interactive - avec des filtres - les adresses sur un plan de Paris de l'époque, représenter la répartition géographique des immigrés allemands était en 2003/2006 techniquement trop complexe. De plus, les données brutes n'étaient pas disponibles en *open data* sur le dispositif initial : à l'époque, on ne pensait pas à proposer le téléchargement des données ; le site Web et les données étaient placés sous copyright.

Environ 15 ans plus tard, le corpus est régulièrement consulté et fait l'objet de demandes d'informations notamment de la part de généalogistes et des historiennes et historiens menant une recherche prosopographique. Le projet *Adressbuch 1854* a donc été relancé en 2020 et depuis 2021 il bénéficie de l'aide de l'*Institut für Digital Humanities* de Cologne. Ce dernier fournit l'infrastructure d'hébergement du dispositif de consultation des données du projet. Les données brutes sont hébergées sur la plateforme Zenodo, le code du dispositif de consultation est déposé sur les comptes GitHub de deux institutions impliquées dans le projet.

---

[5] Le chiffre était probablement plus élevé que compté dans le recensement, voir Grandjonc 1972. *L'Adressbuch* se vante d'être complet (*vollständig*), néanmoins une recherche rapide montre aussitôt les lacunes que les rajouts comme le fameux poète polonais Adam Mickiewics ou encore une librairie espagnole et une librairie polonaise.

[6] Dans ce chapitre, nous mentionnerons les arrondissements d'avant 1860, en vigueur lors de la publication de la source principale.

Les finalités et enjeux du nouveau projet

La transformation de la base obéit à des objectifs et enjeux pluriels. Il s'agit avant tout de nettoyer, structurer et enrichir les données originelles pour permettre des usages plus larges en relation avec les besoins des historiens et historiennes, généalogistes et plus généralement des érudits et érudites. Ensuite, pour respecter les bonnes pratiques, nous désirons proposer un jeu de données qualifiées, accessible en *open data*. Les défis d'une modélisation en diachronie ont été nombreux : si les métiers et contextes socioprofessionnels ont évolué avec la révolution industrielle, il en va de même pour la terminologie associée. De plus, le paysage de la ville de Paris, sa toponymie et son découpage administratif ont subi de profondes mutations en 170 ans notamment avec les travaux de Haussmann dans les années 1850/1860 (Gaillard, 1997 ; Pinon, 2012).

Ensuite, ce corpus mérite une remodélisation à visée humaniste pour correspondre aux sujets et méthodes des sciences humaines et sociales. Cette visée humaniste consiste à offrir une base de connaissances, en l'occurrence une base de données, dont la structure réponde à la fois à la vision historienne et à la possibilité de partage selon les standards FAIR des humanités numériques tout en respectant les règles de modélisation informatique. Pour questionner les données, il faut qu'elles soient finement articulées sous la forme d'un modèle qui valorise les problématiques historiennes. Chaque concept présenté dans la base est ainsi associé aux autres selon un point de vue historique. Les entités et associations telles que modélisées dans le paragraphe sous-titré « Penser les immigrés allemands dans un contexte socioprofessionnel et sur un territoire » seront à même d'éclairer le public cible selon ses propres critères. Pour faire face à la complexité du phénomène, appuyé conjointement sur plusieurs disciplines - tant sur le fond que sur la forme - le modèle structurant les données issues de l'*Adressbuch* a donc été pensé de manière transdisciplinaire. Une fine collaboration entre informaticiens, documentalistes et historiens a permis d'en définir le fond, la forme et les méthodes d'accès. Il n'est pas possible ici de cloisonner le rôle de chacun : chaque aspect a été discuté pour répondre à la fois une vérité sociohistorique, une réalité technique, des besoins de consultation pluriels et enfin une méthode et de bonnes pratiques infodocumentaires : il s'agit là d'une réelle collaboration transdisciplinaire.

L'interface initiale du compagnon Web FileMaker ne correspond plus aux canons de consultation, d'ergonomie, d'accessibilité et de sémantique du Web. Il fallait donc repenser le dispositif de consultation en cohérence avec les nouveaux standards (design responsive, structuration sémantique des contenus...), les besoins de filtrages thématiques historiques, sociaux et cartographiques, mais aussi en cohérence avec les méthodes des humanités numériques.

Matériau de recherche et méthodes

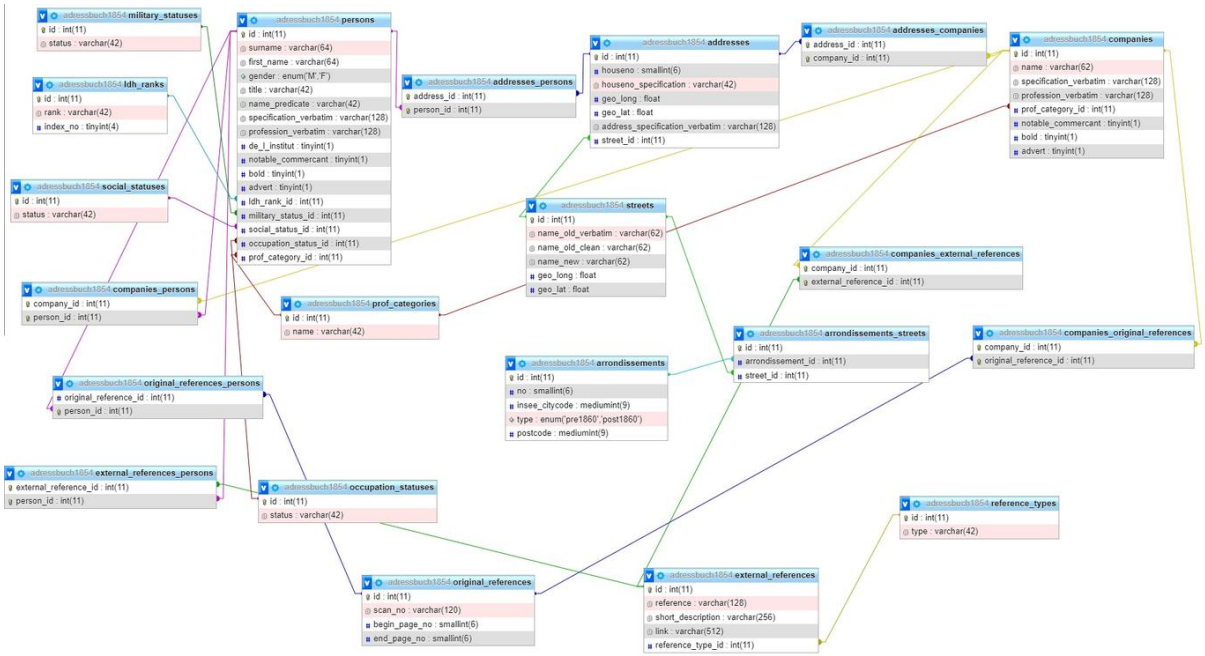
Le modèle de données

Penser les immigrés allemands dans un contexte socioprofessionnel et sur un territoire

La base a été remodélisée dans une optique historienne avec l'outil de conception Mocodo puis implémentée sur l'application de gestion de bases phpMyAdmin. L'étape suivante consistait à créer les tables de données selon notre modèle. L'outil libre OpenRefine a permis de fragmenter le fichier tabulaire initial de données en plusieurs tables qui seraient ensuite exportées au format SQL : la base contient les tables courantes avec les données du projet (personnes, compagnies, professions ou rues...) et les tables associatives qui permettent la jonction entre les différentes tables courantes (ex. : la table qui permet la jonction entre les

personnes et les compagnies : companies_persons). OpenRefine permet d'exporter les tables au format SQL pour ensuite les importer dans la base de données. En tout, ce sont 21 tables qui ont été créées et qui permettent d'interagir efficacement avec la base de données (voir figure 1). Une fois la base de données complète et stabilisée, elle fut transférée sur un serveur en ligne géré par l'*Institut für Digital Humanities* de Cologne.

Figure 1. Le modèle de données



Légende : Le modèle physique de données après une modélisation conceptuelle, licence *Creative Commons* BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/, DHI Paris.

Questions pragmatiques de transfert numérique

Comme spécifié précédemment, les données de la base initiale ont été exportées dans un format tabulaire, dans un fichier unique CSV encodé en UTF-8. Le CSV est un format qui ne nécessite pas de logiciels propriétaires pour être lu. Un simple éditeur de texte suffit pour lire un fichier CSV. Sa polyvalence est un avantage pour la diffusion et la réutilisation des données par les utilisateurs selon les principes FAIR. La norme d'encodage UTF-8 permet que le fichier soit rétrocompatible avec d'autres normes comme la norme ASCII et il permet de prendre l'ensemble du répertoire universel de caractères développé par l'ISO. Cela permet que tous les caractères spéciaux soient lus dans le fichier CSV. Elles ont été ensuite enrichies et réconciliées à l'aide d'OpenRefine avec des notices d'autorité et autres sources de données externes disponibles en ligne comme les coordonnées GPS[7]. Le bruit (parenthèses,

[7] Par exemple : WikiData, BnF, GnD, ViaF, Geonames, etc.

crochets...) a été éliminé des données avec le même logiciel au moyen d'expressions régulières du langage GREL[8]. Enfin, les données ont été segmentées pour correspondre au modèle présenté préalablement et ont été intégrées à une nouvelle base de données relationnelles.

Diplomatique numérique en histoire

Dans une optique historienne, la diplomatique se base sur le lien étroit qu'il peut y avoir entre les informations et l'archive : une base de données ne saurait se substituer dans ce contexte aux documents primaires qu'il convient de sourcer. Il est de plus indispensable d'associer les données et les métadonnées, mais aussi un *fac-similé* de qualité des pages de l'ouvrage. Il a été décidé de procéder à la numérisation des archives mobilisées pour offrir une alternative de qualité à la consultation physique.

La Bibliothèque historique de la Ville de Paris a mis à disposition de l'équipe le document original pour la numérisation. Le format d'encodage JPEG a été choisi pour les images numérisées pour sa compatibilité universelle et son ratio qualité-compression convaincant qui permet des impressions et visualisations de haute qualité (selon la résolution choisie). Les métadonnées de chaque numérisation ont été extraites pour permettre de proposer 2 tailles de vues différentes des pages numérisées : le format vignette pour l'affichage-écran et la haute définition pour les examens minutieux ou les tirages papier.

Documentation et classement de fac-similés

Les métadonnées techniques ont été extraites en ligne de commande avec un script Bash à l'aide de la commande Shell « file » pour obtenir les informations relatives aux images : la résolution, le format et la taille[9]. Ces métadonnées ont été associées au nom de l'image et à son répertoire de stockage selon les règles de nommage que nous définissons pour uniformiser et un plan de classement puis enregistrées dans un fichier *Comma Separated Values* (CSV). Ces informations ont permis, grâce aux propriétés des images, de procéder à leur redimensionnement par lot. Pour l'affichage en vignette, nous avons choisi une résolution de 72 dpi et une taille de 400 x 800 px, soit la moitié de la taille des numérisations d'origine (voir figure 2). Le redimensionnement s'est également effectué en ligne de commande avec un script Bash et à l'aide de la bibliothèque Imagemagick[10] qui permet le traitement d'images en lots pour la modification de leurs résolutions et de leurs tailles. L'objectif est de proposer un premier aperçu sans ralentir le chargement de la page de consultation avec une image volumineuse. Pour l'affichage en haute définition (HD), la résolution en 300 dpi a été retenue comme proposée par les règles de numérisation de la DFG[11]. C'est la résolution minimale pour proposer des visualisations et des numérisations de haute qualité compatible avec l'impression. Pour ne pas surcharger l'interface avec la taille et la résolution de la page numérisée, cette dernière s'ouvre dans un autre onglet lorsque l'on souhaite qu'elle s'affiche en plein écran[12].

---

[8] Google Refine Expression Language, voir https://docs.openrefine.org/manual/grel

[9] Voir la documentation : https://linux.die.net/man/1/file

[10] Voir le site de l'application : https://imagemagick.org

[11] Voir le guide de bonnes pratiques « *DFG-PraxisregelnDigitalisierung* » : https://www.dfg.de/formulare/12_151

[12] Voir le dépôt principal du projet sur Zenodo : https://doi.org/10.5281/zenodo.7427439

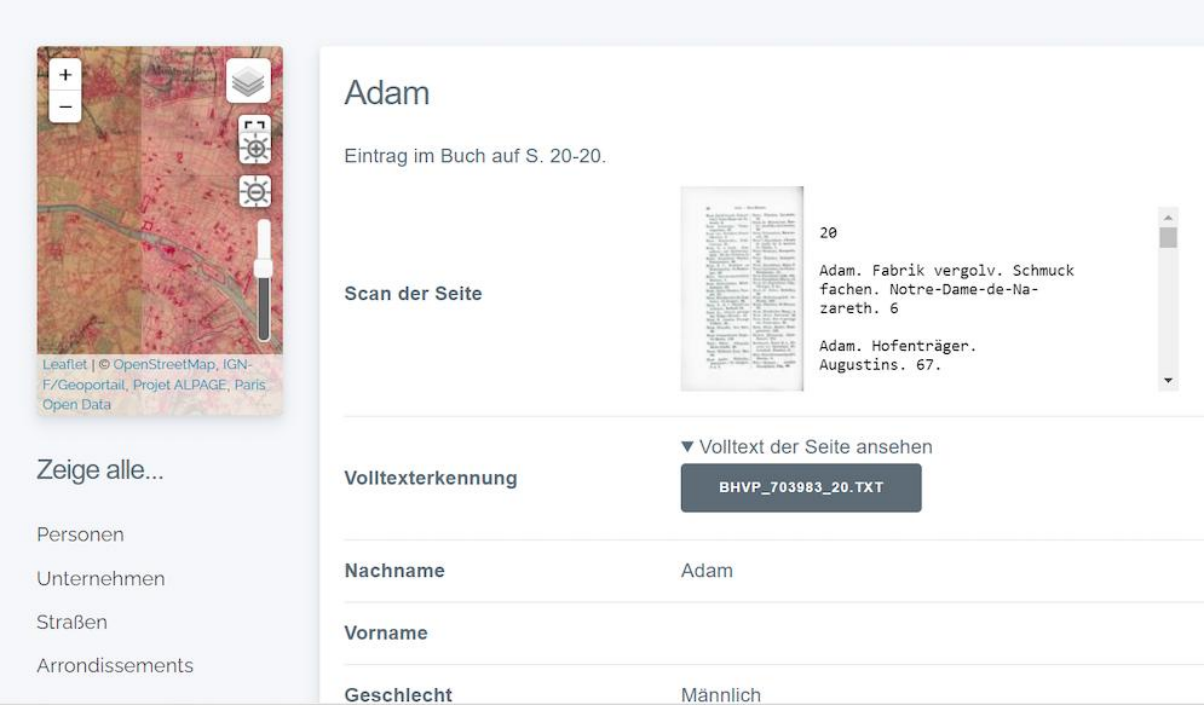
Mise en texte de fac-similés

Les numérisations ont également été OCRisées[13] afin de proposer sur le site une vue affichant le texte d'une page (voir figure 2). L'OCRisation a été réalisée en ligne de commande à l'aide d'un autre script Bash et le package Tesseract[14]. Cette librairie permet d'obtenir le contenu de chaque numérisation au format texte avec un taux d'erreur acceptable sur les documents imprimés datant du XIX^{e} siècle. Le jeu de données ainsi réalisé, déposé sur Zenodo en format ouvert (CSV, utf-8) et sous licence libre, se compose à la fois d'informations prosopographiques (métiers, adresses, grades militaires ou distinctions sont mentionnés) alignées géographiquement et d'une collection de numérisations de haute qualité.

Enrichir et aligner le jeu de données avec des autorités

Le travail de remodélisation s'est accompagné de l'alignement des anciennes données[15] avec le nouveau modèle au moyen du logiciel OpenRefine et de diverses autorités comme Wikidata ou encore des projets de recherche comme HISCO[16] qui recense et aligne en allemand et anglais les métiers anciens (Leeuwen 2002).

Figure 2. Rendu d'une page personnelle



[13] La reconnaissance optique de caractères, ou OCR, permet une traduction textuelle de caractères issus d'une image.

[14] Voir la documentation présentée par Ubuntu : https://doc.ubuntu-fr.org/tesseract-ocr

[15] Les données du projet initial sont également partagées sous la forme d'un classeur, https://zenodo.org/record/4030877

[16] Le projet, sa base et un outil permettant d'aligner jusqu'à 1 000 intitulés de métiers dans un même travail sont disponibles à l'URL https://historyofwork.iisg.nl

Légende : Exemple de rendu d'une fiche d'une personne avec ses données personnelles issues de la base, le contenu textuel brut issu de l'OCR et la source primaire correspondante numérisée pour la gestion de la preuve. Licence Creative Commons BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/, DHI Paris.

OpenRefine permet un enrichissement des données avec une multitude de possibilités. Dans le cadre de notre projet, nous avons pu associer les coordonnées GPS contenues dans chaque fiche de rue depuis Wikidata pour les associer aux différentes adresses des personnes présentées dans la base de données. Le processus se fait de façon semi-automatique, c'est-à-dire qu'une partie est automatisée par OpenRefine et une partie est manuelle si le logiciel ne détecte pas de parfaite correspondance, ou détecte une ambiguïté. Des connaissances en histoire de la ville de Paris furent mobilisées pour s'assurer de la correspondance entre les rues de 1854 et celles d'aujourd'hui[17].

Le dispositif de consultation

Nous avons évoqué précédemment que Web companion de FileMaker, l'outil pour créer la première interface, ne correspondait pas aux besoins des chercheurs d'aujourd'hui pour la consultation, le filtrage, la visualisation, l'accessibilité ou encore l'ergonomie du Web. Pour renouveler le projet, une interface suivant un modèle de conception *Model, View, Controller* (MVC) a été créée avec l'aide d'un *framework* Cake PHP.

Une nouveauté en rapport à l'interface de la première réalisation du projet est la cartographie. Il est possible de visualiser l'adresse des personnes et entreprises sur une carte de Paris grâce aux coordonnées géographiques calculées depuis l'association avec les données IGN. Pour afficher les latitude et longitude des entités de la base, une carte réalisée avec la bibliothèque de cartographie interactive Leaflet charge les coordonnées demandées au format JSON et positionne chaque individu ou société sur la carte.

Les données partagées, exposées et décrites

Un des objectifs dans le cahier des charges de la suite du projet *Adressbuch* a été de proposer notre jeu de données ou des parties du jeu en téléchargement libre. Dans l'interface du dispositif de consultation, la base entière est proposée au téléchargement dans des formats libres : CSV, JSON, SQL ou XML. Les utilisatrices et utilisateurs peuvent également choisir de télécharger les données de plusieurs personnes par lot ou selon les résultats d'une recherche : des données filtrées par personne, ou groupe. Pour une plus large diffusion des données du projet, la totalité est partagée dans les formats précédemment cités sur la plateforme de dépôt scientifique Zenodo sous une licence ouverte. Cela permet de rendre les données accessibles, découvrables, citables et réutilisables, grâce au *digital object identifier* (DOI) que la plateforme attribue aux jeux de données et à la licence CC-BY 4.0.

Dans une logique d'interopérabilité et une optique prosopographique à granularité fine, la fiche de chaque personne référencée dans la base est détectable par Zotero depuis le dispositif. Cette fiche peut donc être enregistrée par les chercheuses et chercheurs ou les généalogistes dans une base de connaissance de type Zotero avec des informations de type

[17] Un manuel très utile est : Jacques Hillaret, Dictionnaire historique des rues de Paris, Paris 1963. Voir aussi le site « *La nomenclature officielle des voies parisiennes* » : https://www.paris.fr/pages/les-voies-de-paris-denominations-et-numeros-d-immeubles-7550.

nom, prénom ou encore adresse. Cela peut être techniquement réalisé grâce à la norme NISO Z3988 (*Context Objects in Spans*) adaptée pour les besoins du projet à la plateforme CakePHP, ceci afin de correspondre au plus près aux recommandations d'ouverture et de description du FAIR[18].

Les résultats de recherche en méthodologie des humanités numériques

Sémiologie graphique et regard pluriel

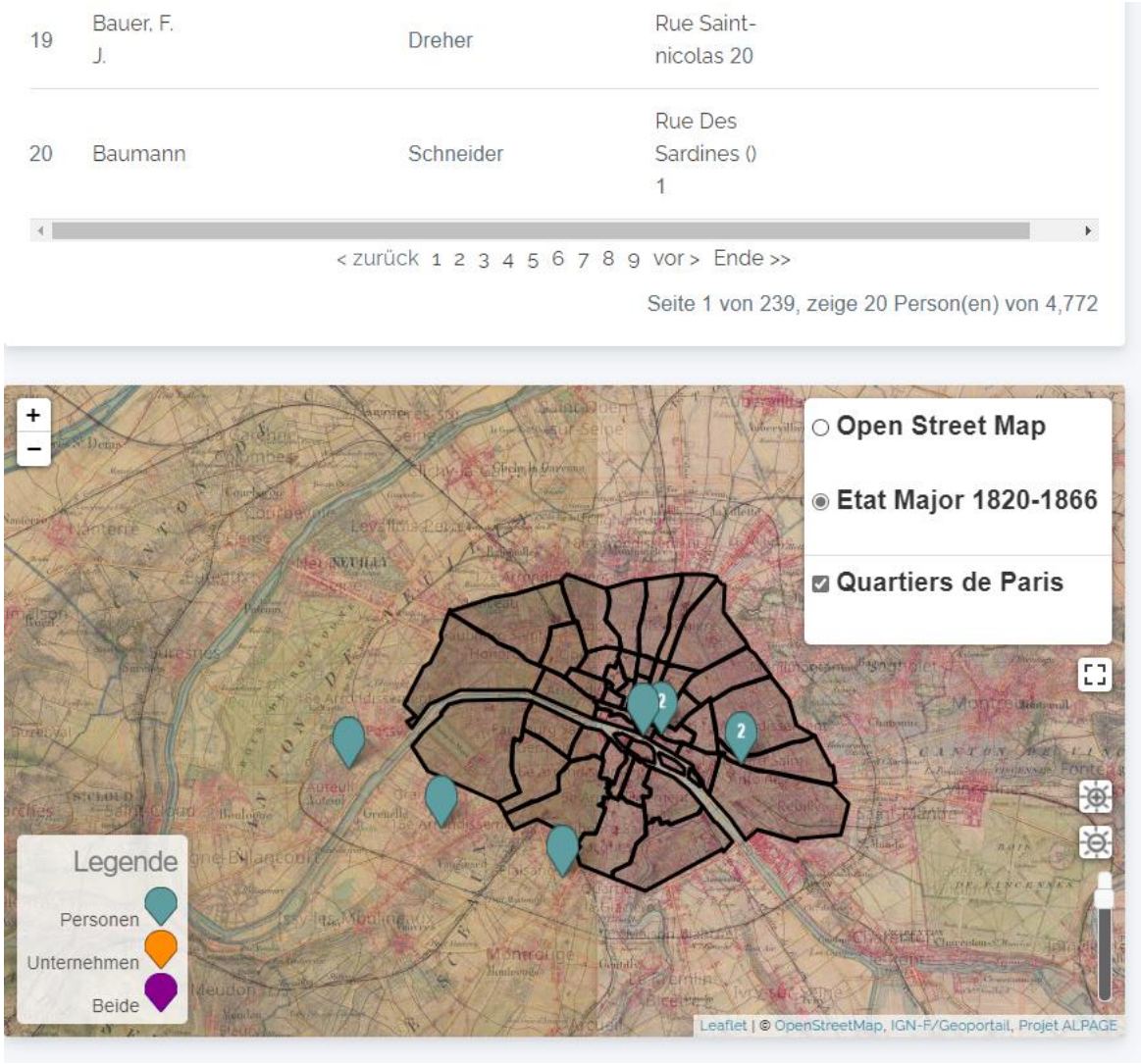
Le renouveau du projet d'analyse de l'*Adressbuch* lancé en 2020 a permis de faire une étude diachronique de Paris, c'est-à-dire étudier la capitale française à deux époques différentes. Rappelons qu'en 1854, les arrondissements parisiens ne correspondent pas à l'actuel découpage qui date de la Loi du 16 juin 1859 pour la création des vingt arrondissements, avec le décret impérial du 31 octobre 1859 sur la dénomination des nouveaux arrondissements[19]. En outre, Paris a connu de profondes mutations à la suite des travaux de Haussmann de 1854-1870, avec une destruction en partie du vieux Paris au détriment des nouveaux grands axes et boulevards. Pour une parfaite compréhension de la problématique, les adresses des personnes ont été indexées selon les deux découpages et il est possible de choisir si l'on souhaite visualiser la carte selon une des deux modalités.

Figure 3. Carte diachronique des immigrés allemands à Paris en 1854

[18] Voir note 1.

[19] Voir à ce propos cet autre document des archives de la ville de Paris :

http://archives.paris.fr/_depot_ad75/_depot_arko/articles/47/correspondance-entre-les-arrondissements-anciens-et-nouveaux_doc.pdf



Légende : Exemple de filtrage de données utilisant Leaflet avec les données de l'*Adressbuch* géo référencées enrichies de celles du projet Alpage et une carte d'état-major IGN. Licence Creative Commons BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/, DHI Paris.

La carte numérique moderne OpenStreetMap[20] a été accompagnée d'un fond de carte historique provenant de l'état-major français entre 1820-1866. Cette carte est disponible dans le catalogue de cartes IGN issues du Géoportail du gouvernement[21]. Les données peuvent donc figurer et être visualisées au choix sur une carte moderne avec le découpage actuel et dans les contextes administratifs et topographiques historiques du Paris de 1854. Cette double vision en diachronie, avec la possibilité d'une surimpression en couches multiples peut permettre d'interpréter l'impact historique potentiel du passage des Allemands à Paris sur la

[20] OpenStreetMap est une plate-forme collaborative et ouverte qui possède une grande communauté de contributeurs en France et ne cesse d'évoluer. OpenStreetMap est le service de carte qui s'inscrit le mieux dans notre projet en suivant les principes d'open data et d'open source en opposition à Google Maps, par exemple, qui est propriétaire de ses données et payant pour certaines fonctionnalités.

[21] Voir les données ouvertes IGN https://www.geoportail.gouv.fr/donnees/carte-de-letat-major-1820-1866

ville moderne et le contexte historique grâce à Opacity Controls[22], une extension de Leaflet. Mettre en place une cartographie multimodale avec un double découpage administratif est une nouveauté dans le cadre du projet. Cela se réfère aux principes d'expressivité graphique des données cartographiques tels qu'exprimés par Johanna Drucker (2011). Les données rassemblées, agencées, construites même, par le projet, ce que Drucker nomme "*capta*", peuvent être réalisées visuellement dans une relation de co-dépendance entre l'observatrice ou l'observateur et la carte : une analyse cartographique multiple est possible. Le choix d'une observation située en diachronie s'ouvre à l'utilisatrice ou l'utilisateur : soit une carte moderne, soit un fond de carte d'époque, soit également le découpage administratif effectif lors de l'observation initiale de la population des immigrés, soit celui qui est apparu quelques années plus tard et qui est toujours d'actualité, soit pas de découpage administratif du tout (voir figure 3).

Fiabilité des sources, reproductibilité des résultats

Dans un esprit d'ouverture et de partage des données, le code source partagé sur GIThub, en plus des simples données, permet de cloner la totalité du projet pour répliquer, s'approprier et modifier l'interface de consultation selon les besoins d'un nouveau projet : les focales spécifiques d'une autre discipline, la période étudiée ou encore les besoins de filtrage et/ou d'affichage[23]. Nous mettons à disposition les jeux de données primaires comme les numérisations de l'*Adressbuch* et la bibliographie qui nous a permis de mener à bien le projet. Puis, nous mettons également à disposition les jeux de données secondaires : la base de données avec les enrichissements issus des réconciliations avec les serveurs d'autorité ou encore les marqueurs cartographiques utilisés pour parfaire le dispositif de consultation. La véracité des données disponibles sur la plateforme en plusieurs formats est donc possible à vérifier et les potentielles erreurs peuvent être signalées. De plus, avec la publication des jeux de données, les analyses quantitatives induites présenteront des résultats qui pourront être reproduits, discutés, confirmés ou infirmés.

Conclusion

Mettre à disposition des chercheuses et chercheurs le contenu enrichi du livre *Adressbuch* leur permet d'explorer, de filtrer, d'analyser et de visualiser les données de façon plurielle et de les adapter à leurs propres besoins et questions de recherche. Ainsi, la base permet des observations statistiques et une représentation significative de la population allemande aisée au milieu du XIX^{e} siècle à Paris. Bien évidemment, sont explorées et visualisées ici seulement les données rassemblées par Kronauge, et non les données exhaustives sur l'immigration allemande à Paris en 1854. Néanmoins, considérant l'ampleur des données, leur représentativité et le fait que celles-ci portent en plus sur une époque sur laquelle la recherche historique repose souvent sur de pures spéculations, les résultats obtenus sont d'une grande importance. Ils permettent également de vérifier des thèses de recherche sur la migration et de corriger de petites erreurs, cette fois sur une base empirique.

La possibilité d'extraction et de téléchargement des données crée la condition préalable pour un *data mining* et une réutilisation individuelle. Pourtant, pour l'instant, les historiennes et historiens, en particulier, sont souvent très réticents à réutiliser les données provenant d'autres

[22] Voir https://github.com/lizardtechblog/Leaflet.OpacityControls

[23] Voir https://github.com/dhi-digital-humanities/Adressbuch1854

projets. Les principales réserves pour la réutilisation des données d'autres chercheurs sont la difficulté d'évaluer la véracité, la conformité aux sources et la qualité des données. Un *data paper* permet de rendre les données compréhensibles, tout comme les décisions prises pour les créer, enrichir et modéliser ainsi que les choix des formats et logiciels.

Nous avons rédigé ce data paper à trois, afin d'exprimer les différents points de vue disciplinaires sur le projet et ses données, et pour que le texte reste compréhensible à tous points de vue. Les valeurs d'une science ouverte sont la motivation fondamentale : un *data paper*, comme discours d'escorte auctorial, rend les données compréhensibles en explicitant leurs structures, les sources associées, les conditions de collecte et de transformation. Nous souhaitons ainsi activement participer à un courant de promotion, par la réutilisation, des mises en données des travaux de recherche qui corresponde à une science durable et éthique. Enfin, nous souhaitions proposer et rendre possibles des scénarios d'utilisation divers des jeux produits, tant dans la sphère de la recherche que dans le monde de la culture : comme le dit Rufus Pollock, fondateur de l'Open Knowledge Foundation : « *The best thing to do with your data will be thought of by someone else* »[24].

Questions aux auteurs

Votre chapitre "Data paper en humanités numériques" se présente sous une forme proche d'un article de recherche. Quelle logique a prévalu dans votre choix de publier un data paper sous cette forme ?

Oui, notre *data paper* est plus proche d'un article de recherche en sciences humaines et sociales que d'un simple *data paper* de modèle STM.

Nous avons pensé que cette forme est plus adaptée au sujet et à la discipline historique, cela faisait également du sens du côté des SIC.

Pouvez-vous nous indiquer comment s'est fait votre choix de licence pour la réutilisation des données et les problèmes éventuels auxquels vous avez été confrontés ?

Nous avons opté pour une licence CC-0 pour les numérisations des pages de l'*Adressbuch* imprimé, parce que celui-ci est dans le domaine public. Aucun problème de ce côté pour convaincre la bibliothèque historique de la Ville de Paris qui est en possession du livre et qui nous l'a prêté pour faire les scans.

Les données tirées de l'*Adressbuch* sont sous simple CC-BY pour pouvoir permettre une réutilisation et adaptation libre. Nous n'avons pas rencontré des problèmes.

Pouvez-vous préciser quels ont été vos choix pour répondre aux principes FAIR ?

Pour les règles du FAIR, nous avons décidé d'aller au maximum sur les données documentées et structurées dans la continuité des règles 4, 5 et 6[25] du *semantic publishing* t-q. présentées par David Shotton (2009, p. 93) à la fois pour proposer des données de qualités, alignées à des référentiels et données externes, mais aussi pour augmenter la visibilité des jeux de données de l'IHA (Kembellec, 2019).

---

[24] Traduction : La meilleure chose à faire avec vos données sera pensée par quelqu'un d'autre, voir, https://rufuspollock.com/misc/.

[25] Règles (4) : Utilisation les normes établies dans la mesure du possible. (5). Publier des ensembles de données brutes sur le Web. (6). Diffusez les métadonnées des articles, en particulier les listes de références, sous une forme lisible par machine.

Vous êtes-vous appuyés sur un template et/ou des recommandations particulières pour la rédaction de ce data paper ? Si oui, lesquels ?

Pour la structure du *data paper*, nous avons appliqué à minima la synthèse des réflexions que nous avons pu avoir par ailleurs avec Olivier Le Deuff (dans le numéro dédié de la *RFSIC*, en suivant la revue de la littérature[26]). Pour rappel : (1 et 2) une succincte description de données produites, leur contexte scientifique ainsi que leurs utilisations potentielles ; (3) une description de leur processus de production. (4) Les analyses ou procédures ayant permis de confirmer la validité des données décrites (confrontation de sources ou de données comparables). (5 et 6) Une note d'usage : une éventuelle procédure de réutilisation des données, la licence, l'accès aux données (url) et un éventuel accès au code informatique de reproduction du jeu de données.

Sur les données, les logiciels et les codes sources mobilisés

Logiciels utilisés

- Mocodo : http ://www.mocodo.net
- OpenRefine : https://openrefine.org
- Cake PHP : https://cakephp.org/
- phpMyAdmin : https://phpmyadmin.net
- Leaflet : Leaflet - a JavaScript library for interactive maps, https://leafletjs.com
- Opacity Controls : https://github.com/lizardtechblog/Leaflet.OpacityControls
- Shapefile Leaflet : https://github.com/calvinmetcalf/leaflet.shapefile
- Commande « file » : commande shell Unix/Linux qui permet essentiellement de déterminer le type MIME d'un fichier. http://darwinsys.com/file/
- ImageMagick : https://imagemagick.org/
- Tesseract : https://doc.ubuntu-fr.org/tesseract-ocr
- Projet Jupyter : https://jupyter.org/

Données mobilisées et partagées

- Référence primaire « Adressbuch » :
  F.A. Kronauge, *Adressbuch der Deutschen in Paris für das Jahr 1854. Vollständiges Adressverzeichnis aller in Paris und seinen Vorstädten wohnenden selbständigen Deutschen in alphabetischer Ordnung. Nebst Angaben der Sehenswürdigkeiten und Wohnungen der Gesandten*, Paris 1854.
  - Bibliothèques spécialisées et patrimoniales de la Ville de Paris, https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0000884072
- Jeux de données du projet « Adressbuch » :
  https://doi.org/10.5281/zenodo.7427439

[26] Cf. Gérald Kembellec et Olivier Le Deuff, « Poétique et ingénierie des data papers », Revue française des sciences de l'information et de la communication [En ligne], 24 | 2022, mis en ligne le 01 janvier 2022, consulté le 05 janvier 2023. URL : http://journals.openedition.org/rfsic/12938 ; DOI : https://doi.org/10.4000/rfsic.12938

- Projet « ALPAGE » : AnaLyse diachronique de l'espace urbain PArisien : approche GEomatique - Alpage (huma-num.fr), https://alpage.huma-num.fr/
- Géoportail, carte de l'état-major (1820-1866). Carte française en couleurs du XIXe siècle en couleurs, superposable aux cartes et données modernes. https://www.geoportail.gouv.fr/donnees/carte-de-letat-major-1820-1866
- Correspondance des anciens (pré-1860) et nouveaux arrondissements de Paris : http://archives.paris.fr/_depot_ad75/_depot_arko/articles/47/correspondance-entre-les-arrondissements-anciens-et-nouveaux_doc.pdf
- *History of Work information system*: https://iisg.amsterdam/en/data/data-websites/history-of-work

Codes sources partagés et interfaces de consultation

- *Adressbuch der Deutschen in Paris aus dem Jahr 1854*, dispositif de consultation du projet « Adressbuch » : http://adressbuch1854.dhi-paris.fr/
- Code source du projet « Adressbuch »: https://github.com/DH-Cologne/Adressbuch1854
- Jupyter Notebooks utilisant les données du projet « Adressbuch »:
  - https://zenodo.org/record/5512502#.Yi99zLjjJpR
  - https://github.com/dhi-digital-humanities/Adressbuch-Notebook

Références

Dietrich, Karin Marie-Héléne Varnier. 1987. « Les Allemands naturalisés en France de 1791-1858 ». *Cahiers d'études germaniques* 3: 4-34.

Drucker, Johanna. 2011. « Humanities Approaches to Graphical Display ». *Digital Humanities Quarterly* 5 (1) : 1-21.

Gaillard, Jeanne. 1997. *Paris, la ville (1852-1870*. Paris: L'Harmattan.

Grandjonc, Jacques. 1972. « Éléments statistiques pour une étude de l'immigration ». In *Archiv für Sozialgeschichte*, 12:487-533.

Grandjonc, Jacques. 1974. « Les étrangers à Paris sous la Monarchie de Juillet et la Seconde République ». In *Population, Cahier de l'INED. Numéro spécial Migrations*, 61-88. Paris mars.

Grandjonc, Jacques. 1983. « Émigrés français en Allemagne, Émigrés allemands en France 1685-1945 ». Paris : Goethe Institut.

Green, Nancy. 1998. « Du Sentier à la 7ème Avenue : la confection et les immigrés ». Paris - New York.

Kembellec, Gérald. 2019. « Semantic publishing, la sémantique dans la sémiotique des codes sources d'écrits d'écran scientifiques », In *Les Enjeux de l'information et de la communication*, vol. 20/2, no. 2, pp. 55-72. DOI : https://doi.org/10.3917/enic.027.0055

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

König, Mareike. 2003. *Deutsche Handwerker Arbeiter und Dienstmädchen in Paris*. Oldenbourg Wissenschaftsverlag. DOI : https://doi.org/10.1524/9783486834383.

König, Mareike. 2006. « Les Allemands à Paris au XIXe siècle. » In *Annuaires de l'École pratique des hautes études*, 387-89. Paris : École pratique des hautes études. https://www.persee.fr/doc/ephe_0000-0001_2004_num_20_1_11528.

Kronauge, F.-A. 1854. *Adreßbuch der Deutschen in Paris für das Jahr 1854 oder vollständiges Adreßverzeichniß aller in Paris und seinen Vorständten wohnenden selbst andigen Deutschen in alphabetischer Ordnung*. https://bibliotheques-specialisees.paris.fr/ark:/73873/pf0000884072.

Leeuwen, M. 2002. *HISCO : Historical International Standard Classification of Occupations*. Leuven: Leuven University Press.

Moch, Leslie Page. 2008. « Frankreich ». In *Enzyklopädie Migration in Europa. Vom 17. Jahrhundert bis zur Gegenwart*, édité par Klaus Bade, 122-41.

Noiriel, Gérard. 1992. *Population, immigration et identité nationale en France : XIXe-XXe siècle*. Paris : Hachette.

Noizet, Hélène ; Bove, Boris et Costa, Laurent. 2013. « Paris de parcelles en pixels. » In *Analyse géomatique de l'espace parisien médiéval et moderne*, édité par Comité d'histoire de la Ville de Paris. Paris : Presses universitaires de Vincennes.

Pinon, Pierre. 2012. « Atlas du Paris haussmannien. La ville en héritage du Second Empire à nos jours ». Parigramme.

Schieder, Wolfgang. 1963. « Anfänge der deutschen Arbeiterbewegung ». In *Die Auslandsvereine im Jahrzehnt nach der Julirevolution von 1830*. Stuttgart.

Schreibman, Susan, Ray Siemens, et John Unsworth. 2004. *A Companion to Digital Humanities*. Blackwell Publishing.

Shotton, David. *(2009). « Semantic publishing: The coming revolution in scientific journal publishing ». Learned Publishing.* ***22*** *(2): 85–94.* DOI : https://doi.org/10.1087/2009202

Thillay, Alain. 1999. « Les artisans étrangers au faubourg Saint-Antoine à Paris (1650–1793) ». In *Les Étrangers dans la ville. Minorités et espaces urbains du Moyen âge à l'époque moderne*, édité par Jacques Bottin et Donatella Calabi, 261-69. Paris.

Werner, Michael. 1995. « Étrangers et immigrants à Paris autour de 1848 : L'exemple des Allemands ». In *Paris und Berlin in der Revolution*, édité par Ilja Mieck, Horst Möller, et Jürgen Voss, 199-213. Stuttgart.